

**MITOCHONDRIAL GENOME INFORMATICS OF SELECTED
SERI-BIORESOURCES AND DATABASE DEVELOPMENT**

A Thesis

Submitted for the Degree of

DOCTOR OF PHILOSOPHY

by

**DEEPIKA
(11610621)**

Under supervision of

Prof. Utpal Bora



March 2018

**Department of Biosciences & Bioengineering
Indian Institute of Technology Guwahati
Guwahati 781 039, Assam, India**



**INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI**

Department of Biosciences & Bioengineering

DECLARATION

I do hereby declare that the content embodied in this thesis entitled **“Mitochondrial Genome Informatics of Selected Seri-bioresources and Database Development”** is the result of investigations carried out by me under the supervision of Prof. Utpal Bora and is submitted to the Indian Institute of Technology Guwahati, Guwahati, Assam, India for the award of degree of *Doctor of Philosophy in Biosciences and Bioengineering*.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made wherever the work described is based on the findings of other investigators.

March, 2018

Deepika

Roll No. 11610621



**INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI**

Department of Biosciences & Bioengineering

CERTIFICATE

It is to certify that the work described in this thesis entitled “**Mitochondrial Genome Informatics of Selected Seri-bioresources and Database Development**” by **Ms. Deepika** (Roll No. 11610621) for the award of degree of Doctor of Philosophy is an authentic record of the results obtained from the research work carried out under my supervision in the Department of Biosciences & Bioengineering, Indian Institute of Technology Guwahati, Guwahati, India. The work embodied in this thesis has not been submitted elsewhere for a degree.

March, 2018

Prof. Utpal Bora

(Thesis Supervisor)

Department of Biosciences & Bioengineering

Indian Institute of Technology Guwahati

Guwahati-781 039, India

***D**edication*

The thesis is dedicated to

My Parents, Brother, Sisters,

Respected Teachers,

&

Almighty God

... Deepika

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor, Prof. Utpal Bora, Department of Biosciences and Bioengineering, for his valuable advice, guidance, support and encouragement throughout my research work. He has always been the leading light of my research endeavor for the meaningful and fruitful research in the fast growing field of biotechnology and bioinformatics. I must acknowledge the freedom he has given me in every step of my research work. I must thank my doctoral committee members, Prof. Venkata Dasu, Dr. Ranjan Tamuli, Dr. Kusum Kumari Singh and Prof. Pinakeshwar Mahanta for their valuable suggestions and constant advice which helped me in improving my work.

I would take this opportunity to acknowledge the present and past Heads of the Department of Biosciences and Bioengineering for providing all the essential facilities and a conducive academic environment. I would also like to acknowledge the Centre for the Environment for providing me all the support and facilities. I sincerely thank the financial support from Ministry of Human Resource Development (Govt. of India) in the form of fellowship and Department of Biotechnology (Govt. of India) for providing research grants to my supervisor.

I would like to extend my thanks to Dr. Kartik Neog, CMER&TI; Azizur Rahman Khan, Khanapara and Mr. Keshav Kalita, Boko for helping me in the sample collection

required for my thesis work. I would also like to thank the departmental staff members Mr. Nurul, Mrs. Prarthana, Mr. Dipankar, Mr. Pankaj, Mr. Bidyut, Mr. Chandan, Mr. Raghuveer and Mr. Niranajan for their constant co-operation.

My heartiest thanks to my lab members Dr. Shekhar, Dr. Anil, Dr. Suradip and Dr. Arghya, Dr. Manav, Dr. Ajoy, Shamim Rahman, Dr. Jintu, Dr. Sunita, Swagata, Papori, Hasnahana, Debajyoti, Vimal, Adhiraj, Biju, Dharitri, Manas, Tinka, Rupanjali, Shamim, Deepjyoti, Krishna, Smita and Jyotirekha for providing a healthy environment in the lab. A special thanks to Pragya Ma'am for her constant support and immense love.

I would like to thank my friends in the institute especially Dr. Muthusivaramapandian, Dr. Basavaraj, Dr. Vikram, Saumya and Mehak who were always supportive to me. I also thank Anuma, Aruna, Nivedita, Mohan, Yoganand, Himanshu, Arun, Thiagarajan, Somaiah, Shrikant, Payal, Damodharan, Nandkishore and Rupak for being supportive during my Ph.D., whenever I needed.

Last but not the least, I would like to specially thank my parents, brother and sisters for their blessings, love, patience, support and understanding throughout my studies and most of all to the Almighty God who made everything possible.

March 2018

Deepika

Contents

Synopsis	i
List of figures	xi
List of tables	xxiii
1. Introduction and Review of Literature	1
1.1 Seribiodiversity	1
1.2 Molecular characterization using DNA Barcoding	9
1.2.1 DNA barcoding in insects and animals	10
1.2.2 DNA barcoding in plants	13
1.2.3 DNA barcoding in bacteria	16
1.2.4 DNA barcoding in fungus	17
1.2.5 DNA barcoding in virus and microsporidians	18
1.3 Mitochondrial genome sequencing of wild silkmoths	20
1.3.1 Genome rearrangement and gene content	22
1.3.2 Strand asymmetry in the mitogenome	23
1.3.3 Transfer RNAs (tRNAs)	25
1.3.4 Protein coding genes (PCGs)	27
1.3.5 Non-coding regions	31
1.3.6 Ribosomal RNAs (rRNAs)	33
1.3.7 Importance of mitogenome studies in Saturniid silkmoths (wild silkmoths)	34
1.4 Seri-informatics databases	36
1.4.1 Silkworm databases	37
1.4.1.1 Nucleotide databases	37
1.4.1.2 Protein databases	44
1.4.1.3 Silkworm genetic resource databases	45
1.4.1.4 Insect pathway databases	45
1.4.2 Silkworm host plant databases	46
1.4.2.1 Databases of mulberry	46

1.4.2.2	Databases of castor	49
1.4.2.3	Databases of papaya	49
1.4.2.4	Databases of <i>Jatropha</i>	49
1.4.2.5	Databases of cassava	50
1.4.2.6	Databases of <i>Quercus</i>	50
1.4.2.7	Other plant databases	50
1.4.3	Pest and pathogen databases	51
1.4.4	Combined databases	51
1.4.5	Need of seri-related databases	53
2.	Molecular characterization of selected Seri- bioresources of Assam through DNA Barcoding	57
2.1	Introduction	57
2.2	Materials and methods	60
2.2.1	Sample collection	60
2.2.2	Genomic DNA extraction	61
2.2.3	PCR amplification and sequencing of barcode regions	64
2.2.4	Sequence analysis and submission to database	66
2.2.5	Phylogenetic analysis	67
2.3	Results and discussion	76
2.3.1	Genomic DNA extraction	76
2.3.2	PCR amplification and sequencing of barcode regions	77
2.3.3	Sequence analysis and submission to database	83
2.3.4	Phylogenetic analysis	92
2.4	Conclusions	100
3.	Mitochondrial genome sequencing of <i>Antheraea assamensis</i>, a semi-domesticated silkmoth endemic to Assam and North-eastern part of India	103
3.1	Introduction	103
3.2	Materials and methods	105
3.2.1	Sample processing, DNA sequencing and assembly	105

3.2.2 Genome annotation, visualization and comparative analysis	108
3.2.2.1 <i>Comparative mitogenome analysis</i>	110
3.2.2.2 <i>Comparative analysis among protein coding genes (PCGs)</i>	110
3.2.2.3 <i>Comparative analysis among transfer RNAs (tRNAs)</i>	110
3.2.2.4 <i>Comparative analysis among ribosomal RNAs (rRNAs)</i>	111
3.2.2.5 <i>Comparative analysis among control region (A+T rich region)</i>	111
3.2.2.6 <i>Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions</i>	111
3.2.2.7 <i>Comparative analysis with respect to nucleotide composition, skewness and codon usage</i>	111
3.2.3 Comparative phylogenetic analysis	112
3.3 Results and discussion	113
3.3.1 Sample processing, DNA sequencing and assembly	113
3.3.2 Genome annotation, visualization and comparative analysis	116
3.3.2.1 <i>Comparative mitogenome analysis</i>	116
3.3.2.2 <i>Comparative analysis among protein coding genes (PCGs)</i>	120
3.3.2.3 <i>Comparative analysis among transfer RNAs (tRNAs)</i>	130
3.3.2.4 <i>Comparative analysis among ribosomal RNAs (rRNAs)</i>	142
3.3.2.5 <i>Comparative analysis among control region (A+T rich region)</i>	142
3.3.2.6 <i>Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS)</i>	144

	<i>regions</i>	
3.3.2.7	<i>Comparative analysis with respect to nucleotide composition, skewness and codon usage</i>	149
3.3.3	Comparative phylogenetic analysis	158
3.4	Conclusions	161
4	Mitochondrial genome sequencing of <i>Cricula trifenestrata</i>, a wild silkmoth from Assam	163
4.1	Introduction	163
4.2	Materials and methods	164
4.2.1	Sample processing, sequencing and assembly	165
4.2.2	Genome annotation, visualization and comparative analysis	165
4.2.3	Comparative phylogenetic analysis	166
4.3	Results and discussion	171
4.3.1	Sample processing, sequencing and assembly	171
4.3.2	Genome annotation, visualization and comparative analysis	173
4.3.2.1	<i>Comparative mitogenome analysis</i>	174
4.3.2.2	<i>Comparative analysis among protein coding genes (PCGs)</i>	177
4.3.2.3	<i>Comparative analysis among transfer RNAs (tRNAs)</i>	183
4.3.2.4	<i>Comparative analysis among ribosomal RNAs (rRNAs)</i>	187
4.3.2.5	<i>Comparative analysis among control region (A+T rich region)</i>	191
4.3.2.6	<i>Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions</i>	193
4.3.2.7	<i>Comparative analysis with respect to</i>	196

	<i>nucleotide composition, skewness and codon usage</i>	
	4.3.3 Comparative phylogenetic analysis	204
	4.4 Conclusions	209
5	Development of Seri-related Web-resources	211
	5.1 Introduction	211
	5.2 Materials and methods	215
	5.2.1 Database content-retrieval, curation and categorization	217
	5.2.2 Design and implementation of SeriPort and Seribase	221
	5.3 Salient features and accessibility of SeriPort and Seribase	223
	5.4 Data access in the portal and database	225
	5.5 Data submission, download and updates	228
	5.6 Cross-referencing and pictorial data in Seribase	229
	5.7 Conclusions	229
6	Summary and Future prospects	231
	Bibliography	237
	List of Publications	287
	List of Conferences/Workshops/Symposia	289
	Vitae	291

Synopsis

Northeastern states of India such as Assam, Meghalaya, Nagaland and Manipur occupy a unique place in the world as they are one of the hot-spots of seribiodiversity. Seribiodiversity refers to the variability in sericigenous insects and their interacting partners like host plants, pests, predators and pathogens. All five kinds of commercially important silks such as mulberry, muga, eri, tasar and oak tasar are cultivated in this region. Assam itself offers a favorable environment for many silkmooths ranging from domesticated (*Bombyx mori*, *Samia ricini*) to wild silkworms (*Antheraea assamensis*, *A. pernyi*, *A. frithi*, *Actias selene*, *Attacus atlas*, *Cricula trifenestrata*, etc). Wild or non-mulberry silkmooths mainly belong to Saturniidae family of the order Lepidoptera with diversity in several characteristics such as medium to large in size, uni- or multi-voltine nature depending upon the climatic conditions, feeding behavior, rich or bright colored silk, etc. Similarly, there exists a great diversity in the host plants which influence the growth and development of silkworms and in turn affect the quality and quantity of silk. There are several other factors which affect the diversity of these silkworms and their host plants such as pests, pathogens, predators, pesticides, climatic conditions, deforestation, etc. Many of these wild seri-bioresources are economically and ecologically important resources and can be utilized for commercial applications. However, despite of huge diversity in sericigenous insects, the current knowledge on their diversity, identification and characterization is limited. This has necessitated immediate need to explore, characterize and document the

resources through various approaches like DNA barcoding, omics (organellar genomics, transcriptomics, proteomics, etc.) tools and informatics. DNA barcoding has achieved significant progress in the identification of various species based on short and standard DNA sequences. Organellar (mitochondrial for animals and chloroplast for plants) genomics greatly contributes towards molecular systematics, evolutionary and comparative genomic studies based on large informative characters by comparing entire genome structure and sequence. However, the characterization data based on the above two approaches on the wild seri-resources is surprisingly sparse. Application of informatics has resulted in a plethora of information and development of various databases related to seri-resources especially domesticated ones. A common platform comprising of all varieties of data offered on sericulture field over the internet is an urgent need. Further, an integrated database comprising of integrated information on the wild silkworms, host plants, pest and pathogens is not yet available.

Formulation of Objectives: Based on the literature review, the objectives were formed and are categorized into two parts- Part A and Part B as follows-

Part A: Molecular characterization and mitochondrial genome sequencing of selected seri-bioresources of Assam

Objective 1: Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

Objective 2: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkmoth endemic to Assam and North-eastern part of India

Objective 3: Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkmoth from Assam

Part B: Database development related to Seri-bioresources

Objective 4: Development of Seri-related Web-resources

Based on the above objectives, the thesis is organized into various chapters as described herein.

Chapter 1: Introduction and Review of Literature

This chapter describes a systematic review of literature conducted as per the objectives framed for the present study. A brief introduction about the seribioresources of North-east India has been provided especially for Assam seribiodiversity and their importance. This chapter further discusses the utility of DNA barcoding approach in the identification and characterization of insects, host plants, bacteria, fungus, etc. The mitochondrial genome sequencing and characterization studies of lepidopteran insects, particularly wild silkmoths and their importance in evolutionary studies have also been discussed. Finally the chapter portrays information of various seri-related databases or web-resources and their importance in sericulture area.

Chapter 2: Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

This chapter describes barcoding of 18 indigenous seri-bioresources which is the first report on muga and eri silkworms, host plants, pests and pathogens. The sequence homology and sequence divergence revealed that the selected barcode markers (*COI*, *matK*, *rbcL*, *trnH-psbA*, ITS, ITS2 and 16S *rRNA*) were efficient in discriminating the selected species. Further, the phylogenetic trees constructed using suitable barcode regions clustered the species into their respective groups (related species and family). All the sequences have been submitted to NCBI GenBank database with their corresponding accession numbers. This study is expected to be useful for the researchers working in the exploration of silkworm resources.

Chapter 3: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkworm endemic to Assam and North-eastern part of India

Muga (*Antheraea assamensis*) is an economically important silkworm endemic to North-eastern part of India and is the producer of the strongest known commercial silk. However, there is a scarcity of genomic and proteomic data for understanding the organism at a molecular level. This chapter aims at decoding the complete mitochondrial genome (mitogenome) of *A. assamensis* using next generation sequencing technology and compares it with other lepidopteran mitogenomes. Mitogenome of *A. assamensis* is an AT rich circular molecule of 15,272 bp (A+T content ~80.2%). It contains 37 genes comprising of 13 protein coding genes (PCGs), 22 tRNA and 2 rRNA genes along with a 328 bp long control region. Its typical $tRNA^{Met}-tRNA^{Ile}-tRNA^{Gln}$ arrangement differed from ancestral insects ($tRNA^{Ile}-tRNA^{Gln}-tRNA^{Met}$). Two PCGs *cox1* and *cox2* were found to have CGA and GTG as start codons, respectively as reported in some lepidopterans. Interestingly, *nad4l* gene showed higher transversion mutations at intra-species than inter-species level. All PCGs evolved under strong purifying selection with highest evolutionary rates observed for *atp8* gene while lowest for *cox1* gene. We observed the typical clover-leaf shaped secondary structures of tRNAs with a few exceptions in case of $tRNA^{Ser1}$ and $tRNA^{Tyr}$ where stable DHU and TΨC loop were absent. A significant number of mismatches (35) were found to spread over 19 tRNA structures. The control region of mitogenome contained a six bp (CTTAGAG) deletion atypical of other *Antheraea* species and lacked tandem repeats. Phylogenetic position of *A. assamensis* was consistent with the traditional taxonomic classification of Saturniidae. The complete annotated mitogenome is

available in GenBank (Accession No. KU379695). This the first report on complete mitogenome of *A. assamensis* to the best of our knowledge.

Chapter 4: Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkmoth from Assam

C. trifenestrata is a wild silkmoth of Assam that produces golden yellow colored silk which can prove as a valuable natural silk resource for textile, biomedicine and other applications. This chapter describes the sequencing of *Cricula* mitogenome using next generation sequencing technology followed by characterization and comparative analysis with other sequenced lepidopteran mitogenomes. Our study revealed that *C. trifenestrata* mitogenome exhibited lowest A+T content as compared to *A. assamensis* and the selected organisms of Lepidoptera order. The typical gene arrangement ($tRNA^{Met}$ - $tRNA^{Ile}$ - $tRNA^{Gln}$) was different from the ancestral type. All protein coding genes exhibited strong purifying selection but *atp8* gene was comparatively higher than others, suggesting its role in energy requirement for silk production. This was concordant with gene-by-gene divergence studies which reflected higher proportion of informative sites as compared to other genes. AT biasness was detected in the codon usage of protein coding genes in *Cricula* mitogenome (15, 425 bp) which showed high frequency of codons rich in A or T nucleotides at the third codon position while low GC rich codons. All the tRNAs folded into typical clover leaf secondary structures except in $tRNA^{Ile}$, $tRNA^{Arg}$ and $tRNA^{Ser1}$ genes which lacked T Ψ C loop, DHU loop and stable DHU arm, respectively. The control region displayed conserved regions like two conserved blocks (greater than 50 bp), poly T-stretch, microsatellite (TA/AT)_n like elements and poly-A stretch consistent with that of *A. assamensis* and other lepidopterans. However, it differed from other

organisms in terms of the presence of TTAGA motif and two 35 bp tandem repeat units in the control region. The phylogenetic analysis revealed that *C. trifenestrata* is close to Saturniidae family members clustering in Saturniidae family. The annotated mitogenome sequence of *C. trifenestrata* has been submitted to GenBank under the accession number KY644697.

Chapter 5: Development of Seri-related Web-resources

The review of literature as well as exploration of the internet for all available web resources culminated into a huge amount of seri-database related data. Some of these databases are not easily available to users due to their minimal turn-up during common searches using popular search engines. Again, some of the related web-resources are not published in literature and hence, are not known to common people. Further, most of the databases are available on domesticated silkworms and their host plants, pests, pathogens, etc. However, currently no database comprises of integrated information on wild silkworms, their host plants, pests and pathogens on a single platform.

In order to address these issues, we have created an HTML based web portal named “**SeriPort**” and a seri-bioresource database named “**Seribase**”. This chapter describes development of a portal and a database which are freely available at www.seriport.in and www.seribase.in. **SeriPort** acts as a common platform for all varieties of data offered on sericulture field over the internet. The database for SeriPort was designed using Basic HTML 5 and CSS 3 for front-end, and PhpMyAdmin and MySQL 5 for the back end. The working language for the construction of website is PHP and HTML 5. The connectivity between front end and back end was done by using PHP. The central data (75 databases) in SeriPort was categorized into four classes; (i) silkworm DB (22 No.s), (ii) host

plant DB (25 No.s), (iii) pest and pathogen DB (02 No.s), and (iv) combined DB (26 No.s). The main features of SeriPort include a user-friendly and dynamic user interface, in-built search engine and data download as well as data submission options. **Seribase** is an integrated repository for more than 1600 wild seri-bioresources (501 species of wild silkmoths, 1023 host plants of the respective silkmoths, 35 disease pathogens and 90 pests of wild silkmoths) of economic or ecological significance. More than 300 patents have been included in the database till now. It comprises of datasets categorized into six different modules- “Silkworm”, “Host plants”, “Silkworm Pests”, “Silkworm Pathogens”, “Patents” and “Related Publications”. Seribase has been implemented in a PHP (Hypertext Preprocessor) module with php and HTML scripts running in the front end whereas MySQL database and Apache web server at the back end. The complete architecture resides upon a Linux-based server hosted at Arvix. It has an easy-to-use web-interface equipped with various options like search, browse and download for easy retrieval of data. Both these web-resources are expected to act as supportive platforms between general users and the niche occupied by sericulture in the internet.

Chapter 6: Summary and Future Prospects

The present thesis describes the identification and characterization of few seri-resources of Assam based on molecular methods like DNA barcoding and mitochondrial genome sequencing. The seri-resources in the present study have been barcoded for the first time based on universal barcode markers. Sequences generated for these resources will be helpful for future characterization and phylogenetic studies of the seri-resources. This study can be integrated with classical taxonomical methods for developing pest control strategies, disease

diagnosis and prior treatment, etc. The mitogenome characterization studies performed for wild silkmoths like *A. assamensis* and *C. trifenestrata* will facilitate utilizing these silkmoths for various studies like comparative genomics, evolutionary and phylogenetic studies with other wild silkmoths. Functional studies can be performed based on the mitochondrial protein coding genes which will help in understanding the expression of genes under various conditions. Further, the portal and database developed in the present study are expected to be extremely useful for the researchers working in the field of sericulture, seribiodiversity and seri-informatics. In future, more databases can be created to explore the area which is still in its infancy.

Published manuscripts

1. **Singh, D.**, Kabiraj, D., Sharma, P., Chetia, H., Mosahari, P. V., Neog, K., & Bora, U*. (2017) The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. *PLOS ONE*, 12(11), e0188077.
2. **Singh, D.**, Chetia, H., Kabiraj, D., Sharma, S., Kumar, A., Sharma, P., Deka, M., & Bora, U.* (2016) A comprehensive view of the current web-resources in sericulture and related fields. *Database*, 2016, 1-31.
3. **Singh, D.**, Sett, A., Sharma, P., Deka, M., & Bora, U.* (2016) DNA Barcoding- Current status and future prospects in '*Biotechnological tools for genetic resources*'. Ed. Prof. P. C. Deka. Daya Publishing House. Astral International Pvt. Ltd. New Delhi, pp 19-58.

Manuscripts under review/preparation

1. **Singh, D.**, Kabiraj, D., Chetia, H., Sharma, P., Neog, K., & Bora, U.* (2017) Complete mitochondrial genome of *Cricula trifenestrata* and its comparative analysis with other lepidopteran insects. (Under preparation).
2. **Singh, D.**, Sharma, P., Neog, K., & Bora, U.* (2017) DNA barcoding of few seribioresources of Assam, India. (Under preparation).

Manuscripts from collaborative work

1. Ojha, S., **Singh, D.**, Sett, A., Chetia, H., Kabiraj, D., & Bora, U.* (2017) Nanotechnology in Crop Protection in '*Nanomaterials in Plants, Algae and Micro-organism: Concepts and Controversies*'. Eds. D. K.Tripathi, P. Ahmad, S. Sharma, D. K. Chauhan, N. K. Dubey. Elsevier, pp 345-391.
2. Chetia, H., Kabiraj, D., **Singh, D.**, Mosahari, P. V., Das, S., Sharma, P., Neog, K., Sharma, S., Jayaprakash, P., & Bora, U.* (2017) De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). *Gene*, 611, 54-65.
3. Kabiraj, D., Kalita, J. J., Chetia, H., **Singh, D.**, & Bora, U.* (2015) Expanding the frontiers of rice research through omics. *Journal of Assam Science Society*, 56(2), 1-28.
4. Bora, U.*, Sett, A., & **Singh, D.** (2013) Nucleic Acid Based Biosensors for Clinical Applications. *Biosensors Journal*, 2, 104.
5. Mosahari, P. V., **Singh, D.**, Kalita, J. J., Sharma, P., Chetia, H., Kabiraj, D., Mahanta, C., & Bora, U.* Nanotoxicity: Impact on Health and Environment. CRC Taylor & Francis. (Under review).

* represents corresponding author



List of figures

- Figure 1.1** Interactions between silkworm, host plants, pest/parasitoid/competitor and pathogens in a seri-ecosystem.
- Figure 1.2** COI gene of mitochondrial genome in insects.
- Figure 1.3** Map of nuclear rRNA genes and their ITS regions in plants.
- Figure 1.4** Structure of mitochondria showing membrane protein complexes of the respiratory chain. Complex I (NADH/ubiquinone oxidoreductase, II (succinate dehydrogenase), III (cytochrome c reductase), IV (cytochrome c oxidase) and V (mitochondrial ATP synthase) [Adapted from Kühlbrandt, 2015]
- Figure 1.5** Arrangement of genes in the mitogenome of lepidopteran insects. (A) Ancestral arrangement with *trnI-trnQ-trnM* (IQM), (B) Arrangement *trnM-trnI-trnQ* (MIQ) in other sequenced lepidopteran insects diverged from ancestors during evolution. Other rearrangements are highlighted in red boxes observed in few lepidopterans.
- Figure 1.6** Origin of replication in control region (in red) and initiation of replication. (A) Vertebrate mitogenome where O_H denotes heavy strand, O_L denotes minor strand. (B) Insect mitogenome where O_J denotes major strand, O_N denotes minor strand. [Source: Arunkumar & Nagaraju, 2006]
- Figure 1.7** An example of secondary structures of tRNA genes in lepidopteran insects. (A) Typical clover leaf like structure (*trnM*) with four defined arms- AA: aminoacyl acceptor arm, DHU: dihydrouridine arm, AC: anticodon arm, TΨC: pseudouridine arm; (B) Secondary structure of tRNA gene (*trnS1*) with unstable DHU arm (stable stem and loop, highlighted in red). cat and gct in brackets denote anticodons of tRNA genes. Mismatches are

highlighted in green (G-U pairing).

Figure 1.8 Trend of publications entitled “Mitochondrial genome of Lepidopteran insects” (Source: Scopus dated 14.08.2017).

Figure 1.9 Frequency of mitochondrial genome sequencing of organisms belonging to various superfamilies of Lepidoptera (Data search from NCBI GenBank database).

Figure 1.10 Number of publications on seri-related databases from the year 2003 to 2015- Continued year (Data shown for 2015-Contd. year as only two databases have been published during 2016-17).

Figure 2.1 Samples collected from different locations of Assam. (A) Eri silkworm (*S. ricini*) (B) Muga silkworm (*A. assama*) (C) Weevils (D) Red ant (E) Som (*P. bombycina*) (F) Soalu (*L. monopetala*) (G) Dighloti (*L. salicifolia*) (H) Payam (*E. fraxinifolia*) (I) Red castor (*R. communis* MCB1 strain) (J) Mejankari (*L. cubeba*) (K-N) Infected *P. bombycina* leaves (O-P) Infected *L. monopetala* leaves (Q) Infected *L. salicifolia* leaf (R-V) Herbarium sheets of the collected host plants.

Figure 2.2 A schematic representation of the workflow for DNA barcoding.

Figure 2.3 Fungus isolates from infected leaves of Muga host plants. (A-D) fungus isolated from *Persea bombycina*/Machilus *bombycina* (E&F) fungus isolated from *Litsea monopetala* (G-K) fungus isolated from *Litsea salicifolia*.

Figure 2.4 Agarose gel electrophoresis of PCR products of isolated DNA of the collected samples. (A) 1.5% agarose gel electrophoresis of PCR product of *S. ricini* DNA sample. Lane 1: 1 Kb DNA ladder, Lane 2 & 3: PCR product of COI gene of *S. ricini* using TY-J-1460F and CI-N-2191R primer set; (B) 1.5% agarose gel electrophoresis of PCR product of *A. assama* DNA sample. Lane 1: 1 Kb ladder, Lane 2: PCR product of COI gene of *A. assama* (Two wells were clubbed together to obtain good yield of DNA for sequencing) using TY-J-1460F and CI-N-2191R primer set; (C) 1.5% agarose gel electrophoresis of amplified COI gene of Red ants and Weevils. Lane 1&2: Amplified COI gene of Red ants

using LCO1490F and HCO2198R primer set, Lane 3, 4 & 7: Empty, Lane 5 & 6: Amplified COI gene of Weevils using same primer set, Lane 8: 100 bp DNA ladder; (D) 1% agarose gel electrophoresis of amplified *rbcL* gene products of *Persea bombycina*. Lane 1: 100 bp Ladder, Lane 2: PCR product of *rbcL* gene of *Persea bombycina* using *rbcLaF* and *rbcLaR* primer set; (E) 1% agarose gel electrophoresis of amplified *matK* gene products of *Persea bombycina*. Lane 1: 1 Kb DNA ladder, Lane 2-6: PCR product of *matK* gene using *matK 2.1F* and *matK 5R* primer set; (F) 1.5% agarose gel electrophoresis of amplified products of *Litsea monopetala*. Lane 1: PCR product of *rbcL* gene of *Litsea monopetala* using *rbcL 1F* and *rbcL 724R* primer set, Lane 2: 100 bp ladder; (G) 1.5% agarose gel electrophoresis of amplified products of *Litsea monopetala*. Lane 1: PCR product of *matK* gene of *Litsea monopetala* using *matK 2.1F* and *matK 5R* primer set, Lane 2: 100 bp ladder; (H) 1.5% agarose gel electrophoresis of amplified products of *Ricinus communis*. Lane 1: PCR product of *rbcL* gene of Red Castor using *rbcL 1F* and *rbcL 724R* primer set, Lane 2: 100 bp ladder; (I) 1.5% agarose gel electrophoresis of amplified products of *Evodia fraxinifolia*. Lane 1: 100 bp DNA ladder, Lane 2-3: PCR product of *matK* gene of *Evodia fraxinifolia* using *matK 2.1F* and *matK 5R* primer set, Lane 4-5: PCR product of *rbcL* gene of *Evodia fraxinifolia* using *rbcL 1F*, *rbcL 724R* primers; (J) 1.5% agarose gel electrophoresis of amplified products of *Litsea salicifolia*. Lane 1: 100 bp DNA ladder, Lane 2-6: PCR products of *matK* gene of *Litsea salicifolia* using *matK 2.1F* and *matK 5R* primer set; (K) 1.5% agarose gel electrophoresis of amplified products of *Litsea salicifolia*. Lane 1: 100 bp DNA ladder, Lane 2-5: PCR products of *rbcL* gene of *Litsea salicifolia* amplified using *rbcL 1F* and *rbcL 724R* primer set; (L) 1.5% agarose gel electrophoresis of amplified products of *Zizyphus mauritiana*. Lane 1: PCR products of *psbA-trnH* gene of *Z. mauritiana* amplified using *psbAF* and *trnHR* primer set, Lane 2:

Empty, Lane 3: 100 bp DNA ladder; (M) 1.5% agarose gel electrophoresis of amplified products of *Litsea citrata*. Lane 1: 100 bp DNA ladder, Lane 2: PCR product of *matK* gene of *Litsea citrata* using *matK* 2.1F and *matK*5R primer set, Lane 3: PCR product of *rbcL* gene of *Litsea citrata* using *rbcL* 1F and *rbcL* 724R primer set; (N) 1.5% agarose gel electrophoresis of amplified products of *Ricinus communis*. Lane 1: 100 bp ladder, Lane 2-4: PCR product of ITS2 region of Red Castor using ITS2F and ITS3R primer set.

Figure 2.5 Agarose gel electrophoresis of PCR products of isolated DNA of fungal and bacterial colonies. (A) 1.5% agarose gel electrophoresis of amplified products of fungal colonies isolated from *L. monopetala* and *L. salicifolia* leaves using ITS1 and ITS4 primer set. Lane 1-2: PCR products of ITS region of fungus isolated from *L. monopetala* leaf, Lane 3, 4, 7, 8, 11, 12, 13, 15, 16, 19: Empty, Lane 5, 6, 9, 10, 14, 17 and 18: PCR products of ITS region of fungus isolated from *L. salicifolia* leaf, Lane 20: 100 bp DNA ladder; (B) 1.5% agarose gel electrophoresis of amplified products of fungal colonies isolated from *P. bombycina* leaves using ITS1 and ITS4 primer set. Lane 1-12: PCR products of ITS region of fungus isolated from *P. bombycina* leaf, Lane 13: 100 bp DNA ladder; (C) 1% agarose gel electrophoresis of amplified products of bacterial colonies isolated from gut of muga silkworm using 8F and 1492R primer set. Lane 1: 1 Kb DNA ladder, Lane 2-9: PCR products of 16S rRNA gene of bacterial colonies.

Figure 2.6 Sequence similarity of COI gene of silkworms. (A) *A. assama* Nongpoh with selected species (B) *S. ricini* Nongpoh with selected species. *Bombyx mori* was taken as outgroup in the study.

Figure 2.7 Sequence similarities of COI gene of weevils and ants with other related species. (A) *Odoiporus longicollis* with other related species taking *Lyctus brunneus* as outgroup (B) *Oecophylla smaragdina* with other related species taking *Eumenis crucifera* as outgroup.

- Figure 2.8** Sequence homology of 16S rRNA gene of *Micrococcus endophyticus* strain DS04 with other related species. *Bacillus* sp. strain AB1 here denotes an outgroup species.
- Figure 2.9** Phylogenetic tree inferred using COI sequences of *Antheraea assama* Nongpoh and *Samia ricini* Borduar with other selected silkworms through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.
- Figure 2.10** Phylogenetic tree inferred using COI sequences of sequenced pests through ML method in Mega 6.0 tool (1000 bootstrap replicates). (A) *Odoiporus longicollis* (weevil) with other selected organisms (B) *Oecophylla smaradina* (red ant) with other selected organisms. Tree was drawn to scale with bootstrap values above the line.
- Figure 2.11** Phylogenetic tree inferred using matK sequences of sequenced host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.
- Figure 2.12** Phylogenetic tree inferred using rbcL sequences of sequenced host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.
- Figure 2.13** Phylogenetic tree inferred using trnH-psbA and ITS2 regions of host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). (A) Tree based on trnH-psbA region of *Z. jujuba* with other selected plants (B) Tree based on ITS2 region of *R. communis* with other selected plants. Tree was drawn to scale with bootstrap values above the line.
- Figure 2.14** Phylogenetic tree inferred using ITS sequences of the sequenced fungal isolates and other selected species through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.
- Figure 2.15** Phylogenetic tree using 16S rRNA gene sequence of sequenced

bacterial isolate through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

Figure 3.1 A pictorial overview of the methodologies used for sequencing and analysis of *A. assamensis* mitogenome.

Figure 3.2 Bioanalyzer profiles after sonication of enriched mitochondrial DNA sample and libraries.

Figure 3.3 *A. assamensis* mitochondrial genome (A) Circular map (B) Characteristic features of the mitogenome. tRNAs are represented as trn followed by the IUPAC-IUB single letter amino acid codes e.g. trnM denote tRNA^{Met}. IGN represents (+) values as intergenic nucleotides and (-) values as overlapping regions.

Figure 3.4 Mitogenome arrangement in selected species. Ancestor here denotes *Thitarodes renzhiensis*.

Figure 3.5 Evolutionary rates (Ka/Ks) of individual PCG of *A. assamensis* with selected species. w.r.t. here denotes with reference to-.

Figure 3.6 G+C content versus Ka/Ks ratios of 13 concatenated PCGs in *A. assamensis* mitogenome.

Figure 3.7 Secondary structures of 22 tRNAs of *A. assamensis*. (A) trnM to trnR (B) trnN to trnV. trnY lacked TΨC loop and trnS1 lacked DHU loop. tRNA here are represented as trn. AA denotes amino acyl arm and AC denotes anticodon arm.

Figure 3.8 (A & B) Variation in the secondary structures of tRNAs in *A. assamensis* and selected Bombycoids. AA denotes Aminoacyl arm and AC denotes Anticodon arm. S-L denotes Stem-Loop of tRNA gene.

Figure 3.9 Multiple sequence alignment of control region of *A. assamensis* mitogenome showing conserved structural elements. Yellow and green highlighted blocks represent 1st and 2nd conserved blocks, respectively.

Figure 3.10 Multiple sequence alignment of spacer tRNA^{Ser2}-nad1 of selected Bombycoids showing a conserved 7 bp motif "ATACTAA".

Figure 3.11 Sequence alignments of relatively large intergenic spacers (IGS)

of *A. assamensis*. (A) Alignment of the IGS located between *tRNA^{Gln}* and *nad2* with its neighboring partial gene *nad2*. (B) Alignment of the IGS located between *cytb* and *tRNA^{Ser2}* with partial neighboring genes *cytb* and *nad1* genes. (C) Alignment of the IGS located between *tRNA^{Ser2}* and *nad1* with partial *cytb*, *nad1* and *rrnS* genes. Asterisks represent the consensus sequences between the IGS and partial genes. Percentage values in () denote the sequence homology between IGS and partial genes.

Figure 3.12 Nucleotide compositions of different regions of mitogenome of *A. assamensis* with respect to other Bombycoids.

Figure 3.13 Codon usage in *A. assamensis* mitogenome. (A) Codon distribution where CDspT = codons per thousand codons (B) Relative synonymous codon usage (RSCU). Termination codons were excluded in the study.

Figure 3.14 The usage rate of five frequently used codons in Bombycoidea members.

Figure 3.15 Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms based on maximum likelihood (ML) in RaxmlGUI v1.3 (1000 bootstrap replicates). The tree is drawn to scale, with bootstrap values indicated above the branches.

Figure 3.16 Phylogenetic tree inferred from 13PCGs+2rRNAs of 34 organisms based on maximum likelihood (ML) in RaxmlGUI v1.3 (1000 bootstrap replicates). The tree is drawn to scale, with bootstrap values indicated above the branches.

Figure 3.17 Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms based on bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

Figure 3.18 Phylogenetic tree inferred from 13PCGs+2rRNAs of 34 organisms based on bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

Figure 4.1 Tapestation profiles of the ladder (A) and prepared libraries (B).

- Figure 4.2** Circular mitochondrial genome map of *C. trifenestrata* with its features. (+ values denote Intergenic spacer while –ve values denote Overlapping spacers)
- Figure 4.3** Size variation in different regions of mitogenome in the selected species (Whole mito: Whole mitogenome, PCGs: Protein coding genes, tRNAs: Transfer RNAs, rRNAs: Ribosomal RNAs; CR: Control region).
- Figure 4.4** Secondary structures of 22 tRNAs of *C. trifenestrata*. (A) trnM to trnR (B) trnN to trnV. trnY lacked TΨC loop and trnS1 lacked DHU loop. tRNA here are represented as trn. AA denotes amino acyl arm and AC denotes anticodon arm.
- Figure 4.5** Conserved regions in the control region of *C. trifenestrata* where **TTAGA** (orange) denotes conserved motif followed by **poly T stretch** (T)₁₉ (location 15123-15141) in green, **ATTTA** (blue) followed by **(TA)₈** in pink denotes microsatellite like element, **(AT)_n** microsatellite (location 15348-15362) and finally **poly A stretch** (A)₁₆ with intermittent T nucleotides (in red) (location 100-120). The highlighted sequence in grey denotes the **repeat element** occurring twice in the region (location 77-112 & 15390-15425).
- Figure 4.6** Multiple sequence alignment depicting a conserved 7 bp motif “ATACTAA” identified in tRNA^{Ser2}-nd1 spacer of *C. trifenestrata* mitogenome similar to other selected Bombycoids. * indicates identical bases.
- Figure 4.7** Frequency of AT and GC content at first, second and third codon positions of PCGs of *C. trifenestrata* mitogenome and other selected organisms.
- Figure 4.8** Codon usage analyses in protein coding genes (PCGs) of *C. trifenestrata* mitogenome. (A) Relative synonymous codon usage (RSCU). Codon families are represented on X-axis and RSCU values on Y-axis. (B) Codon distribution where CDspT = codons per thousand codons. Termination codons were excluded in the study.

- Figure 4.9** Maximum likelihood (ML) phylogenetic tree inferred using PCGs dataset in RaxmlGUI v1.3 among *Cricula* and other selected lepidopterans. The tree was obtained after 1000 bootstrap replicate analysis. Various selected families of Lepidoptera order are represented in different colors. The organism of the present study (*C. trifenestrata*) is highlighted in box.
- Figure 4.10** Maximum likelihood (ML) phylogenetic tree inferred using PCGs+rRNA dataset in RaxmlGUI v1.3 among *Cricula* and other selected lepidopterans. The tree was obtained after 1000 bootstrap replicate analysis. Various selected families of Lepidoptera order are represented in different colors. The organism of the present study (*C. trifenestrata*) is highlighted in box.
- Figure 4.11** Bayesian inference (BI) phylogenetic tree inferred using PCGs dataset in MrBayes v3.2.6 among *Cricula* and other selected lepidopterans. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.
- Figure 4.12** Bayesian inference (BI) phylogenetic tree inferred using PCGs+rRNAs dataset in MrBayes v3.2.6 among *Cricula* and other selected lepidopterans. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.
- Figure 5.1** Timeline of the existing seri-databases from the year 2003 to 2017 generated using respective publication in the literature and database creation year from websites, where (*) indicates database first published in 1999 and its updated versions considered from period 2003-2017; (**) indicates the same database with updated information. (Abbreviations of the web-resources are detailed in **Figure 5.5**).
- Figure 5.2** Percentage of databases with the features (Analysis was based on detailed review on various features of all available seri-related databases).
- Figure 5.3** Workflow for the construction of SeriPort.
- Figure 5.4** Workflow for the construction of Seribase.

Figure 5.5 Schematic representation of Seri-databases classified into four categories- Silkworm Databases- 22 No.s, Host Plant Databases- 25 No.s, Pest and Pathogen Databases- 02 No.s, Combined Databases- 26 No.s.

(**Abbreviations-** **SilkDB:** Silkworm Knowledgebase, **EST DB:** Expressed Sequence Tag Database, **BmMDB:** Bombyx mori Microarray Database, **SilkTransDB:** Silkworm Transcriptome Database, **SilkSatDb:** Silkworm Microsatellite Database, **DBMP:** Database of Bombyx mutant photographs, **BmTEdb:** Transposable elements database for B. mori, **Insect Proteome DB:** Insect Proteome Database, **SilkProt:** Annotated protein database of silkworm, **SilkPPI:** Silkworm Protein- Protein Interaction database, **SilkTF:** Silkworm Transcription Factor Database, **SGRDB:** Silkworm Gene Resources database, **iPathDB:** Insect Pathway Database, **MorusDB:** Morus Genome Database, **MulSatDB:** Mulberry Microsatellite Database, **CastorDB:** A comprehensive knowledgebase database for R. communis, **Papaya-DB:** Papaya Genomic Resources Online, **CPR-DB:** Papaya Repeat Database, **CGDB:** Cassava Genome Database, **CCDB:** Chinese Cassava Genome Database, **HOSTS:** a Database of the World's Lepidopteran Host plants, **PlantGDB:** Resources for Comparative Plant Genomics, **ChromDB:** The Chromatin Database, **PlantTFDB:** Plant Transcription Factor Database, **SilkPathDB:** Silkworm Pathogen Database, **SilkDis:** Silkworm Disease Database, **BOLD:** Barcode of Life Data System, **DBIF:** Database of Insects and their Food Plants, **UCD:** Universal Chalcidoidea Database, **DBMW:** Database of Butterflies and Moth of the World, **CNIDB:** Common Names of Insects Databases, **BAMONA:** Butterflies and moths of North America, **EOL:** Encyclopedia of Life, **ITIS:** Integrated taxonomic information system, **INPN:** Inventaire National du Patrimoine Naturel, **ACG:** Area de Conservación Guanacaste, **SRDB:** Spatio-temporal database of Silk Road, **SFSDB:** Silk Fabric Specification

Database, **SILKS**: Sericulture Information Linkages and Knowledge System, **miRNEST**: An integrative microRNA resource, **miRBase**: The microRNA database, **MEROPS**: the peptidase database).

Figure 5.6 Data content in Seribase.

Figure 5.7 Schematic of database design for SeriPort.

Figure 5.8 Schematic of database design for Seribase.

Figure 5.9 Home Page of SeriPort.

Figure 5.10 Home Page of Seribase.

Figure 5.11 Search option in SeriPort and snapshots of search results based on keyword “Antheraea”.

Figure 5.12 Browse Page of Seribase.

Figure 5.13 Search Page of Seribase and search results based on keyword “Antheraea pernyi”.



List of tables

- Table 1.1** *Diversity of sericigenous insects recorded in North eastern region of India.*
- Table 1.2** *List of host plants of few silkworms from North East India.*
- Table 1.3** *List of silkworm databases and their data contents.*
- Table 1.4** *List of silkworm host plant databases and their data contents.*
- Table 2.1** *Collection of different silkworm resources and their corresponding location.*
- Table 2.2** *List of primers used in the study.*
- Table 2.3** *List of organisms retrieved from NCBI GenBank for phylogenetic analysis. (A) Organisms selected for phylogenetic analysis of silkworms based on COI gene. (B) Organisms selected for phylogenetic analysis of weevils based on COI gene. (C) Organisms selected for phylogenetic analysis of red ant based on COI gene. (D) Organisms selected for phylogenetic analysis of host plants based on ITS2 region. (E) Organisms selected for phylogenetic analysis of host plants based on matK region. (F) Organisms selected for phylogenetic analysis of host plants based on rbcL gene. (G) Organisms selected for phylogenetic analysis of host plants based on trnH-psbA region. (H) Organisms selected for phylogenetic analysis of bacteria based on 16S rRNA gene. (I) Organisms selected for phylogenetic analysis of fungi based on ITS region.*
- Table 2.4** *Models selected for various datasets under study.*
- Table 2.5** *Optimized PCR cycling conditions for each sample.*
- Table 2.6** *BLAST similarity results for the sequenced species in the present study.*
- Table 2.7** *Barcode sequences submitted in NCBI GenBank database.*
- Table 3.1** *List of species considered for comparative mitogenome study.*
-

(Highlighted ones are the organisms of Bombycoidea superfamily used for comparative mitogenome analysis with respect to *A. assamensis*).

Table 3.2 Comparative gene lengths (in bp) in Bombycoids.

Table 3.3 Initiation and termination codons of PCGs in *A. assamensis* with respect to selected members of Bombycoidea superfamily.

Table 3.4 Gene by gene divergences in the protein coding genes at different taxonomic levels, both at nucleotide level and amino acid levels (in percentage). C here denotes conserved sites, PI denotes phylogenetically informative sites, V denotes variable sites, N denotes nucleotide level, A denotes amino acid level. CVI percentage was calculated as $\text{No. of sites} / \text{total gene length} \times 100$.

Table 3.5 Nucleotide variation pattern in terms of transitions (ts) and transversions (tv) ratio among PCGs of *A. assamensis* and other selected members.

Table 3.6 Anticodons (AC) of tRNAs and number of mismatches present in Amino-acyl (AA) stem, T Ψ C-stem, DHU-stem and Anticodon (AC) stem of mitochondrial tRNAs of *A. assamensis* with respect to the selected Bombycoids.

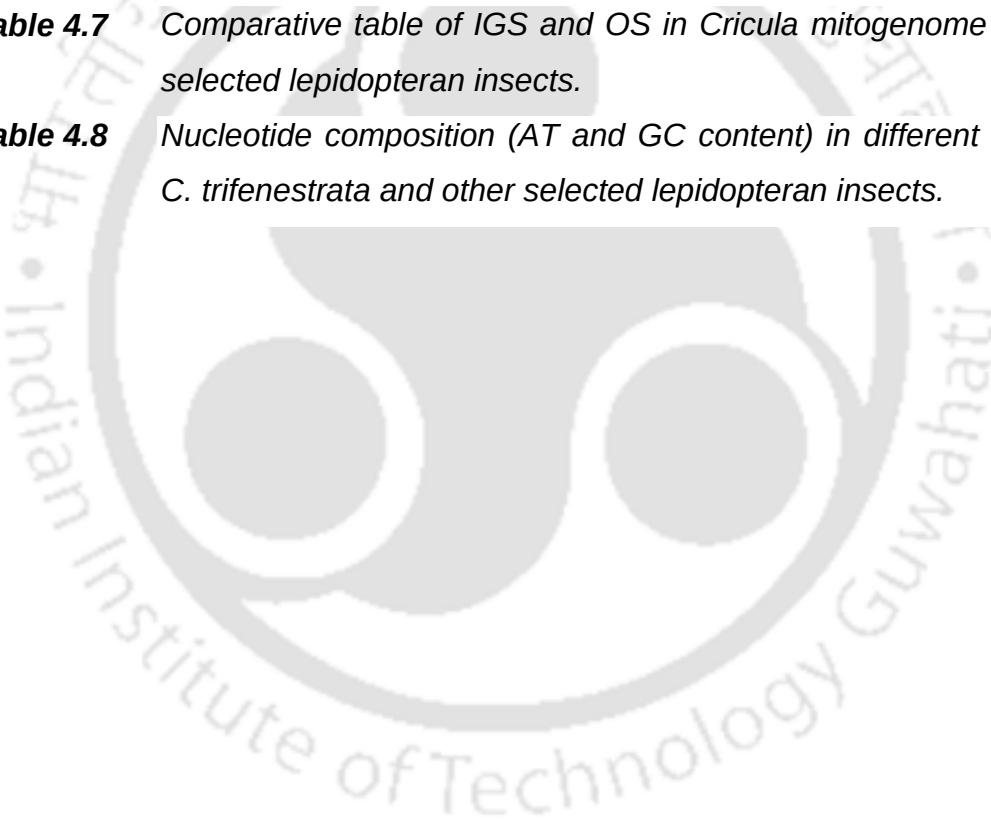
Table 3.7 Intergenic spacers (IGS) and overlapping sequence (OS) between different genes in *A. assamensis* with respect to the selected Bombycoid species.

Table 3.8 A+T content of individual PCGs in Bombycoids.

Table 3.9 Base composition at each codon position of the concatenated 13 protein-coding genes in Bombycoidea superfamily members.

Table 3.10 AT skewness and GC skewness in the whole mitogenome sequence, protein coding genes (PCGs), tRNAs, rRNAs and control region of *A. assamensis* and the selected Bombycoid species.

Table 4.1 List of the organisms with completely sequenced mitogenome belonging to the order Lepidoptera and the outgroup orders Trichoptera, Diptera and Hymenoptera. (Highlighted ones represent the selected organisms used for comparative mitogenome analysis with *C. trifenestrata*).

-
- Table 4.2** Initiation codons of 13 protein coding genes in the mitogenome of *C. trifenestrata* and selected lepidopteran insects.
- Table 4.3** Termination codons in the protein coding genes of *C. trifenestrata* mitogenome and other selected lepidopteran insects.
- Table 4.4** Gene-by-gene divergence in protein coding genes of *C. trifenestrata* with respect to other selected organisms at amino acid level. Conserved sites (C), variable sites (V) and phylogenetically informative sites (PI) represented in percentage of total gene length.
- Table 4.5** Ka/Ks ratio of individual protein coding genes (PCGs) of *C. trifenestrata* with respect to the members of different families.
- Table 4.6** Number of mismatched pairs in the tRNAs of *C. trifenestrata*.
- Table 4.7** Comparative table of IGS and OS in *Cricula* mitogenome and other selected lepidopteran insects.
- Table 4.8** Nucleotide composition (AT and GC content) in different regions of *C. trifenestrata* and other selected lepidopteran insects.
- 

Chapter 1

Introduction and Review of Literature

CHAPTER 1

Introduction and Review of literature

1.1 Seribiodiversity

Sericulture is an agro-based cottage industry with immense employment potential. It earns foreign exchange (>2000 crores per year) and India exports silk items to over 50 countries. Globally India, the second largest producer of silk after China, contributes about 15.5% to the total world raw silk production. It generates employment to 8 million rural and semi-urban people covering silkworm seed producers, farmers-cum-rearers, reelers, twistors, weavers, spinners of silk waste, traders, etc. This is mainly because of availability of commercially exploited silk in India as it is a habitat for silkworm biodiversity. Diverse cultural and climatic conditions of India also support the increased usage of silkworms in India (Jay Prakash, 2011; <http://www.inserco.org/en/statistics>).

Among the two biodiversity hot spots of India - The Eastern Himalayas and The Western Ghats, the former is the best hot spot of seri-biodiversity. North-East (NE) India is endowed with rich seri-biodiversity which has great impact in its economic growth and culture (Thangavelu, 1991). Seri biodiversity refers to the variability in seri-bioresources which involve a vast array of ecologically and economically important silkworms and their host plants (Srivastava & Thangavelu, 2005). These silkworms interact with other partners like host plants, pest, pathogens and parasites forming a seri-ecosystem (**Figure 1.1**).



Figure 1.1 Interactions between silkworm, host plants, pest/parasitoid/competitor and pathogens in a seri-ecosystem.

All five kinds of commercially exploited silks- mulberry, eri, muga, tasar and oak tasar are produced in this region by *Bombyx mori*, *Samia ricini* (formerly known as *Philosamia ricini*), *Antheraea assamensis* (or *Antheraea assama*), *Antheraea mylitta* and *Antheraea pernyi*, respectively (Peigler & Naumann, 2003). Nassig *et al.*, 1996 have reported around 1200-1500 wild species worldwide, out of which 50 species are still unexplored and undocumented which have been suggested to be lying in Indian sub-continent from Himalayas to Sri Lanka. Hence, the study on exploration and documentation of silkworms is needed to preserve seribiobiodiversity and to boost the country's economy. Sericulture is a part of tradition and culture for the people of Assam where muga and eri culture have imparted significant contribution towards employment generation, thus improving socio-economic status of this region (De & Das, 2010). The pupal stages of non-mulberry and mulberry silkworm pupae are also nutritive source for many communities and malnourished population of Assam (Mishra *et al.*, 2003).

Silkworms can be categorized into three groups namely domesticated (e.g. *B. mori*, *S. ricini*), semi-domesticated (e.g. *A. assamensis*) and wild (*A. mylitta*, *A. pernyi*, *Cricula trifenstrata* and many more) based on their cultivation method or growth (Kalita & Dutta, 2014). Silkworms belong to Lepidoptera order which includes wide varieties of butterflies and moths. Non-mulberry silkworms are wild or semi-domesticated 'charismatic fauna' of the insect world that belong to family Saturniidae of the order Lepidoptera. The wild silkworms show great diversity in life history with characteristically different physiological, morphological and feeding behavior. Few reports available on sericobiodiversity of NE India suggesting its potential as a source of natural silk in Indian continent are given in **Table 1.1**.

Table 1.1 Diversity of sericigenous insects recorded in North eastern region of India.

S. No.	No. of wild sericigenous species reported	Distribution	Reference
1	19 species	Mainly Sikkim and Assam	Seitz, 1933
2	40 wild species	India	Arora & Gupta, 1979
3	9 sericigenous insects	NE region	Thangavelu, 1991
4	24 Saturniidae species	NE India	Singh & Chakravorty, 2006
5	14 wild species (8 genera) and their host plants	Nagaland	Kakati & Chutia, 2009
6	15 sericigenous insects (7 genera, 2 families)	Manipur	Devi <i>et al.</i> , 2011
7	12 sericigenous species (8 genera, 2 families)	Assam	Kalita & Dutta, 2014
8	14 wild sericigenous species (9 genera)	Khasi Hills of Meghalaya	Shangpliang & Hajong, 2015
9	6 wild sericigenous species (5 genera)	Assam	Saikia <i>et al.</i> , 2017

Muga, *A. assamensis* (Helfer) is a semi-domesticated multivoltine silkmoth and producer of precious golden silk endemic to Assam province (Unni *et al.*, 2009; Tikader *et al.*, 2011). The muga silk had received a geographical indication (GI) tag in 2007 and GI logo in 2012 due to its unique golden silk texture and durability.

The eri silkmoth (*S. ricini*) produces white colored silk known as “Ahimsa silk” as no pupa are sacrificed during its production. *Attacus atlas* (also known as Atlas Moth, Deo Muga, Kotkari muga) produces durable brownish colored silk. Another wild silkmoth available in Assam is *Cricula trifenestrata* Helfer (*Amphutukoni muga*) which produces perforated cocoon and biocompatible golden yellow colored silk (Ahmed & Rajan, 2011).

The quality and yield of silk produced by the silkworms primarily depends upon the varieties and condition of host plants which have great diversity in their morphology and leaf content. For instance, muga silkworm produces very attractive and several times costlier silk when fed on the host plant Mejankari, in comparison with other host plants (Tikader *et al.*, 2013). Hence the study of host plants is equally important as silkworms. These plants may be categorized as primary, secondary and tertiary host plants based on the preference by silkworms as listed in **Table 1.2**.

Table 1.2 List of host plants of few silkworms from North East India.

Silkworm	Common Name	Host Plants	Reference
Antheraea assama	Muga silkworm (Semi-domesticated)	<i>Persea bombycina</i> or <i>Machilus bombycina</i> , <i>Litsea monopetala</i> (1°), <i>Litsea salicifolia</i> , <i>Litsea citrata</i> , <i>Cinnamomum camphora</i> , <i>Cinnamomum tamala</i> (2°), <i>Actinodaphne augustifolia</i> , <i>Actinodaphne obovata</i> , <i>Celastrus monosperma</i> , <i>Cinnamomum cecicodaphne</i> , <i>Cinnamomum glanduliferum</i> , <i>Cinnamomum obtusifolium</i> , <i>Gmelina arborea</i> , <i>Zizyphus jujuba</i> , <i>Machilus odoritissima</i> , <i>Magnolia pterocarpa</i> , <i>Michelia oblonga</i> , <i>Michelia champaca</i> , <i>Symplocos oxyphylla</i> , <i>Symplocas ramosissima</i> , <i>Symplocas paniculata</i> , <i>Zanthoxylum rhesta</i> , <i>Litsea nitida</i> , etc. (3°)	Neog et al., 2005; Ahmed & Rajan, 2011; Kalita & Dutta, 2014; Saikia et al., 2017
Samia ricini	Eri silkworm (Domesticated)	<i>Ricinus communis</i> , <i>Heteropanax fragrans</i> (1°), <i>Evodia fraxinifolia</i> , <i>Manihot utlitissimn</i> (2°), <i>Ailanthus excelsa</i> , <i>Ailanthus glandulosa</i> , <i>Ailanthus grandis</i> , <i>Ailanthus tryphysa</i> , <i>Carica papaya</i> , <i>Jatropha curcas</i> , <i>Jatropha multifida</i> , <i>Plumeria acutifolia</i> , <i>Xanthoxylum rhesta</i> , etc. (3°)	Ahmed & Rajan, 2011; Kalita & Dutta, 2014
Antheraea mylitta	Tropical tasar (Wild)	<i>Terminalia arjuna</i> , <i>Terminalia tomentosa</i> , <i>Shorea robusta</i> , <i>Anogeissus latifolia</i> , <i>Hardwickia binata</i> , <i>Lagerstroemia parviflora</i> , <i>Zizyphus jujuba</i>	Ahmed & Rajan, 2011; Kalita & Dutta, 2014
Cricula trifenestrata	Amphutukani muga (Wild)	<i>Persea bombycina</i> , <i>Litsea cubeba</i> , <i>Mangifera indica</i> , <i>Psidium guajava</i> , <i>Betula alnoides</i>	Ahmed & Rajan, 2011; Kalita & Dutta, 2014; Saikia et al., 2017
Antheraea proylei	Temperate tasar (Wild)	<i>Lithocarpus dealbata</i> , <i>Quercus acutissima</i> , <i>Quercus floribunda</i> , <i>Quercus griffithii</i> , <i>Quercus leucotrichophora</i> , <i>Quercus semicarpifolia</i> , <i>Quercus semiserrata</i>	Ahmed & Rajan, 2011

Silkworm	Common Name	Host Plants	Reference
<i>Actias maenas</i>	Moon moth (Wild)	<i>Azadirachta indica</i> , <i>Terminalia arjuna</i>	Saikia <i>et al.</i> , 2017
<i>Attacus atlas</i>	Atlas Moth, Deo Muga, Kotkari muga (Wild)	<i>Litsea salicifolia</i> , <i>Dillenia indica</i> , <i>Maesa indica</i>	Ahmed & Rajan, 2011; Kalita & Dutta, 2014; Saikia <i>et al.</i> , 2017
<i>Actias selene</i>	(Wild)	<i>Azadirachta indica</i> , <i>Terminalia arjuna</i> , <i>Persea bombycina</i> , <i>Evodia fraxinifolia</i>	Kalita & Dutta, 2014; Saikia <i>et al.</i> , 2017
<i>Antheraea frithi</i>	(Wild)	<i>Shorea robusta</i>	Kalita & Dutta, 2014

Where, (1°), (2°) and (3°) denote primary, secondary and tertiary host plants, respectively.

Other than nutritional source of silkworms, host plants also have economic importance such as medicinal value, industrial value, etc. (Bhuinya *et al.*, 2010; Lalhmingliani *et al.*, 2015). For instance, fruits and oils of *Litsea cubeba* can be used for treatment of headache, stomach ache, dizziness, hysteria and memory loss (Lalhmingliani *et al.*, 2015). *Jatropha* is more popular as a biofuel crop and most of the molecular and genetic research has been focused on this aspect. Similarly, *Ricinus communis* is known for the production of castor oil having applications as lubricant, food, medicine, etc. (Scarpa & Guerici, 1982).

The productivity of silk is being affected by the attack of various disease pathogens, pests and predators on either silkworms or their host plants (Choudhury, 1981; Rath *et al.*, 2001; Das *et al.*, 2003). For instance, on an average 40% loss was observed by some common diseases of silkworms such as Grasserie, Flacherie, Pebrine and Muscardine (green, white, black, etc.) which are caused by virus, bacteria, protozoa and fungi, respectively. High incidence of Grasserie and Flacherie diseases occurs in summer causing 20-30% of silkworm crop loss. Infestation of Muscardine disease mainly occurs in winters that may result in 25-35% of losses (Selvakumar & Datta, 2013; Saikia *et al.*, 2016). Pebrine, a deadly disease common in Assam, is caused by *Nosema* species and infects majority of the silkworms or even kills the whole culture (Rath *et al.*, 2001). In addition to diseases, several pests and predators like white flies, bugs, ants, termites, beetles, weevils, scale insects, ants, wasps, hornets, praying mantids, birds, owls, bats, monkeys, etc. contribute to productivity loss (Veer *et al.*, 1996; Subharani & Jayaprakash, 2015). For example, a serious endo parasitoid, uzi fly (*Exorista bombycis*), has been reported to result in great loss (up to 90%) by

attacking the silkworms such as *A. assamensis*, *A. mylitta*, etc. (Singh *et al.*, 1993; Rath & Sinha, 2005).

Similarly, several diseases have been reported for host plants of silkworms. Few of these include anthracnose, rust disease, stem rot, leaf spot, leaf blight, etc. caused by various fungal or bacterial pathogens (Das *et al.*, 2003; Paliwal & Paliwal, 2015). Along with diseases, various pests and insects also cause loss to silk productivity. For instance, *Cricula trifenestrata* is a competitor of muga silkworm and considered as gregarious pest of primary host plant of muga (*P. bombycina* or Som) thus resulting in insufficient leaves for muga silkworm rearing. This in turn affects the survival rate of silkworms and silk production rate (Negi & Sengupta, 1993; Ahmed *et al.*, 2012).

Many important genetic resources of sericigenous insects are declining due to various factors like large scale deforestation, threat from various pest and predators, soil and air pollution by chemical insecticides, biopesticides and adverse climatic conditions (Devi *et al.*, 2011). The disease pathogens are major obstacle to sericulture which need proper identification for easy detection and management of diseases. Hence, proper conservation and utilization of important genetic resources of wild silk moths and their host plants require immediate attention for commercial exploitation for higher production of silk. Further, despite of such huge diversity, biodiversity of sericigenous insects particularly wild silkworms, their host plants, their pests and pathogens of north-eastern India including Assam is not yet fully understood (Peigler & Naumann, 2003). Exploration, documentation and characterization of wild silk varieties and their host plants is thus important for laying the foundation for increasing productivity.

To accomplish this task, systematics and taxonomy are the prerequisites to effectively describe and understand the world's biodiversity which is quite challenging (Packer *et al.*, 2009). Traditional systematic based taxonomy suffers from complications in proper identification of organisms with phenotypic plasticity, genetic variability, immaturity and crypticity. Advanced alternative methods (DNA sequence-based analyses) have therefore emerged as essential tools for correct identification and to establish the field of taxonomy (Padial & De La Riva, 2007; Padial *et al.*, 2009).

1.2 Molecular characterization using DNA Barcoding

DNA barcoding is a molecular and bioinformatical tool first proposed by Paul Hebert (2003) to aid in diagnosis of biodiversity. However, it was not a novel concept. Earlier it had been used in the discrimination of microbial species and later been tested among various taxa (Eggert *et al.*, 2002; Floyd *et al.*, 2002). DNA barcoding is a rapid, cost-effective and reliable species identification tool that uses short, standardized and conserved gene regions known as DNA barcodes. Since every species most likely have a unique DNA signature code, this approach is based on the postulate that inter-species variation exceeds intra-species variation (Hebert *et al.*, 2003). It performs species level identification by comparing and assessing the degree of similarity between barcode sequences derived from unknown specimens and a set of reference taxa. The major advantage of DNA barcoding is the identification of organisms which are even cryptic in nature, microscopic in size and inaccessible or complex in morphology (Hebert *et al.*, 2003). It is effectively being used for the identification of species in animals, plants, microorganisms, protists, etc. (Purty & Chatterjee, 2016). The recognized

DNA barcode should have following salient features as explained by Taberlet *et al.*, (2007)-

- (i) The barcode region should exhibit variation between species but nearly identical within species.
- (ii) It should provide enough phylogenetic information for easy identification and classification of unknown species.
- (iii) It should be highly robust and consistent for DNA amplification and sequencing.
- (iv) It should be short in length to amplify even the degraded DNA.

DNA barcodes can be useful in many fields such as ecology, biomedicine, epidemiology, evolutionary biology, biogeography, conservation biology, bio-industry and biomonitoring (Ball & Armstrong, 2006; Parveen *et al.*, 2016). Selection of a suitable barcode region is an important prerequisite for DNA barcoding that varies amongst various taxa. Several mitochondrial, chloroplast and nuclear regions have been studied for their efficiency to fulfill the requirements of barcoding of diverse biosphere.

1.2.1 DNA barcoding in insects and animals

Insects are the most diverse among various live forms on earth distributed worldwide and have very complex evolutionary history ranging from agricultural pests, disease agents or vectors, predators, parasites and parasitoids of other insects, model organisms for research, etc. (Speight *et al.*, 2005; Price *et al.*, 2011).

DNA barcoding of animals including insects is being done by sequencing specific gene present in mitochondria. The mitochondrial genome (mitogenome or mtgenome) of animals including insects is devoid of introns, maternally inherited,

and, has limited exposure to recombination, high mutation rate and haploid type of inheritance. Therefore, it is a better candidate for analysis than the nuclear genome (Saccone *et al.*, 1999). The above properties allow universal primer development which enables routine amplification of specific regions of the mitogenome. A 650-base fragment of the 5' end of the mitochondrial *cox1* gene (cytochrome c oxidase I or COI) has been considered as gold standard barcode region for animals including insects (Folmer *et al.*, 1994; Simmons & Weller, 2001; Hebert *et al.*, 2003). This is due to lack of insertions or deletion mutations, more conserved region of sequence and a longer range of phylogenetic signal. Cytochrome c oxidase I is one of the constitute subunits of respiratory complex IV enzyme which participates in oxidative phosphorylation taking place on electron transport chain of mitochondria. The structural position of *cox1* gene in mitogenome of insect is shown in **Figure 1.2**. DNA barcoding has been widely applied for differentiating various insect species belonging to orders- Lepidoptera, Coleoptera, Diptera, Hemiptera, Hymenoptera, etc. (Park *et al.*, 2011; Jinbo *et al.*, 2011; Jalali *et al.*, 2015; Halim *et al.*, 2017). These include different developmental life stages of mostly silkworms, pests, parasites and parasitoids. For instance, *cox1* gene was used to study phylogenetic relationships among *Bombyx* species (*B. mori* and *B. mandarina*) and *Antheraea* species (*A. yamamai* and *A. pernyi*) (Hwang *et al.*, 1999). Similarly, barcoding was also found to be effective in identifying different life stages of the Lepidoptera and sex (males/females) of dimorphic species (Hajibabaei *et al.*, 2006; Floyd *et al.*, 2009).

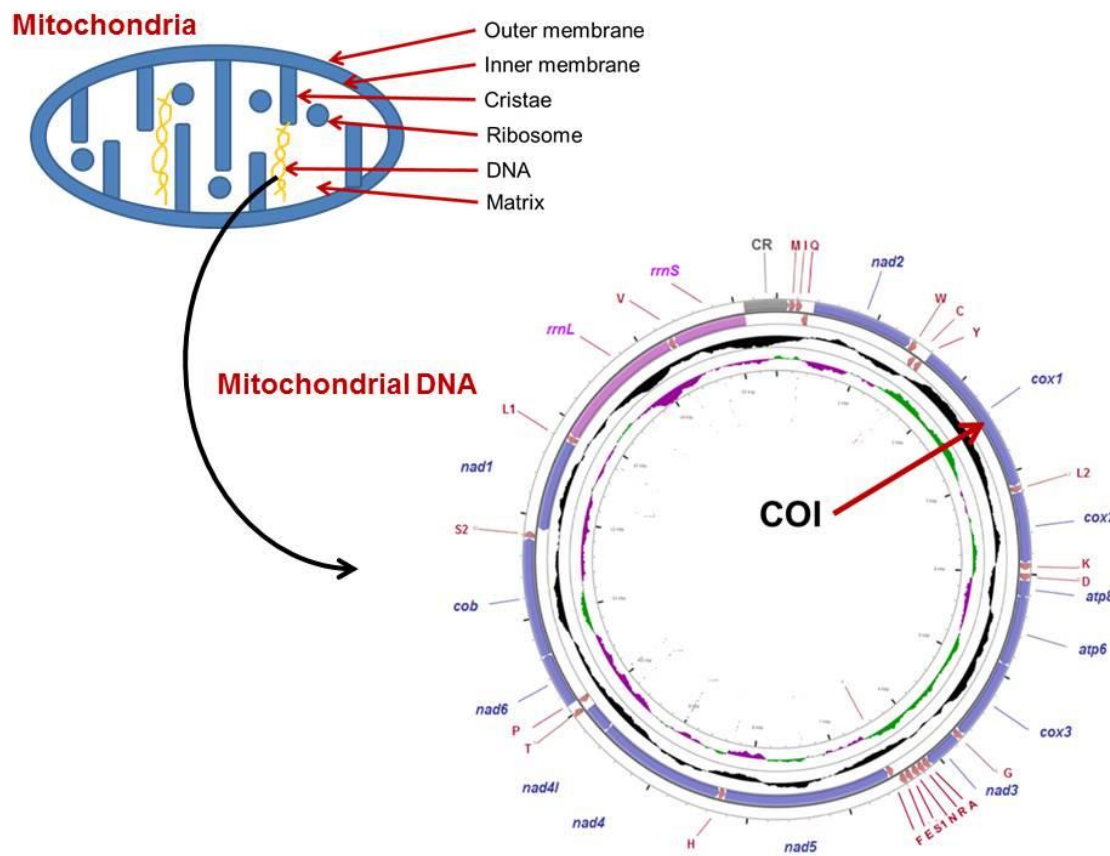


Figure 1.2 COI gene of mitochondrial genome in insects.

The diversity of ants, termites, aphids, weevils, bugs, etc. has also been assessed through DNA barcoding which can be effective in developing pest management strategies (Hirsch *et al.*, 2010; Rebijith *et al.*, 2013). For example, Ojha *et al.*, 2014 successfully discriminated 16 ant species belonging to 4 subfamilies (Formicinae, Myrmicinae, Ponerinae and Dolichoderinae) collected from various locations of Karnataka. Similarly, studies reported the genetic divergence and population variant studies of weevils like *Copturus aguacatae* (avocado stem weevil) in Mexico and *Anthonomus grandis* (boll weevil) in North America belonging to family Curculionidae (Order: Coleoptera) through DNA barcoding (Engstrand *et al.*, 2010; Barr *et al.*, 2013).

There are few reports available which used other genes along with *cox1* gene for the identification of pests and insect species. For instance, Arunkumar *et al.*, 2006

proposed the use of mitochondrial genes including 12S *rRNA*, *cox1* and control region (CR) to resolve phylogenetic relationships among members of Bombycoidea superfamily. Similarly, the accuracy of different regions like 18S and 28S nuclear genes as complementary barcodes to mitochondrial *cox1* gene was studied across 10 scale insect families (Hemiptera: Coccoidea) which were mostly cryptic. The study showed that identification success rate was enhanced to 91.5% from 72.7% when combined gene set (*cox1* and 28S *rRNA*) was used (Sethusa *et al.*, 2014).

1.2.2 DNA Barcoding in plants

Finding a universal, effective and robust barcode region in plants is always challenging due to their high genome diversity. Mitochondrial *cox1* gene is not primarily used for species identification due to low substitution rate of mitochondrial genes than plastid or nuclear gene in plants (Palmer & Herbon, 1988; Yang *et al.*, 1999). Therefore, the nuclear and plastid/chloroplast genomes could provide suitable barcode regions for plants as they are variable enough to discriminate amongst species and have less intraspecific variation than interspecific variation (Chase *et al.*, 2005; Kress *et al.*, 2005).

Various molecular markers have been tested in a wide variety of plants to find a potential barcode for discriminating plant species. These include coding genes such as *rpoB*, *rpoC1*, *matK*, *rbcL*, *UPA* and non-coding regions such as *trnH-psbA*, *atpF-atpH*, *trnL-trnF*, *psbK-psbI*, ITS2, ITS (Kress *et al.*, 2005; Kress & Erickson, 2007; Pennisi, 2007; Wang *et al.*, 2010). While barcodes ITS2 & ITS are present in nuclear genome of plant, rests are chloroplast markers. The *rbcL* (ribulose-bisphosphate carboxylase) and *matK* (maturase K) barcodes include a 599 bp fragment located at 5' end of the gene (1-599 bp) and a 841 bp region

located at the center of the gene (205–1046 bp) in the complete plastid genome sequence of *Arabidopsis thaliana* (gi 7525012:54958–56397) and *A. thaliana* (gi 7525012:2056–3570), respectively (Hollingsworth *et al.*, 2011). The *rbcL* gene is the most characterized plastid coding region in GenBank which has been evaluated for molecular identification of plants due to its universality, ease of amplification and alignment. It is a broad spectrum candidate for all major groups and provides a good baseline for barcoding of land plants as compared to other plastid genes (Newmaster *et al.*, 2006). Similarly, the plastid gene, *matK* has also been proposed as a barcode marker for land plants. It has higher substitution rates than *rbcL* in angiosperms/flowering plants but not in all bryophytes and ferns (Lahaye *et al.*, 2008; Yu *et al.*, 2011). Further, due to high level of variation in length at inter-species level, the spacer between *trnH* and *psbA* genes (119 to 1000 bp length) can be best suited for discriminating closely related organisms accurately (Bruni *et al.*, 2012). Amongst all the nuclear regions, internal transcribed spacer region (ITS) could be used as potential barcode candidates in plant phylogenetic analysis. It comprises of ITS1 and ITS2 regions, distinguished by the 5.8S gene and is covered by the small subunit or 18S (SSU) and large subunit or 28S (LSU) genes in the nuclear ribosomal DNA (rDNA) repeat units (**Figure 1.3**).

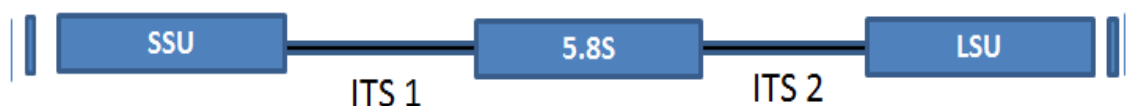


Figure 1.3 Map of nuclear *rRNA* genes and their ITS regions in plants.

Although presence of paralogous copies of ITS has been observed as a problem in some genera (Campbell *et al.*, 2005), ITS2 region has been proved to be the most promising universal barcode candidate for identification of medicinal plants

(Selaginellaceae family) (Chen *et al.*, 2010; Gu *et al.*, 2013). This is due to several distinct features like availability of conserved regions to design universal primers, ease of amplification and significant variability at species or lower levels (Chen *et al.*, 2010). ITS2 has been found to be the most potential markers for assessment of diversity and species identification of Euphorbiaceae, Rutaceae, Lauraceae, Moraceae and Brassicaceae families (Pang *et al.*, 2010; Luo *et al.*, 2010; Sun *et al.*, 2015; Zeng *et al.*, 2015; Liu *et al.*, 2017).

However, no single plant marker is as efficient as *COI* observed in animals and thus combination of different markers (*trnH-psbA*, *matK*, *rpoC*, *rpoB*, *rbcL*) is being employed for barcoding of plants. Moreover, the level of barcode performance varied with plant taxa (Consortium for the Barcode of Life-Plant Working Group (CBOL-PWG). In 2009, CBOL-PWG recommended a two-locus standard barcode-*rbcL* and *matK*, with 70% discriminatory power, for barcoding of plant species based on two criteria: efficient recovery of good-quality sequences and high species discriminatory potential (Hollingsworth *et al.*, 2011). In addition, *trnH-psbA* and ITS regions were suggested as supplementary barcodes for plant DNA barcoding (Ren *et al.*, 2010). Multi-locus barcodes (*matK+rbcL+trnH-psbA*, *rpoC1+rpoB+matK*, *rpoC1+matK+trnH-psbA*, etc.) are also being employed for efficient identification of plant species (Kress *et al.*, 2009; Heise *et al.*, 2015; Yang *et al.*, 2017).

Thus, the barcode markers (either single or in the combination) showed variable discriminatory power in different plant taxa and can be used for the identification efficiently. Similar to silkworms and their pests, the barcoding studies on silkworm host plants are still limited and thus need further attention to explore their diversity.

1.2.3 DNA Barcoding in bacteria

Apart from silkworms, pests and their host plants, bacteria are another component of seri-ecosystem which may act both as useful or harmful organisms. It has been reported that gut microbiota plays a pivotal role in the growth and development of silkworms as well as improving silk production capacity (Anand *et al.*, 2010). Apart from gut microbiota, studies on disease causing bacteria pathogens are important which contribute towards losses in sericulture.

16S *rRNA* gene is used as a marker for rapid and efficient assessment of bacterial diversity and phylogenetic studies. This is attributed to several advantages like non-variable function, ~1500 bp size sufficient for informatics work and presence in almost all bacterial species (Janda & Abbott, 2007; Woo *et al.*, 2008). This gene has been employed in studying diversity of bacterial communities from the gut of insects belonging to different orders such as Lepidoptera, Coleoptera, Diptera, Hymenoptera, etc. (Gusmao *et al.*, 2007; Kelley & Dobler, 2011; Chen *et al.*, 2016; Iskender *et al.*, 2017). Studies reported the identification of cellulolytic, amylase, protease and lipase producing bacteria in *B. mori* silkworms using 16S *rRNA* gene which can be exploited for probiotic preparation in order to enhance both quality and yield of silkworms (Feng *et al.*, 2011; Wang *et al.*, 2016). Similarly, this gene has also been employed in the identification of bacterial pathogens in silkworms. For instance, flacherie causing bacterial pathogens were identified in *B. mori* based on 16S *rRNA* gene sequencing (Swathi *et al.*, 2015).

Apart from 16S *rRNA* gene, other markers (conserved protein coding genes) including *rpoB* (beta subunit of DNA polymerase), *gyrB* (structural gene for DNA gyrase β subunit), *cpn60* (chaperonin-60) and *cox1* have also been suggested to be suitable and effective for bacterial DNA barcode development process (Case *et*

al., 2007; Smith *et al.*, 2012; Links *et al.*, 2012). However, utility of these markers in the identification of bacteria from either gut or infected silkworms is yet to be explored.

1.2.4 DNA Barcoding in fungus

The occurrence of fungal diseases has been studied based on mainly morphological characteristics that identify fungal strains in silkworms and their host plants (Sen *et al.*, 1970; Das *et al.*, 2010). Morphological systems alone are not reliable due to misidentification in many cases; therefore molecular systems have been employed to study fungal diversity.

Use of standard *COI* barcode has also been excluded in fungi due to its difficulty in amplification and insufficient variation. ITS, a multi-copy and easily amplified region of the ribosomal DNA has been considered as the most ideal barcode for the identification and characterization of most of the fungal taxa at species or lower level (Hillis & Dixon, 1991; Schoch *et al.*, 2012). For instance, Chowdappa *et al.*, 2009 reported molecular characterization of *Colletotrichum* species that causes anthracnose in cereals, vegetables, grasses, cashew, fruit crops like mango, grapes, etc. using ITS barcode. Both ribosomal subunits *SSU* and *LSU* have also been employed for the identification of novel fungal clade alone or in combination with ITS region (Arfi *et al.*, 2012). Further, phylogenetic analysis using *SSU* was found to be successful in the identification of *Mucor* sp. (cellulolytic fungus) isolated from larval gut of *B. mori* silkworms (Ma *et al.*, 2014). However, ITS shows low variation and lacks species resolution, especially in the economically important microfungal groups in Ascomycota division, which constitutes 60% of the presently recognized fungal taxa (Seifert, 2009). Therefore, several protein coding genes e.g. DNA-dependent RNA polymerase II largest

subunit (*rpb1*), DNA-dependent RNA polymerase II second largest subunit (*rpb2*), β -tubulin (*tub2*) and translation elongation factor- 1 α (*tef1*) have been proposed for species identification and phylogenetic analyses in fungi particularly in Ascomycota division. They are more informative providing better species level resolution in most of the fungi (Frisvad *et al.*, 2004; Schoch *et al.*, 2009; O'Donnell *et al.*, 2010).

Thus for simultaneous species identification and phylogenetic studies in fungal taxa, multigene approach including ITS supplemented with additional sequences such *LSU* or *SSU* and use of protein coding genes can be effective as barcoding markers (Schoch *et al.*, 2012; Imoulan *et al.*, 2016). Most of the research has been done in barcoding different fungal species; however, the utility of barcode markers in exploring the fungal diversity in sericulture field is limited.

1.2.5 DNA Barcoding in virus and microsporidians

Apart from bacterial and fungal pathogens, viruses (flacherie virus) and microsporidians (especially *Nosema* species) are highly pathogenic which threaten sericulture. Various molecular markers have been employed as identification tools for species identification in viruses (Awasthi *et al.*, 2008; Razmaraii *et al.*, 2013). For instance, Qanungo *et al.*, 2002 reported the characterization of genome segment 9 of cytoplasmic polyhedrosis virus (CPV), a virus that infects midgut of many insects including silkworm and suggested a new CPV that infects different *Antheraea* species (*A. mylitta*, *A. assamensis* and *A. proylei*) in different geographical regions. Similarly, PCR method based on primers targeting polyhedrin/granulin (*polh*) and late expression factor 9 (*lef9*) genes was also found to be effective in the detection of baculovirus infecting *Actias selene*

(Indian moon moth) and Grasserie disease virus (BmNPV) in *B. mori* silkworms (Skowron *et al.*, 2015; Joshi & Raja, 2016).

Further, different molecular markers like *SSU-rRNA* gene, complete RNA cistron (ITS and Intergenic spacer- IGS), *16S rRNA* gene and *rpb1* have also been found to be highly effective in the identification and classification of various microsporidians like *Nosema* sp. (Ironsides, 2013; Gisder & Genersch, 2013). Barcode markers like *LSU-ITS-SSU*, *RPB1* and combined set of *LSU-ITS-SSU* and *RPB1* have also been employed for phylogenetic studies of *Nosema disstriae* where the former set placed *N. disstriae* in *N. bombycis* clade while later two placed it with *N. antheraea* (Kyei-Poku & Sokolova, 2017).

Based on the available literature, significant progress has been made for studying diversity of disease pathogens in *B. mori* while the studies on non-mulberry silkworms are limited. Even the DNA barcode records available on BOLD and the number of publications are negligible. Work has to be done in exploring barcode regions for accurate identification of disease pathogens in non-mulberry silkworms and hence proper management. This will further pave a way towards improving the growth and development of silkworm larvae and thus sericulture productivity.

As discussed above, various mitochondrial genes have proven to be potential markers for phylogenetic and other diversity studies in various taxa. Therefore, mitogenome sequencing of silkworms is necessary which can be a good source of DNA barcode markers. The next section describes the details on mitochondrial DNA and its genome, its role in conservative and evolutionary studies and finally the sequencing status of mitogenome of important silkworm resources.

1.3 Mitochondrial genome sequencing of wild silkmoths

Mitochondria are thought to be originated from alpha-proteobacterium through endosymbiosis which arose only once in evolution around 2 billion years ago (Gray *et al.*, 2001; Martin & Mentel, 2010). They are semi-autonomous eukaryotic organelles which produce energy in the form of ATP (adenosine triphosphate) by aerobic breakdown of organic molecules through oxidative phosphorylation mechanism in order to drive cell functions (Brand, 1997). They are also responsible for other processes like apoptosis, ageing, disease, etc. (Kroemer *et al.*, 1998; Lane, 2005). Mitochondrion is a double membrane organelle (**Figure 1.2**) comprising of a porous outer membrane which separates mitochondria from cytosol and an inner membrane, consisting of highly convoluted and transverse folds known as cristae, main sites of mitochondrial energy conversion (Kühlbrandt, 2015).

Due to the extensive use of individual mitochondrial genes in numerous barcoding studies, mitogenome has huge impact on entomological research area. Whole mitogenomes are found to be more informative compared to individual genes due to many additional genome-level features like genome arrangement, relative position of different genes, secondary structures of RNA, control region involved in initiation of replication and transcription (Cameron, 2014).

In insects including lepidopterans, the mitogenome is a circular double stranded molecule of typical size ranging from 14-20 Kb (14000 -20000 basepairs). It encodes 37 genes including 13 protein coding genes (PCGs), 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNAs) involved in protein translation (Boore, 1999). Thirteen PCGs encode protein subunits that form four of the five

membrane protein complexes of respiratory electron transport chain (ETC)- complex I (NADH/ubiquinone oxidoreductase subunits 1–6 and 4L), complex III (cytochrome c reductase or cytochrome *bc1* complex), complex IV (cytochrome c oxidase subunits 1-3) and complex V (subunits 6 and 8 of the ATPase) (**Figure 1.4**). They work in combination with complex II (succinate dehydrogenase) in oxidative phosphorylation to harness energy for the cell. Complexes I, III and IV play a role in creating proton gradient by pumping protons across the cristae membrane while complex V is involved in ATP generation from ADP (adenosine diphosphate) and phosphate ions (P_i). However, majority of mitochondrial active proteins are encoded in the compartment of nucleus (Kühlbrandt, 2015).

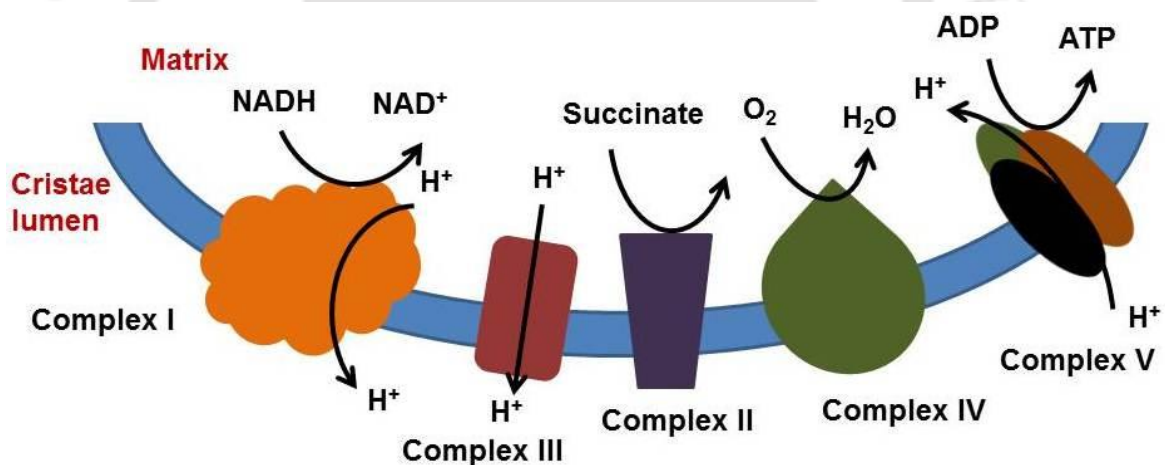


Figure 1.4 Structure of mitochondria showing membrane protein complexes of the respiratory chain. Complex I (NADH/ubiquinone oxidoreductase, II (succinate dehydrogenase), III (cytochrome c reductase), IV (cytochrome c oxidase) and V (mitochondrial ATP synthase) [Adapted from Kühlbrandt, 2015]

In addition to the above genes, mitogenome also contains several non-coding regions among which control region is the largest region which contains origin of mitochondrial replication and translation. It is also known as A+T rich region due to its high AT content (Adenine and Thymidine nucleotides) (Boore, 1999).

Various characteristics of the mitogenome of lepidopteran insects are discussed as below-

1.3.1 Genome rearrangement and gene content

Genome rearrangement has great potential in elucidating evolutionary history of insects as they serve as type of phylogenetic signatures known as rare genomic changes at order or lower taxonomical levels. Studies reveal that the mitogenome rearrangement is multifactorial which may be caused by various gene movements like inversion, transposition, inverse transposition, tandem duplication, etc. (Dowton *et al.*, 2002; Tan *et al.*, 2011). Based on the type of genes rearranged, it may be classified as major (for PCGs and rRNAs) and minor rearrangements (for tRNAs). Also, based on the movement of gene within the genome, it may be classified as short range movement i.e. local rearrangement (within multiple tRNA gene blocks) or shuffling rearrangement (gene movement at adjacent position) and long range movement i.e. across one or more blocks of PCG rearrangements where the former two arrangements are thought to be more frequent than the later (Dowton *et al.*, 2009). The mechanisms underlying this genome rearrangement are based on tandem duplication–random loss (TDRL) model, tandem duplication and nonrandom deletion (TDNR) model, and recombination model (Moritz & Brown, 1987; Boore, 2000; Lavrov *et al.*, 2002). Most of the mitogenome rearrangement in insects is based on TDRL model where a part of mitogenome is duplicated due to many replication errors like slipped strand mispairing, imprecise initiation or termination of replication, genome recombination or dimerization (Moritz *et al.*, 1987). This is followed by random elimination of one copy (non-functional) of the duplicated gene due to the selection pressure for genome size reduction. TDNR differs from TDRL in the point that gene loss is non-random as

the genes duplicated around the location of transcriptional blocks become non-functional and are thus deleted (Lavrov *et al.*, 2002). Apart from the above two models, recombination model explains the occurrence of translocation in mitogenomes. The reduction or expansion of mitogenome size is thus attributed to gene duplication or gene loss events.

In lepidopteran insects, rearrangements have been observed in two tRNA gene clusters of the mitogenome- $tRNA^{Met}$ - $tRNA^{Ile}$ - $tRNA^{Gln}$ (*trnM-trnI-trnQ* or *MIQ*) cluster and $tRNA^{Ala}$ - $tRNA^{Arg}$ - $tRNA^{Asn}$ - $tRNA^{Ser1}$ - $tRNA^{Glu}$ - $tRNA^{Phe}$ (*trnA-trnR-trnN-trnS1-trnE-trnF* or *ARNS1EF*) cluster (Cao *et al.*, 2012; Timmermans *et al.*, 2014; Wang *et al.*, 2014; Park *et al.*, 2016; Liu *et al.*, 2017). The arrangement of genes in the mitogenome of ancestral and other sequenced lepidopterans is shown in **Figure 1.5**. The gene content i.e. presence of typical 37 genes and their arrangement is conserved in most of the lepidopterans.

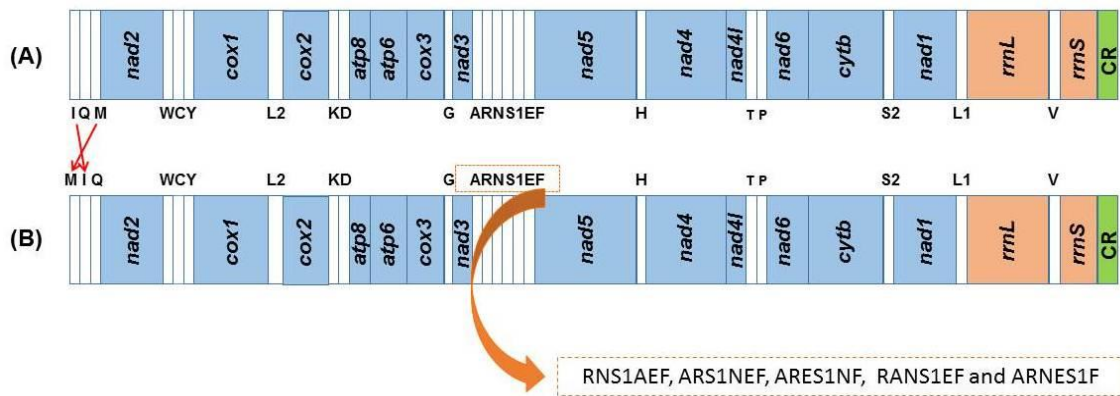


Figure 1.5 Arrangement of genes in the mitogenome of lepidopteran insects. (A) Ancestral arrangement with *trnI-trnQ-trnM* (IQM), (B) Arrangement *trnM-trnI-trnQ* (MIQ) in other sequenced lepidopteran insects diverged from ancestors during evolution. Other rearrangements are highlighted in red boxes observed in few lepidopterans.

1.3.2 Strand asymmetry in the mitogenome

The mitogenomes of insects including lepidopteran insects exhibit strong nucleotide compositional bias with high A+T content than G+C content

($A+T \neq G+C$) (Wei *et al.*, 2010; Kim *et al.*, 2012). Based on the number of genes on each strand, the two strands of mitogenome are known as J (major) strand and N (minor) strand in insects which differ from vertebrate strands (L strand or light strand and H-strand or heavy strand, respectively) (Garesse, 1998; Taanman, 1999). Strand asymmetry (strand composition biasness) is reflected in terms of skewness, AT skew ($(A-T)/(A+T)$) or GC skew ($(G-C)/(G+C)$) where A, T, G, C denotes adenine, thymidine, guanine and cytosine bases which vary between the strands (Wei *et al.*, 2010; Kim *et al.*, 2012). One strand of the mitogenome in insects is AT skewed while the other strand is GC skewed. Positive values of AT skew indicate the presence of more As than Ts and vice-versa. Similar is the case with GC skewness values. Various studies reported nucleotide composition biasness in both the strands and for different genes in Bombycoid silkmoths and other lepidopteran insects (Liu *et al.*, 2008; Pan *et al.*, 2008; Dai *et al.*, 2016). The mechanism reported by previous studies explains that this strand asymmetry underlies in asymmetrical replication of mitogenome which has greater effect than transcription process (Shioiri & Takahata, 2001; Wei *et al.*, 2010). In vertebrates, replication is unidirectional which initiates from replication origin of major strand (O_H) and the replication of minor strand begins only when the major strand reaches $2/3^{rd}$ distance from O_H where O_L (replication origin for minor strand) is located (Taanman, 1999; Saito *et al.*, 2005). This differs in insects where replication begins from O_N of minor strand of control region which then proceeds unidirectional. This is followed by initiation of major strand synthesis when 97% of the minor coding strand synthesis gets completed i.e. at the location of replication origin of major strand (O_J) (**Figure 1.6**). During the replication process, the second

strand stays in single stranded state remaining exposed to mutations thus resulting in strand asymmetry (Reyes *et al.*, 1998; Shioiri & Takahata, 2001).

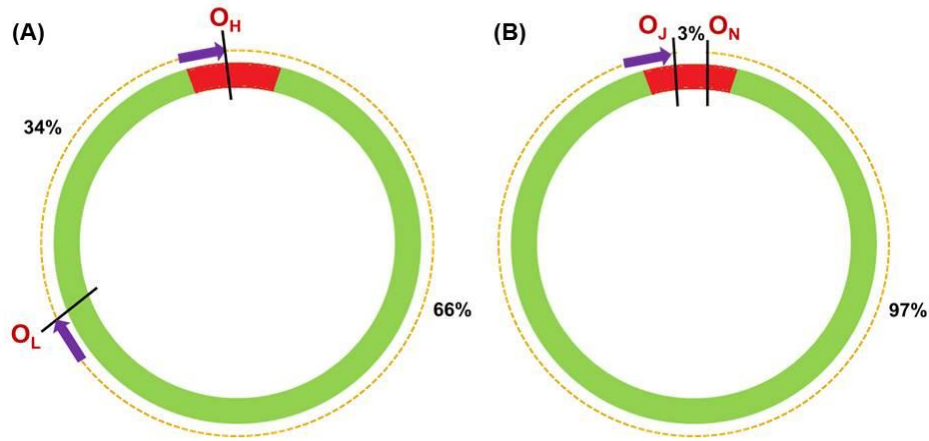


Figure 1.6 Origin of replication in control region (in red) and initiation of replication. (A) Vertebrate mitogenome where O_H denotes heavy strand, O_L denotes minor strand. (B) Insect mitogenome where O_J denotes major strand, O_N denotes minor strand. [Source: Arunkumar & Nagaraju, 2006]

1.3.3 Transfer RNAs (tRNAs)

tRNA gene rearrangements are more commonly observed in insects including lepidopterans as compared to other genes and have potential in evolutionary studies (Beckenbach, 2011). It has been reported that the genes (*IQM*) located near non-coding regions are more prone to rearrangements as explained above. There are twenty two *tRNA* genes in mitogenome of lepidopteran insects including Saturniid silkmoths. These *tRNA* genes can fold into typical clover leaf like secondary structures (Liu *et al.*, 2008; Cameron & Whiting, 2008; Dai *et al.*, 2016). These structures consist of four defined arms with defined functions (**Figure 1.7**) - (i) aminoacyl acceptor stem that holds the isotype specific amino acid, (ii) dihydrouridine (DHU) arm that serves as a recognition site for amino acyl-*tRNA* synthetase enzyme which activates the amino acid, (iii) anticodon arm that

determines tRNA isotype, and (iv) the pseudouridine (TΨC) arm that serves as a ribosome binding site (Juhling *et al.*, 2012).

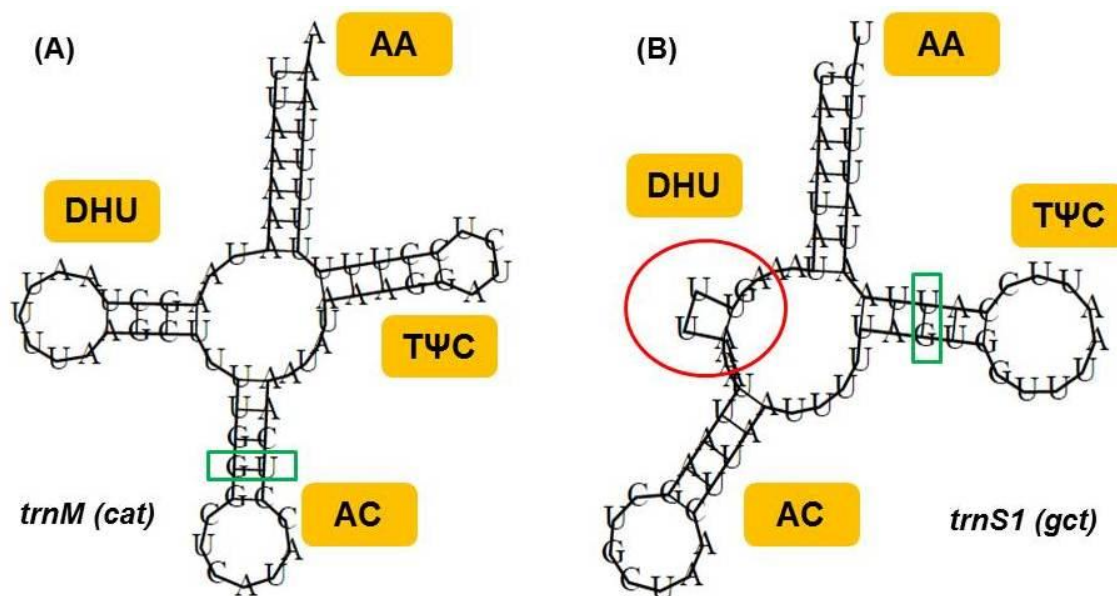


Figure 1.7 An example of secondary structures of tRNA genes in lepidopteran insects. (A) Typical clover leaf like structure (*trnM*) with four defined arms- AA: aminoacyl acceptor arm, DHU: dihydrouridine arm, AC: anticodon arm, TΨC: pseudouridine arm; (B) Secondary structure of tRNA gene (*trnS1*) with unstable DHU arm (stable stem and loop, highlighted in red). *cat* and *gct* in brackets denote anticodons of tRNA genes. Mismatches are highlighted in green (G-U pairing).

Studies on lepidopteran insect mitogenomes reveal that structures of one or a few tRNAs diverge from their canonical pattern (Wolstenholme, 1992; Yang *et al.*, 2009). For instance, lack of stable DHU arm in the secondary structure of *tRNA^{Ser1}* gene in *A. pernyi*, *A. yamamai*, *A. atlas*, *B. panterineria*, etc. (Liu *et al.*, 2008; Kim *et al.*, 2009; Yang *et al.*, 2013; Chen *et al.*, 2014). Further, mismatches are commonly found in tRNAs of lepidopteran insects, the number and type (GU, UU, GA, AA, AC, CU, etc.) of which vary across tRNA genes and across species (Kim *et al.*, 2012; Yang *et al.*, 2013; Dai *et al.*, 2016). Studies suggest that these mismatches are likely to be corrected through the mechanism of RNA editing required for proper functioning of tRNAs (Lavrov *et al.*, 2000). RNA editing refers

to post-transcriptional process that involves programmed alteration (indels or insertion-deletion, nucleotide substitution) in the primary structure of RNA to generate a sequence which could be directly encoded at DNA level. It is a common phenomenon that occurs in all eukaryotic cells, mostly in mitochondria and less frequently in chloroplast or nucleus (Gray, 2003). Apart from mismatches, conserved bases in stem and loop of different tRNAs like 7 bp aminoacyl stem, 5 and 7 bp anticodon stem and loop, respectively can be observed in lepidopteran insects. Variations are most common in the DHU arm and T ψ C arm which further varies with the type of tRNAs and across species (Kim *et al.*, 2009; Kim *et al.*, 2012; Liu *et al.*, 2012). Moreover, tRNAs reported in lepidopteran insects have identical anticodons which recognize complementary codons for corresponding amino acids (Kim *et al.*, 2009; Yang *et al.*, 2013).

1.3.4 Protein coding genes (PCGs)

Protein coding genes (PCGs) are the most widely used mitochondrial genes for phylogenetic, barcoding and other evolutionary studies (Carapelli *et al.*, 2007). The mitogenome of lepidopteran insects comprises of thirteen PCGs- *cox1*, *cox2*, *cox3*, *nad1-6*, *nad4L*, *cytb*, *atp6* and *atp8*. *cox1-3* genes encode cytochrome c oxidase subunits from complex IV, *atp6* and *atp8* genes encode ATP synthase subunits from complex V, *cytb* (*cob* or cytochrome b) encode complex III and *nad1-6*, *4L* encode NADH subunits from complex I (Kim *et al.*, 2012; Cameron, 2014). These genes are encoded by major and minor strand with different initiation and termination codons. Most of the PCGs of lepidopteran insects are initiated with typical ATN codons (ATG, ATA, ATC, ATT) and terminated with complete termination codons (TAA or TAG) (Liu *et al.*, 2008; Liu *et al.*, 2016). However, divergence from canonical codon patterns has commonly been

observed in many lepidopteran insects such as CGA for *cox1* gene in *Thitarodes* sp., *Eriogyna pyretorum*, etc.; TTG in *Caligula boisduvalii*; TTAG in Chinese *B. mandarina*, etc. (Hong *et al.*, 2008; Pan *et al.*, 2008; Jiang *et al.*, 2009; Cao *et al.*, 2012). Similarly, truncated termination codons (T- or TA) have been reported in many lepidopteran insects such as *A. pernyi*, *A. yamamai*, Chinese *B. mandarina* and so on (Liu *et al.*, 2008; Pan *et al.*, 2008). Studies reveal that these codons are completed by post-transcriptional polyadenylation. This involves processing of polycistronic transcript molecule into mature RNA through precise endonucleolytic cleavage by using recognition signals of secondary structures of tRNAs (Cameron & Whiting, 2008).

Various studies reported that different selection pressures act on mitochondrial PCGs which may be associated with energy metabolism, thermoregulation, climatic conditions, etc. Therefore, estimating evolutionary rates of protein coding genes can be important for mitochondrial evolution and gene expression studies (Oliveira *et al.*, 2008; Lawrie *et al.*, 2013; Zhang *et al.*, 2017). It may be measured in terms of Ka/Ks ratios as denoted below:

$Ka = \text{Non-synonymous substitutions} / \text{Non-synonymous site}$

$Ks = \text{Synonymous substitutions} / \text{Synonymous site}$

Synonymous substitutions refer to the mutations in coding sequence (CDS) without changing its amino acid sequence while non-synonymous substitutions refer to the mutation in coding genes which in turn changes its amino acid sequence (Song *et al.*, 2016). The ratio (Ka/Ks) less than 1 (<1), greater than 1 (>1) and equals to 1 (=1) denote negative (purifying) selection, positive (adaptive) selection and neutral evolution, respectively (Meiklejohn *et al.*, 2007; Song *et al.*, 2016). A study showed smaller Ka/Ks values of mitochondrial PCGs of migratory

and cold climate teleost fishes than non-migratory and tropical teleost fishes while no such trend was observed for nuclear genes. This suggests that different functional constraints due to energy and thermoregulation affect the evolutionary rates of mitochondrial PCGs (Sun *et al.*, 2011). Similar study performed in various genera of malacostraca demonstrated that the genes encoding protein complexes (I and V) exhibited higher Ka/Ks ratios as compared to that of complex (III and IV) which indicates that the former complexes evolved under accelerated (positive) selection while the later evolved under purifying (negative) selection (Zhang *et al.*, 2017). A comparative study performed on domesticated and wild type *A. pernyi* silkmoths also revealed different purifying selection pressures acting on PCGs (Liu *et al.*, 2012). Evolutionary studies based on Ka/Ks ratio of mitochondrial PCGs in Bombycoid silkmoths are negligible. However, few studies have reported nucleotide substitution studies in terms of transition-transversion rates (ts/tv) to study evolutionary patterns among silkmoths (Li *et al.*, 2010; Liu *et al.*, 2012). Transitions refer to changes between purines (A↔G) or between pyrimidines (T↔C) while transversions refer to changes between purines and pyrimidines (A↔T, C↔A, T↔G, C↔G). Transitions are more commonly found in closely related species while transversions are observed in distantly related species (Yang & Yoder, 1999; Belle *et al.*, 2005). The comparative studies on wild and domesticated silkmoths revealed that biasness in transitional and transversional mutations may be associated with different environmental selection pressures acting on PCGs (Li *et al.*, 2010; Liu *et al.*, 2012). The gene by gene divergence in PCGs in terms of conserved, variable and phylogenetically informative sites have also been reported in Bombycoid silkmoths like *Manduca sexta* which has

significance in molecular systematics in selecting potential phylogenetic barcode markers (Cameron & Whiting, 2008).

Codon usage biasness has been well studied in many insects where few synonymous codons are more frequently used than others during translation process. Such codons are known as preferred or optimized codons (Akashi, 1997; Sun *et al.*, 2009). Studies revealed that 22 codons are recognized by 22 mitochondrial encoded tRNAs while others are imported from nucleus to carry out protein synthesis in lepidopteran insects (Mirande, 2007). Biasness in codon usage ensures translational efficiency and accuracy by avoiding mis-incorporation of amino acids. The factors affecting codon usage may be natural selection pressure (gene length and function, translational selection) or mutational pressure (base composition, base position mutation) (Sharp *et al.*, 1993; Akashi, 1997; Powell & Moriyama, 1997). It varies among species and even among different genes of mitogenomes and thus is relevant for gene expression and evolutionary studies among species under different conditions and at different time scales (Wei *et al.*, 2014). Based on degeneracy of genetic code, one amino acid can be encoded by multiple codons which typically differ at third codon position. Compared to first and second codon position, third codon position greatly influences codon usage biasness. Previous reports on the effect of GC content at third codon position revealed that GC mutational bias influences the codon and amino acid usage in many prokaryotes and eukaryotes (Knight *et al.*, 2001; Palidwor *et al.*, 2010). Similarly, the most frequent codons were found to be rich in C nucleotides at 3rd position in *B. mandarina* and *Ostrinia furnacalis* suggesting that mutational pressure mainly affects codon usage biasness in few lepidopteran insects (Sharma *et al.*, 2014). However, a study performed on *B. mori* revealed

that natural selection majorly affects codon usage biasness while mutational biasness plays only a minor role. Unlike dipteran insects (mostly GC rich), AT rich codons were more frequently used in most of the lepidopteran insects than GC rich codons similar to hymenopteran insects (Sun *et al.*, 2016). The AT content was higher at third codon position than first and second positions of most of the sequenced lepidopterans. Occurrence of mostly AT rich codons encoding amino acids Phenylalanine, Leucine and Isoleucine have been reported in *A. yamamai*, *Andraca theae*, *Biston panterinaria*, *Thitarodes* species and many other lepidopteran insects (Kim *et al.*, 2009; Cao *et al.*, 2012; Yang *et al.*, 2013; Gu *et al.*, 2016). Thus, AT biasness greatly influences codon usage pattern which contributes a major role in AT richness of the mitogenome.

1.3.5 Non-coding regions

Among all non-coding regions of the mitogenome, control region, also known as “A+T-rich” region is the major and longest non-coding region in insects including lepidopteran insects. This region is heavily-biased towards A/T nucleotides and contains origin sites for mitochondrial replication and transcription (Cameron & Whiting, 2008; Liu *et al.*, 2008). Studies reveal that high A+T (>90%) content may be due to high AT mutational pressure acting on this region during evolution or the enzymes involved in replication and transcription functioning with high AT content (Clary & Wolstenholme, 1985). Diversity in the location of control region resulting from translocation of genes flanking this region and its high A+T content creates difficulty in primer designing and phylogenetic studies. Among lepidopteran insects, this site also varies in length (~300 to 700 bp) which may be due to the presence or absence of tandem repeats (TEs) (Zhang & Hewitt, 1997; Cameron & Whiting, 2008). Presence of several conserved structures such as ‘ATAGA’ motif

followed by poly-T stretch (T_n), ATTTA elements, microsatellite like elements (AT_n or TA_n), poly A tail, etc. are known to be characteristic features of this region in insects (Yukuhiro *et al.*, 2002; Jiang *et al.*, 2009; Dai *et al.*, 2016). ATAGA motif and poly T stretch are mostly conserved in lepidopteran mitogenomes except in ancestral species (Cao *et al.*, 2012) where former serves as site of origin of replication while later acts as a signal for recognition of proteins involved in the initiation of replication of minor-strand of mitogenome (Taanman, 1999; Saito *et al.*, 2005). Studies suggest that poly A tail upstream of *trnM* may have role in controlling transcription and/or replication initiation or may have other unknown functions (Zhang *et al.*, 1995). In addition, TEs also occur in various Bombycoid silkmths like *A. pernyi* (six 38 bp repeats), Japanese *B. mandarina* (3 tandem repeats of 126-bp), etc. and have also been reported in ancestral species (4.9 and 8.5 units of TEs in *Thitarodes* sp.) (Yukuhiro *et al.*, 2002; Liu *et al.*, 2008; Pan *et al.*, 2008; Cao *et al.*, 2012; Kong & Yang, 2015). This may be due to slipped strand mispairing which takes place during replication in mitogenome resulting in variable length of control region in lepidopteran insects (Levinson & Gutman, 1987). However, many Saturniid moths lack such conspicuous repeat elements except short repeats present in poly T stretch (Jiang *et al.*, 2009; Chen *et al.*, 2014). Another feature observed in some lepidopteran insects including Saturniids is that they harbor some additional stem and loop structures or tRNA like structures (Hong *et al.*, 2008; Jiang *et al.*, 2009). Such extra tRNAs have also been reported in few other lepidopteran insects (Cha *et al.*, 2007; Kim *et al.*, 2011). Studies suggest that such tRNA structures are remnants of tRNAs which showed illicit priming within the control region and failed to cleave from the nascent DNA strand, thus resulting in its addition into the mitogenome. Also,

sequence divergence of the newly incorporated tRNA as compared to the original tRNA may be explained in terms of non-coding nature of the control region and structure divergence of incorporated tRNA (Cantatore *et al.*, 1987). This region thus serves as a source of polymorphic markers which can be used for population genetics and phylogenetic studies of closely related taxa.

Apart from control region, mitogenome of lepidopteran insects consists of many intergenic spacers (IGS) and overlapping sequences (OS). These regions vary in length and number in different species. Studies revealed that IGS may result in mitogenome size variation either through deletion or duplication. These IGS and OS contain conserved structural elements in several lepidopteran insects. Sequence homology of OS with their adjacent genes or distant genes has also been reported in many studies (Salvato *et al.*, 2008; Cameron & Whiting, 2008; Kim *et al.*, 2009).

1.3.6 Ribosomal RNAs (rRNAs)

The mitogenome of lepidopteran insects comprises of two rRNA genes, *12S rRNA* (*sRNA* or *rrnS*) and *16S rRNA* (*lRNA* or *rrnL*) which are essential for protein synthesis and vary in length across species (Brimacombe & Stiege, 1985). *12S rRNA* is generally located between *tRNA^{Val}* and control region while *16S rRNA* is located between *tRNA^{Leu1}* and *tRNA^{Val}* genes (Liu *et al.*, 2008; Kim *et al.*, 2009). These rRNA genes can fold into stem and loop secondary structures and have been predicted in few lepidopteran insects like *Manduca sexta*, *Grapholita molesta*, etc. (Cameron & Whiting, 2008; Gong *et al.*, 2012). These rRNA genes have potential as molecular markers and have been employed for various phylogenetic studies.

1.3.7 Importance of mitogenome studies in Saturniid silkmoths (wild silkmoths)

Advancement in sequencing technologies has led to the generation of huge amount of mitogenome data in various taxa. Both sanger based sequencing and next generation sequencing (NGS) technologies have been used for many mitogenomics studies in insects (Timmermans *et al.*, 2010; Timmermans *et al.*, 2014; Liu *et al.*, 2016). Utility of mitochondrial *COI* gene as a universal barcode at a broad spectrum has led to an increased rate of sequencing of mitochondrial genomes. Till date, mitogenome of ~300 insects belonging to different orders have been sequenced and registered in NCBI GenBank database (<https://www.ncbi.nlm.nih.gov/>). There is a rapid accumulation of mitogenome sequencing data of Lepidoptera (approximately more than 180 organisms), one of the most diverse orders consisting of butterflies and moths (**Figure 1.8**).

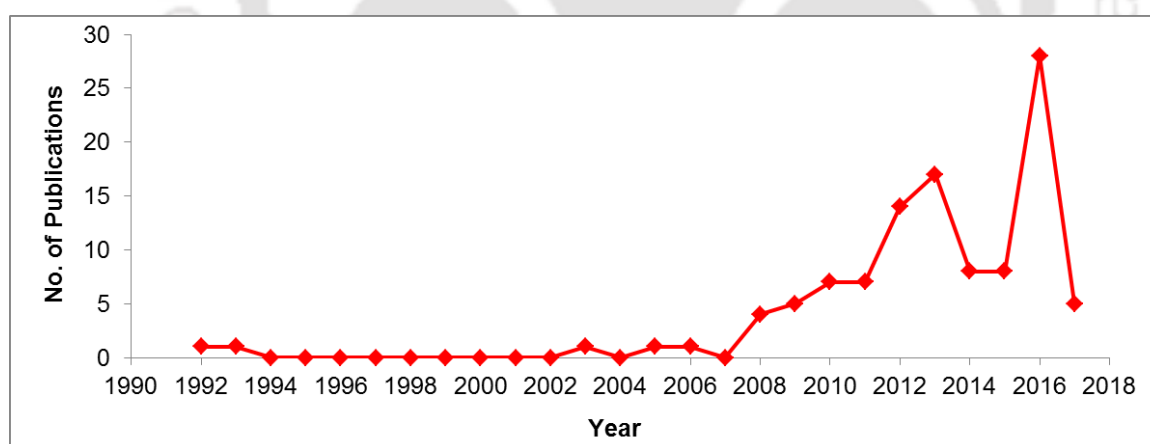


Figure 1.8 Trend of publications entitled “Mitochondrial genome of Lepidopteran insects” (Source: Scopus dated 14.08.2017).

Based on the search on various species belonging to different families and superfamilies of order lepidoptera (**Figure 1.9**), we found that among various superfamilies, the number of species with either complete or partial mitogenomes were maximum in Papilionoidea followed by Bombycoidea and Noctuoidea.

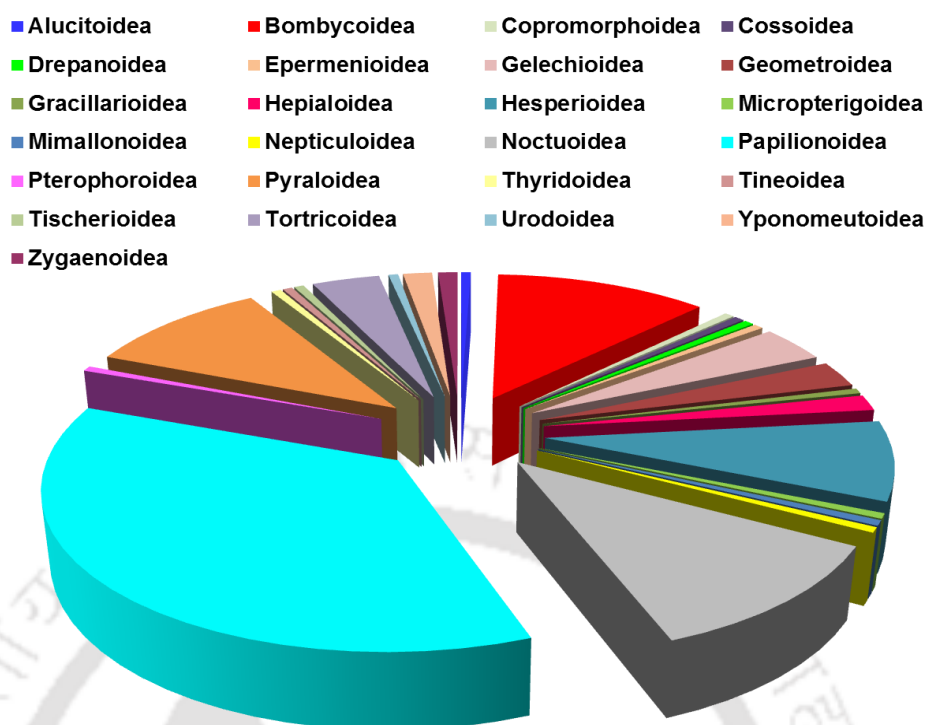


Figure 1.9 Frequency of mitochondrial genome sequencing of organisms belonging to various superfamilies of Lepidoptera (Data search from NCBI GenBank database).

Within Bombycoidea superfamily, Bombycidae and Saturniidae are the most economically important families. Saturniidae itself consists of more than 1800 species as per published records. The mitogenome information of the wild silkworm species may provide fundamental information for future comparative, phylogenetic and evolutionary studies (Kim *et al.*, 2009; Jiang *et al.*, 2009; Timmermans *et al.*, 2014). Moreover, studying phylogenetic relationships among the members of Saturniidae as well as with Bombycidae members will help in better understanding of physiological variations at molecular level and creating transgenic silkworms capable of high quality silk production. Considering the diversity of these silkworms and their economic importance, knowledge of the mitogenomic sequences is surprisingly sparse. Only few mitochondrial sequences of wild silkmoths have been registered in GenBank. Therefore, there is a need of

sequencing mitochondrial genomes of wild silkmoths which has become a gold standard for molecular taxonomy and phylogenetic studies.

1.4 Seri-informatics databases

More than fifty years have passed since the time when the term “database” was coined. However, it was only during the massive digitalization of many resources like archives of music, books, etc. in the 90s that the same term started reflecting its primary usage in today’s world as a data organizational model (Marie-Laure *et al.*, 2014). During these years, the databases have also inherited the ability to retrieve and filter data in various ways. Integration of these databases with biology has brought digital revolution to life science. The amalgamation of biology with information technology for data dispersion and statistics for data analytics has led to the development of some highly successful databases like GenBank, RCSB Protein Data Bank (PDB), etc. (Bernstein *et al.*, 1997). Now, there are databases in almost every field of biology ranging from diseases, whole organisms, taxonomy, bioactive products, etc. making them indispensable for the researchers (Flybase Consortium, 2003; Bilimoria *et al.*, 2008; Federhen, 2012; Kumar *et al.*, 2015). One of the research fields in which databases are being constructed actively is “Sericulture”.

Silkworms and their respective host plants are the key players of sericulture and silk is its prime yield. Researchers also developed mutants of these organisms for improving silk quality and quantity, understanding their physiology and exploiting them as bioreactors for recombinant proteins (Goldsmith *et al.*, 2005; Inoue *et al.*, 2005; Daimon *et al.*, 2012). The need of databases in sericulture field cannot be emphasized more. Firstly, numerous organisms are involved in this field and scientists have uncovered minuscule information about most of them while some

are yet to be identified. Secondly, the data that are generated in this field are of dissimilar type. Each data ranging from nucleotide and protein sequence to gene maps, expression profiles and biomaterials, is unique and vital. Thirdly, the amount of data generated is huge owing to the fast-evolving techniques of sequencing, analysis, imaging, etc. Especially, the sequencing techniques have progressed beyond shotgun-sequencing to more quick and efficient next generation sequencing (NGS), chromatin immuno-precipitation sequencing or ChIP-Seq, etc. which produce millions of sequence data at a go (Aparicio *et al.*, 2005; Mardis, 2008). Till now (2003-present year), the total number of published databases in this seri-bioresource field is 50 out of which 27 were created in the last five years (**Figure 1.10**) and are briefly discussed as below-

1.4.1 Silkworm databases

Silkworms form the backbone of seri-ecosystem and huge research has been done on it. The databases comprising of silkworm specific information such as nucleotide, protein, genetic resource and pathway databases are listed in **Table 1.3** and classified into different categories based on their data content as follows-

1.4.1.1 Nucleotide databases

Nucleotide databases constitute diverse nucleotide information like genomic sequences, expressed sequence tags (ESTs), microarray, microsatellites, transcriptomic data, etc. They provide fast and easy accessibility of sequence information for biological, functional, comparative genomics and phylogenetic studies. The available nucleotide databases are as follows-

Silkworm genome databases

The first draft genome of the lepidopteran model organism, *B. mori* was published in 2004 with 3X and 6X coverage by two independent research groups from Japan

and China, respectively (Mita *et al.*, 2004; Xia *et al.*, 2004). As a result, the databases comprising genome information of *B. mori* were developed such as SilkDB, SilkDB v2.0 and KAIKObase detailed in **Table 1.3** (Wang *et al.*, 2005; Duan *et al.*, 2010; Shimomura *et al.*, 2009). Although well-developed genome databases of *B. mori* are available, there is a scarcity of genomic data related to other silkworms, particularly the wild silkworms. This may be attributed to the issues related to their domestication due to which they are still unexplored. Genome information is necessary for various downstream studies like mutation, mapping studies, etc. Development of new databases or integration of such data in the available databases will help in analysis of their genome function and evolution.

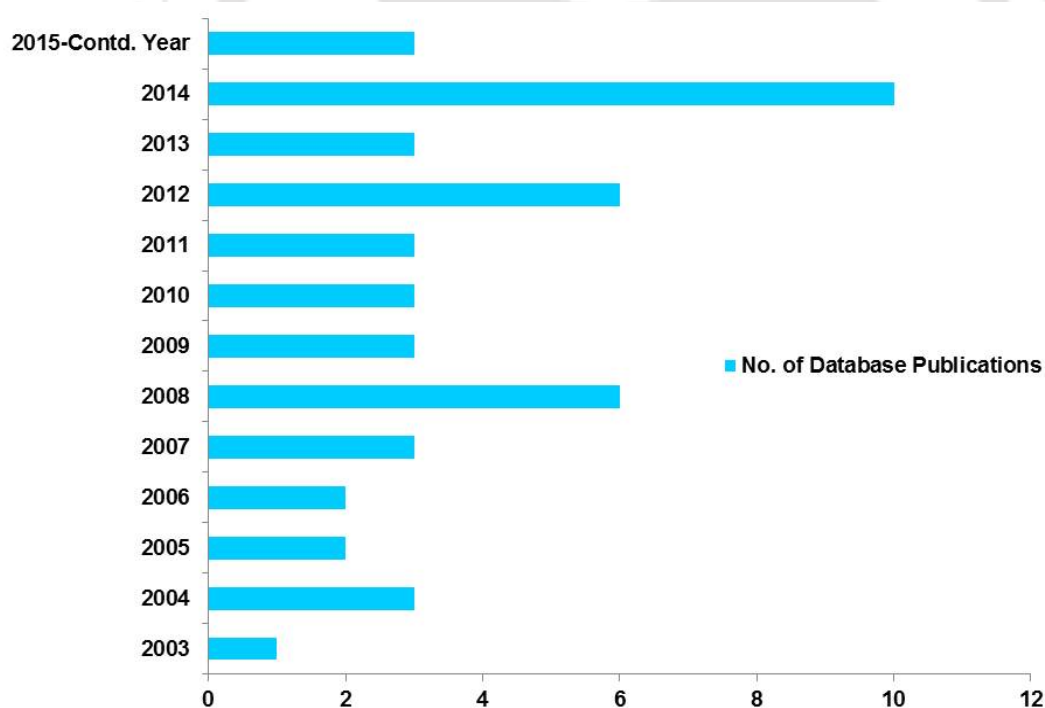


Figure 1.10 Number of publications on seri-related databases from the year 2003 to 2015- Continued year (Data shown for 2015-Contd. year as only two databases have been published during 2016-17).

Table 1.3 List of silkworm databases and their data contents.

Database		Datasets	Developed by	References
Silkworm Nucleotide Databases				
SilkDB	Genomic data	428.7 Mb <i>B. mori</i> genomic sequences (23156 scaffolds for 28 chromosomes; 18510 annotated gene sequences and full cDNA sequences of 212 known silkworm genes), 16 425 EST clusters; around 601225 TEs	Beijing Genomics Institute (BGI), China	Wang <i>et al.</i> , 2005
SilkDB (v2.0)	Genome, functional annotation, microarray, repeats, experimental data	43622 scaffolds spanning ~432 Mb; 1668 repeat sequences; a consensus non-redundant dataset with 14623 PCGs; predicted noncoding genes including 206 miRNAs, 147 rRNAs and 498 tRNAs; 184509 ESTs and full-length cDNAs	Institute of Sericulture and Systems Biology, Southwest University (SWU), China	Duan <i>et al.</i> , 2010
Kaikobase (v 3.2.2)	Genomic Sequence, map data	43462 assembled data corresponding to total genome size of 482 Mb in <i>B. mori</i> ; 81705 BAC-end sequences, 174222 fosmid-end sequences, 166757 ESTs; 16209 gene models, 1532 SNP markers, 770 trait markers, etc.	National Institute of Agrobiological Sciences (NIAS), Japan	Shimomura <i>et al.</i> , 2009
SilkBase	EST data	35000 ESTs from 36 cDNA libraries, grouped into ~11000 nonredundant ESTs with the average length of 1.25 kb in <i>B. mori</i>	NIAS Japan, University of Tokyo	Mita <i>et al.</i> , 2003
WildSilkBase	EST data	57113 ESTs clustered and assembled into 4019 contigs and 10019 singletons; 118 cSNPs predicted in 1412 EST sequences; 14038 putative unigenes identified in 3 species (<i>A. assama</i> , <i>A. mylitta</i> and <i>S. ricini</i>) of	Centre for DNA Fingerprinting and Diagnostics (CDFD), India	Arunkumar <i>et al.</i> , 2008

ButterflyBase	EST, annotation, protein prediction	wild silkmoths A total of 273 077 ESTs from more than 30 different butterflies & other lepidopteran insects clustered to generate stable unigene sets and robust protein translations derived from each unigene cluster	Max Planck Institute for Chemical Ecology, Germany	Papanicolaou <i>et al.</i> , 2008
BmMDB	Microarray data	22987 silkworm gene probes covering the genes predicted from 6x draft genome sequences of <i>B. mori</i> ; 10 393 active transcripts; Raw data for all microarray experiments	Institute of Sericulture and Systems Biology (ISSB), Southwest University, China	Xia <i>et al.</i> , 2007
SilkTransDB	Silkworm transcriptome data	3.3 gigabase (Gb) reads covering around 7-fold of <i>B. mori</i> genome and PCGs that constitute 81.3% of all the predicted genes in SilkDB; 5500 novel transcripts and 13,195 new exons	Chinese Academy of Agricultural Sciences, China	Li <i>et al.</i> , 2012
InsectBase	Integrated insect genome & transcriptome	>12 millions of sequences (genomes of 138 insects, transcriptomes of 116 insects, gene sets of 61 insects, 36 gene families of 60 insects, 7544 miRNAs of 69 insects, 96925 piRNAs from two insects, 22536 pathways of 78 insects, 679881 UTR of 84 insects and 160905 CDS of 74 insects)	Li Lab: Insect Genomic and Bioinformatics Lab, China	Yin <i>et al.</i> , 2016
SilkSatDB	Microsatellite data	Information pertaining to 12000 WGS contigs downloaded from DDBJ (http://www.ddbj.nig.ac.jp/anoftp-e.html), 9300 EST and 50 genomic sequences of <i>B. mori</i>	CDFD, Hyderabad, India	Prasad <i>et al.</i> , 2005

InSatDB	Microsatellite data	Microsatellite characteristics of five fully sequenced insect genomes (<i>B. mori</i> , <i>Drosophila melanogaster</i> , <i>Anopheles gambia</i> , <i>Apis mellifera</i> , <i>Tribolium castaneum</i>)	CDFD, Hyderabad, India	Archak et al., 2007
Silkworm genome database	Mutant images	Photographs of mutants (having mutations on different chromosomes) in <i>B. mori</i>	IGB Lab, University of Tokyo, Japan	http://papilio.ab.a.u-tokyo.ac.jp/genome/
ABURAKO	Mutant data	List of <i>B. mori</i> mutants, linkage maps	NIAS, Japan	http://cse.nias.affrc.go.jp/natuo/en/aburako_top_en.htm
Bombyx Trap Database	Mutant data	288 transposon insertion lines (enhancer-trap lines, gene-trap lines), the positions of insertions in the genomic sequence, gene expression profiles and associated photos of <i>B. mori</i>	NIAS, Japan	Uchino et al., 2008
BmTEdb	Transposable elements data	1308 TE families identified in <i>B. mori</i> silkworm genome	Chongqing University, China	Xu et al., 2013
Silkworm Protein Databases				
Kaiko2DDB	Silkworm protein data	116 images of 2D-PAGE of different tissues of <i>B. mori</i> and at different developmental stages	NIAS, Japan	Kajiwar et al., 2006
Insect Proteome Database	Silkworm protein data	2-DE information of proteins from different tissues of <i>B. mori</i> silkworm including silk gland, fat body, mid gut and hemolymph	Kyushu University, Japan	http://shigen.nig.ac.jp/ispd/
SilkProt	Annotated silkworm protein data	Annotated silkworm proteins	Bioinformatics Centre, CSR&TI, Mysore, India	http://btismysore.in/silkprot/
SilkPPI	Protein-protein interaction data	7736 predicted protein-protein interaction pairs including 2700 unique proteins of the silkworms	Bioinformatics Centre, CSR&TI, Mysore, India	Sumathy et al., 2014

SilkTF	Transcription factor data	Transcription factors of <i>B. mori</i>	Bioinformatics Centre, CSR&TI, Mysore, India	http://btismysore.in/SilkTF/index.html
Silkworm Genetic Resource Databases				
Silkworm Gene Resources database	Silkworm Gene resources	Characterization information of 321 silkworm varieties of six different regions-Korea, Japan, China, Europe, Tropical and non-classified group; 1132 images; univoltine (32 races), bivoltine (188 races), multivoltine (11 races) and other races (90 races)	National Academy of Agricultural Science (NAAS), Korea	Kim <i>et al.</i> , 2010 (Web link-Not accessible)
SilkwormBase	Silkworm genetic resource and mutant data	456 phenotypically classified strains and a total of 419 genes	Part of NBRP between the resource centre (Graduate School of Agriculture, Kyushu University) and the information centre (National Institute for Genetics)	http://shigen.nig.ac.jp/silkwormbase/about_kaiko.jsp
Silkworm Pathway Databases				
iPathDB	Insect pathway data	12111 pathways of molecular interactions of 52 insects including 37 genome-sequenced and 15 transcriptome-sequenced ones	Li Lab: Insect Genomic and Bioinformatics Lab, China	Zhang <i>et al.</i> , 2014

Abbreviations: ESTs: expressed sequence tags; IGB Lab: Laboratory of Insect Genetics and Bioscience; NBRP: National BioResource Project; PCGs: protein coding genes; SNP: single nucleotide polymorphism; TEs: Transposable elements; 2-DE: Two-dimensional gel electrophoresis; 2D-PAGE: Two dimensional polyacrylamide gel electrophoresis.

Silkworm gene expression databases

Gene expression studies have been applied on *B. mori* and other silkworms, leading to the development of databases such as EST databases (“SilkBase”, “WildSilkbase” and “ButterflyBase”), microarray database- “*Bombyx mori* Microarray Database” (BmMDB) and transcriptome databases (“SilkTransDB” and “InsectBase”). The data content and developers of these databases are given in

Table 1.3.

The gene expression data, viz. EST, microarray and transcriptome, from these databases will help in the identification of unique lepidoptera-specific ESTs, unique genes and proteins as well as development of molecular markers and identification/annotation of unknown proteins. These data can also have implications on evolutionary studies on insects.

Microsatellite databases

Microsatellites are repeated sequences of 1-6 bp length which are widely used in fingerprinting, linkage analysis, marker studies, etc. (Toth *et al.*, 2000). Similar to genome and gene expression databases, microsatellite databases (“SilkSatDb” and “InSatDb”) are available only for *B. mori* (**Table 1.3**) while being unavailable for other lepidopteran insects, particularly wild silk moths (Prasad *et al.*, 2005; Archak *et al.*, 2007). Genomic information can greatly assist in microsatellite studies as discussed above and thus opens new scope for constructing more microsatellite databases.

Silkworm mutant databases

Being a model organism for insects, especially silkworms, *B. mori* mutants are generated for research as well as commercial interests and are expected to fulfill requirements like enhanced production of fibroin, different colored silks, etc.

(Uchino *et al.*, 2008). Web search currently shows three databases (**Table 1.3**) that host information on such mutants, namely, “Silkworm Genome Database”, “ABURAKO Database” and “Bombyx Trap database” (Shimomura *et al.*, 2009).

Transposable elements (TEs) databases

“BmTEdb” is the only database exclusively available for transposable elements (1308 TE families) of *B. mori* which have been further classified into sub-families (Xu *et al.*, 2013). TEs are said to represent ~40% of the silkworm genome (Osanai-Futahashi *et al.*, 2008). Study of transposons including their identification, characterization and annotation is crucial as it provides insight into genome variation and evolution (Fedoroff, 2012). In addition to *B. mori* which has been studied well in BmTEdb, there is a great scope of developing databases for many other related silkworms.

Other web-resources

Apart from the above databases, the sequence and frequency information of the ovarian small RNAs in *B. mori* can be retrieved from a web-platform “Silkworm sRNA” (http://papilio.ab.a.u-tokyo.ac.jp/small_RNA/all_smallRNA.txt) supported by National Bioresource Project (<http://www.nbrp.jp/>).

1.4.1.2 Protein databases

The interest in studying silkworms is deeply rooted in the proteins (fibroin and Sericin) that it produces. So, protein databases serve as an essential platform in studying gene expression, post translational modifications and other biological processes related to silkworm proteins (Kajiwara *et al.*, 2006). The available databases directly related to this area are “KAIKO2DDB” (Silkworm proteome database or SPD), “Insect Proteome Database”, “SilkProt”, “SilkPPI” (Silkworm Protein-Protein Interaction database) and “SilkTF” (Silkworm Transcription Factor

Database) (**Table 1.3**). The former two databases house 2D gel-electrophoresis information of proteins from tissues of *B. mori* silkworm while the later host annotated protein data, protein-protein interaction data and transcription factor (TF) data of *B. mori* silkworm, respectively (Shimomura *et al.*, 2009; Sumathy *et al.*, 2014).

Apart from these, an upcoming database “Silkgpcr” has also been reported on the web page of *Bioinformatics Centre, CSR&TI, Mysore, India*. Unlike nucleotide databases, there is dearth of databases related to protein structure, sequence, protein-protein interactions, etc. New databases focused on silkworm protein structure or sequences can be developed by researchers. Thus, the scope of database development in this field is huge. Implementation of combinatorial approach (Proteomics and Transcriptomics) can be one of the ways to understand how variation in the proteome and transcriptome is associated with physiological changes in silk production, to characterize strains, etc.

1.4.1.3 Silkworm genetic resource databases

Silkworm genetic resource databases which deal with data like varieties, strains, races, etc. other than genomes, transcriptomes and genes, also form an integral part of seri-databases. “Silkworm Gene Resources database” and “SilkwormBase” are the databases that have been developed to provide such information and are listed in **Table 1.3** (Kim *et al.*, 2010). SilkwormBase also links to a new website which deals with collection, preservation and distribution of few wild silkworms.

1.4.1.4 Insect pathway databases

“iPathDB” is the only insect pathway database that houses pathway data on Lepidoptera (10 different orders) with a total of ~52 insects (Zhang *et al.*, 2014). Its strongest feature is the inclusion of a pathway construction software

“iPathCons” which facilitates users to construct pathways from transcriptome as well as official gene sets (OGSs) data of insects. These pathways will be helpful for entomological research community and it is available at URL- <http://ento.njau.edu.cn/ipath/>.

1.4.2 Silkworm host plant databases

Host plants are the most important resources in seri-ecosystem as they provide food and nutrition to the silkworms. Certain host plants of silkworms have economic importance other than sericulture and have been studied with different focus. The host plant databases/web-resources available so far are listed in **Table 1.4** along with their data contents and are briefly discussed as below-

1.4.2.1 Databases of mulberry

Recent advances in HT sequencing technology have led to generation of mulberry (primary host plant of *B. mori*) specific databases that are “Morus Genome Database” (MorusDB) and “Mulberry Microsatellite Database” (MulSatDB) housing information on mulberry genome as well as EST based microsatellite (SSR) markers (Li *et al.*, 2014; Krishnan *et al.*, 2014). Other databases include Mulberry Genome Database, Database of DNA sequences for important plant genes in mulberry, MulDis (A Comprehensive Mulberry Disease and Pest Database), Sample Web Application for Analysis of Molecular ID and MulTF. Except Mulberry Genome Database (<http://www.btismysore.in/MulGenomeDb/mghome.html>), other databases are under construction (URL <http://www.btismysore.in/pgene.html>, <http://www.btismysore.in/dbase.html>) aiming to provide biological data on mulberry.

Table 1.4 List of silkworm host plant databases and their data contents.

Database		Datasets	Developed by	References
MorusDB	Mulberry genome data	<i>M. notabilis</i> genome data-110 760 scaffolds, 27 085 PCGs, 81 rRNA genes, 560 tRNA genes, 311 snRNA genes, 233 microRNA genes; repeats-324 736 repetitive sequences (53 725 known TEs, 271 011 others)	SWU, China	Li <i>et al.</i> , 2014
MulSatDB	Microsatellite data	217,312 microsatellites from whole genome and 961 microsatellites from EST sequences of <i>M. notabilis</i> (SSR marker location, repeat kind, motif type and size)	CSR&TI, Mysore, India	Krishnan <i>et al.</i> , 2014
Castor Database	Different phenotypic characters of different Castor germplasms	294 different germplasms including 20 FC5 plants and YRCH plants; Qualitative and quantitative data	TNAU (Tamil Nadu Agricultural University), Coimbatore, India	http://www.tnaugenomics.com/cast (link not accessible)
JCVI Castor Bean Genome DB	Genome data	4X draft of the ~400 Mbp castor bean (<i>R. communis</i>) genome, ~31221 putative proteins, ~50000 ESTs from different tissues	J. Craig Venter Institute, USA	Chan <i>et al.</i> , 2010
CastorDB	Gene and protein data	Gene expression and gene ontology annotation; predicted cellular localization, domains, pathways, protein-protein interactions, sumoylation sites and biochemical properties	M. S. University of Baroda, India	Thakur <i>et al.</i> , 2011
Papaya-DB	Genome data	Papaya physical and genetic maps, shotgun WGS and EST sequences and sex-determining region	Center for Applied Genetic Technologies, USA	http://www.plantgenome.uga.edu/papaya/

CPR-DB	Papaya repeat data	Repetitive elements (56% of the genome) of <i>C. papaya</i> - 52% TEs, 1.3% TRs and 3% high copy number genes	University of Maryland, USA	Nagarajan & Navajas-Pérez, 2014
Jatropha genome DB	Genome data	Genome: total length of sequences- 297,661,187 bp, average G + C content was 33.7%; DNA markers of <i>J. curcas</i>	Kazusa DNA Research Institute, Japan	Sato <i>et al.</i> , 2011
Cassava Genome Database	EST, SNP data	BAC-based fingerprint map of an inbred cultivar of Cassava-70000 BAC clones, Assembled Cassava ESTs, SNP information	Institute for Genome Sciences, USA	http://cassava.igs.umaryland.edu/cgi-bin/index.cgi
Chinese Cassava Genome Database	Genome data	Assembled draft genome sequence and genome annotation of Cassava	Fudan University, China	http://www.cassava-genome.cn/
Cassava-Base	Genome data	Cassava genome, annotation, phenotype, markers, breeding programmes, etc.	NEXTGEN Cassava Breeding project, hosted by Boyce Thompson Institute for Plant Research, USA	https://www.cassavabase.org/
Quercus portal	Integrated data (genome, EST, marker, map, phenotype, genotype)	Genome, genetic resources, biodiversity, evolution, phylogeny and taxonomy of <i>Quercus</i>	EvoTree project at Institut National de la Recherche Agronomique (INRA), France	Ehrenmann <i>et al.</i> , 2014 (URL: https://arachne.pierroton.inra.fr/QuercusPortal/)
HOSTS	Taxonomy data	180,000 records comprising taxonomically 'cleaned' hostplant data for about 22,000 Lepidoptera species drawn from about 1600 published and manuscript sources	Natural Museum of History, UK	Robinson <i>et al.</i> , 2010

NEXTGEN: Next Generation; TRs: tandem repeats; WGS: whole genome sequence; YRCH: Yethapur *Ricinus communis* Hybrid

The exploration of germplasm diversity of mulberry genome would be useful in the molecular characterization of mulberry germplasm for their genetic improvement. Development of these proposed databases would help in future research in mulberry and development of cost-effective and eco-friendly management approaches.

1.4.2.2 Databases of castor

Much of the genomic and molecular studies on castor (*R. communis*- a primary host plant of *S. cynthia*) are completed focusing on its economic significance (DaSilva *et al.*, 2006). This information can be helpful in understanding silkworm and host plant interactions. Databases have therefore been developed and published on castor (Chan *et al.*, 2010; Thakur *et al.*, 2011), however, few of these database URLs are currently not accessible. The databases on Castor include “*Castor Database*”, “*JCVI Castor Bean Genome Database*” and “*CastorDB*”, the information of which is provided in **Table 1.4**. A namesake of Castor database also exists at <http://glbrc.bch.msu.edu/castor/login> hosted and maintained by Michigan State University, but it is not an open source and is accessible to registered users with no option for new registration on the homepage.

1.4.2.3 Databases of papaya

The 3X draft WGS of *C. papaya* Linnaeus (a secondary host plants of *S. cynthia*) was first published in 2008 with genome size of 372 Mb (Arumuganathan & Earle, 1991). It has facilitated the molecular and genetic study of the plant which has industrial and agricultural significance. The databases describing information on papaya include “*Papaya-DB*” and “*CPR-DB*” (Papaya Repeat Database) and their data content is discussed in **Table 1.4** (Nagarajan & Navajas-Pérez, 2014).

1.4.2.4 Databases of *Jatropha*

Much research on *Jatropha curcas* (a secondary host plant of *S. cynthia*) is available but at present, only one open-source database exists for this plant. “Jatropha Genome Database” (current version 4.5), hosts genomic information and DNA markers of *Jatropha* (Sato *et al.*, 2011).

1.4.2.5 Databases of cassava

Besides *Papaya* and *Jatropha*, *Manihot esculenta* (Cassava) is also an important secondary host plant of *S. cynthia*. The databases available for cassava are “Cassava Genome Database”, “Chinese Cassava Genome Database” and “Cassavabase”. These include information on genome, transcriptome, BAC and cDNA libraries, ESTs, SNPs, genetic maps, etc. of Cassava (**Table 1.4**).

1.4.2.6 Databases of Quercus

Quercus or the oak, is a popular timber tree and also a 1° host plant of oak tasar silkworm (*A. proylei*). “Quercus Portal” is the first integrative web resource which deals with almost all facets of *Quercus* data as listed in **Table 1.4** (Ehrenmann *et al.*, 2014). Based on the information it carries, the portal is divided into eleven sub-databases available at [URL https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=fmap](https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=fmap). Besides being highly resourceful, Quercus portal also has a global search bar which allows query search across all the above-mentioned databases.

1.4.2.7 Other plant databases

Apart from the host plant specific databases, several other databases exist which contain genomic, proteomic and taxonomic information of some silkworm host plants. One such database is “HOSTS”, a database of host plants of lepidopteran insects (around 15%) around the world (Robinson *et al.*, 2010). Other plant databases include “PlantGDB”, “Phytozome”, “PLAZA”, “The Chromatin

Database” (ChromDB), “Plant Transcription Factor Database” (PlantTFDB) and “PLANTS” comprising of information related to genome, chromatin linked proteins, transcription factors, ecology, etc. (Dong *et al.*, 2004; Gendler *et al.*, 2008; Goodstein *et al.*, 2012; Jin *et al.*, 2014; Proost *et al.*, 2015; USDA, 2015).

Overall, host plant specific databases are fewer in number than silkworm or rather insect databases. There is a need to bring together the scattered data on these host plants together in one piece.

1.4.3 Pest and pathogen databases

Silkworms in association with host plants inhabit diverse niches and get affected with several viruses, bacteria, fungi and parasites ranging from mutualistic symbiosis to pathogenesis. Study on these organisms is equally essential to understand host-pathogen interactions, studying molecular mechanisms involved in the pathogenesis, host immune response, developing new strategies against infectious pathogens, etc. (Kaito *et al.*, 2005; Ishii *et al.*, 2015). “SilkPathDB” was the first pathogen database developed by State Key Laboratory of Silkworm Genome Biology (SKLSGB) at Southwest University, China (Li *et al.*, 2017). This database deals with genomic and biological data of a variety of silkworm pathogens including fungi, bacteria, virus and microsporidia and is accessible at <http://silkpathdb.swu.edu.cn/>.

Apart from SilkPathDB, a new comprehensive silkworm disease and pest database “SilkDis”, has been mentioned at URL <http://www.btismysore.in/dbase.html>.

1.4.4 Combined databases

In addition to silkworm, host plant and pest/pathogen databases there are few other databases available that comprise of generalized information of seri-

resources. These databases are not only specific to any individual silkworm, plant or pest databases but contain integrated information of these resources. The barcode data of various taxa including insects and their resources can be accessed through Barcode databases such as “*BOLD*” (*The Barcode of Life Data System*), “*Barcode of Life Data Portal*” (*BDP*) and “*Insect Barcode Informática*” (*IBIn*) (Ratnasingham & Hebert, 2007; Sarkar & Trizna, 2011). Similarly, information on various aspects like silkworm species, rearing details, food plants production technologies, weaving information, best sericulture practices, sericulture marketing, etc. is available in *SILKS or Sericulture information linkages and knowledge system* (URL: <http://silks.csb.gov.in/>).

Other web-resources related to proper identification, classification, taxonomy, distribution and geographical information are also available which can play a pivotal role in exploring, identifying and classifying species. These include- Database of Insects and their Food Plants, Universal Chalcidoidea Database, Common Names of Insects Database, Butterflies and moths of North America, Catalogue of Life, Moths of Borneo, INPN, EPPO Global Database, Encyclopedia of Life (EOL) Database, Integrated taxonomic information system (ITIS) database, Butterflies and Moths of the World, Discover Life, Bill's Lepidoptera Photos, Butterflies and Moths, Area de Conservación Guanacaste (ACG) northwestern Costa Rica, The Global Lepidoptera Names Index or LepIndex, etc.

Apart from the above databases, silk related databases like “*Biomat_dBase*”, “*Spatio-temporal database of the Silk Road*” and “*Silk Fabric Specification Database*” have also been reported by different groups to cover the available information related to silk (biomaterial, Silk Road and fabric characteristics) (Kundu, 2012; Yang *et al.*, 2012; Bi *et al.*, 2014). Moreover, during our search, we

found few other databases such as “miRNEST” (miRNA database), “miRBase” (miRNA database), “MEROPS” (peptidase database) and “Pherobase” (database of semiochemicals i.e. pheromones and allelochemicals) which directly or indirectly contain information of few seri-resources (Szcześniak & Makałowska, 2014; Kozomara & Griffiths-Jones, 2014; Rawlings *et al.*, 2014; El-Sayed, 2014).

1.4.5 Need of seri-related databases

The importance of seri-databases cannot be emphasized more. They can play crucial roles in the conservation of silkworm species, especially the wild varieties. For example, *A. assamensis* is a semi-domesticated silkworm which is endemic to North-Eastern part of India, mainly owing to the climatic and environmental conditions of the place. WildSilkBase which provides the complete EST set of *A. assamensis* can assist us in understanding the functional part of its genome. This might help in engineering the organism to be able to survive in unfavorable conditions. Similarly, plant databases providing information on gene and protein sequences, diseases, etc. will facilitate the conservation of host plants, which is further important for conservation of the silkworms.

Among other scientific benefits are studies on complex interactions among silk moths, host plants and microbes as a model system to understand ecological balance; studies to understand genetics of silk materials from different silkworms; development of SNP-based molecular markers to aid in species differentiation; host plant improvement via genetic engineering; development of effective and cheap methodologies for detection and elimination of pest and pathogen infestations in silkworms and host plants; etc. Also, proper documentation of huge information in a secure but accessible place is another reason to have more seri-

related databases in the future. Implementation of informatics and HT technologies can aid in this regards to a great extent as discussed above.

From an economic standpoint, we believe that the databases can indirectly influence the economy of a region which is dependent on silkworms or their host plants for daily income. India and China are some of the sericulture-intensive countries in the world providing employment to around 8 million people, according to International Sericultural Commission.

We have already witnessed the devastating consequences of Colony Collapse Disorder of honey bees in Europe. If an epidemic of pest infestation is to occur in these places, a huge number of families will be hit by it. These seri-bioresource databases will facilitate more studies on understanding the likeliness of such epidemics and development of future strategies to tackle them. Moreover, seri-bioresources will be highly benefited by the databases and it is imperative that the process of developing more databases goes on. New databases can be created for pest and pathogens, geographical locations, silkworm and host plant disease statistics, taxonomy, compounds, pathways, and so on. There is a need of a resource database that provides extensive information of the explored species, their host plants, diseases, pests and predators. Unfortunately, there is no such a database available till now which provides such comprehensive information.

Based on the literature review, following objectives were formed which are divided into two parts- Part A consisting of three chapters and Part B consisting of one chapter as follows-

Part A: Molecular characterization and mitochondrial genome sequencing of selected seri-bioresources of Assam

Chapter 2: Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

Chapter 3: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkmoth endemic to Assam and North-eastern part of India

Chapter 4: Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkmoth from Assam

Part B: Database development related to seri-bioresources

Chapter 5: Development of seri-related web-resources





Part A

Molecular characterization and mitochondrial genome sequencing of selected seri- bioresources of Assam

Chapter 2

Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

CHAPTER 2

Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

2.1 Introduction

Assam occupies a unique place in sericulture scenario for non-mulberry silk (wild or vanya silk) production (Ahmed & Rajan, 2011; Kalita & Dutta, 2014). It has become a popular source of employment and income for many families in rural areas of Assam like Barpeta, Goalpara, Lakhimpur, etc. (De & Das, 2010). Muga, a unique golden silk, is produced by semi-domesticated muga silkworm *Antheraea assama* which is endemic to Assam and North-eastern region of India (Unni *et al.*, 2009). Similarly, white colored eri silk known as “Ahimsa silk” is produced by *Samia ricini* (eri silkworm which is now a domesticated silkworm) (Ahmed & Rajan, 2011). Many wild silkworms belonging to genera *Antheraea*, *Samia*, *Actias*, *Attacus* and *Cricula* have been recorded from Assam while there exist diverse species of silk moths which are yet to be explored (Saikia *et al.*, 2017). Non-mulberry silkworms show a great diversity in their physiological and morphological characteristics as well as feeding behavior (Tikader *et al.*, 2013). Their gut is rich in probiotic microorganisms which produce various digestive enzymes required for larval development and improving silk production capacity (Dillon & Dillon, 2004; Bhuyan *et al.*, 2014). They also have widespread industrial applications like textile, biofuel, food, etc. (Kuhad *et al.*, 2011; Singh *et al.*, 2016). Unlike mulberry silkworm, muga and eri silkworms are polyphagous in nature, thriving on three kinds of host plants- primary, secondary and tertiary depending upon their

preference and availability. These host plants have a great diversity in their morphology and leaf content which affect both quality and quantity of silk (Tikader & Kamble, 2010; Tikader *et al.*, 2013). The growth, fecundity and survival rates of the silkworms are affected by various pests, parasitoids and pathogens that infect both silkworms and their host plants (Choudhury, 1981; Subharani & Jayaprakash, 2015; Saikia *et al.*, 2016). Similarly, the host plants of eri and muga are infected by various diseases like leaf blight, leaf spot, brown or grey blight, red rust, anthracnose and root rot resulting in the reduction of total leaf yield (Das *et al.*, 2003; Paliwal & Paliwal, 2015).

Knowledge on pests and disease causing pathogens is thus necessary for proper management and conservation of these resources. Conservation of seri-resources will further help in protecting the seri-ecosystem and utilizing them for various commercial applications like medicine, textile, cosmetics, etc. Considering the importance of sericulture in employment security, income generation and poverty alleviation, it is necessary to explore and understand silkmoths and their interacting partners. Identification and classification of these resources through morphological attributes is time consuming and face problems as these characters may vary with environment (Packer *et al.*, 2009). Therefore, to catalogue these resources, DNA barcoding, a DNA based taxonomic characterization tool is extremely useful (Hebert *et al.*, 2003). This tool has provided better taxonomic resolution than that achieved through morphological studies as it overcomes the morphological based limitations. DNA barcoding showed promising results in many insect groups including Coleoptera, Diptera, Hemiptera and particularly Lepidoptera (Dai *et al.*, 2012; Taylor & Harris, 2012; Karthika *et al.*, 2016). The ability of mitochondrial COI sequences to correctly identify lepidopteran species

was primarily established for five families of moths (Hebert *et al.*, 2003). Studies revealed the use of cytochrome c oxidase I (COI) as DNA barcodes to effectively discriminate among species in three families of Lepidoptera- HesperIIDae (skipper butterflies), Sphingidae (sphinx moths), and Saturniidae (wild silk moths). Around 97.9% of 521 species were unambiguously distinguishable from other species as their barcode sequences formed distinct, non-overlapping clusters in a neighbor-joining (NJ) analysis (Hajibabaei *et al.*, 2006). Mitochondrial DNA has also been used to identify and study phylogenetic relationships among the members of Bombycidae family, few members of Saturniidae family and some pest species (Arunkumar *et al.*, 2006; Li *et al.*, 2009; Jalali *et al.*, 2015). Also, DNA barcode markers like *matK* (encoding maturase K), *rbcL* (encoding Ribulose bis-phosphate carboxylase), ITS (Internal transcribed spacer) consisting of ITS1-5.8S ribosomal RNA gene-ITS2), *trnH-psbA* spacer, etc. have been extensively explored in the differentiation of plant species and their phylogenetic studies (Kress *et al.*, 2005; Hollingsworth *et al.*, 2011; Liu *et al.*, 2017). Molecular characterization and phylogenetic analysis has been reported in *Morus*, *Ricinus* and *Persea* species using markers like *ISSR*, *RAPD*, ITS, *matK*, etc. (Enan *et al.*, 2012; Rahman *et al.*, 2012; Zeng *et al.*, 2015). Similarly, DNA barcoding is also effectively used in species identification of fungal and bacterial taxa using barcode regions like ITS, *18S rDNA*, *16S rDNA*, etc. (Chowdappa *et al.*, 2009; Swathi *et al.*, 2015). Many studies revealed the diversity and species richness of silkworms, their host plants, and nutritional and medicinal properties of host plants. However, only few studies report the molecular identification and characterization of silkmths and their resources.

In the present study, we have used various DNA barcode markers and studied their applicability in discriminating few seri-resources of Assam. Further, phylogenetic trees were inferred based on maximum likelihood (ML) method using different markers to study their evolutionary relationships. We expect this study to be useful for the researchers working in the field of sericulture to explore the resources for future studies.

2.2 Materials and methods

2.2.1 Sample collection

Eleven samples including silkworms, their host plants and pests were collected from different locations of Assam, India (Boko, Nongpoh, Khanapara and Jorhat) as given in **Table 2.1**.

Table 2.1 Collection of different silkworm resources and their corresponding location.

S. No.	Silkworm	Sample type	Collection Place	Sample type/ Voucher
1	<i>Samia ricini</i> Borduar strain (Eri)	Eri silkworm	Central Silk Board (CSB), Assam	Egg
2	<i>Antheraea assama</i> Nongpoh strain (Muga)	Muga silkworm	Boko of Kamrup district of Assam	Larva
3	<i>Persea bombycina</i> (King ex hoof. F.) Kost. Or <i>Machilus bombycina</i> (Som)	Som (1° host plant of Muga)	Flora of Assam (Boko, kamrup District)	Leaves/ 17705
4	<i>Litsea monopetala</i> (Roxb.) Pers. Syn. <i>L. polyantha</i> Juss (Soalu)	Soalu (1° host plant of Muga)	Flora of Assam (Boko, Kamrup District)	Leaves/ 17704
5	<i>Litsea cubeba</i> (Lour) Syn. <i>L. citrata</i> BL. Pers. (Mejankari)	Mejankari (2° host plant of Muga)	Flora of Assam (Lahdoigarh, Jorhat)	Leaves/ 917723
6	<i>Litsea salicifolia</i> (Roxb. Ex. Nees.) Hook. F. (Dighloti)	Dighloti (2° host plant of Muga)	Lahdoigarh, Jorhat	Leaves/ 917722
7	<i>Evodia fraxinifolia</i> Hk. F. (Payam)	Payam (2° host plant of Muga)	Lahdoigarh, Jorhat	Leaves/ 917724

8	<i>Ricinus communis</i> strain MCB1 (Red Castor)	Eri) Red castor (1° host plant of Eri)	Flora of Assam	Leaves/ 17773
9	<i>Zizyphus jujuba</i> (Ber)	Ber (3° host plant of Muga)	IITG, Assam	Leaves/
10	Red ants	Pest	Boko, Assam	CMERI-Rd-001
11	Weevils	Pest	Boko, Assam	CMERI-Wv-001

The leaves of infected host plants (*L. monopetala*, *P. bombycina* and *L. salicifolia*) were also collected for barcoding studies in order to identify the disease causing organism (bacteria or fungi). Herbarium of each plant sample was prepared and the voucher specimens were deposited in the herbarium of Botany department, Gauhati University (GUBH). The collected samples were photographed (**Figure 2.1**), documented and properly preserved in 70-95% ethanol at -80°C for future studies. A schematic representation of the workflow followed for barcoding is depicted in **Figure 2.2**.

2.2.2 Genomic DNA extraction

Total DNA was extracted from silk glands/pupa/tissue of collected silkworms and their pests using SDS (Sodium dodecyl sulphate) based DNA extraction protocol with slight modifications (Astruc *et al.*, 2004; Singh *et al.*, 2011). In this protocol, tissue was ground in liquid nitrogen followed by incubation in lysis buffer (100 mM Tris-HCl buffer, pH 8.0, 25 mM EDTA, 2.5% (w/v) SDS, 1.4 M NaCl, 0.8 % (v/v) β -mercaptoethanol, 100 μ g/ ml Proteinase K) at 65 °C for 1 hour. The lysate was subjected to extraction with an equal volume of phenol-chloroform-isoamyl alcohol (25:24:1) and subsequently with chloroform-isoamyl alcohol (24:1). The upper aqueous layer containing DNA was transferred to a fresh tube with subsequent addition of 3M sodium acetate (1/4th volume) and chilled absolute ethanol (2

volumes) to precipitate DNA by keeping it for incubation at -20 °C. The precipitated DNA (DNA pellet) was then separated by centrifugation, washed with chilled 70% (v/v) ethanol and air dried for 2-4 hours. The centrifugation speed and time ranged from 10,000-12,000 rpm for 15-20 minutes. The dried pellet was finally resuspended in 50-100 µl of TE buffer (10 mM Tris-HCl buffer, pH 8.0; 1 mM EDTA, pH 8.0; 100 µg/ ml RNase A) and incubated at 37 °C for 1 hour.

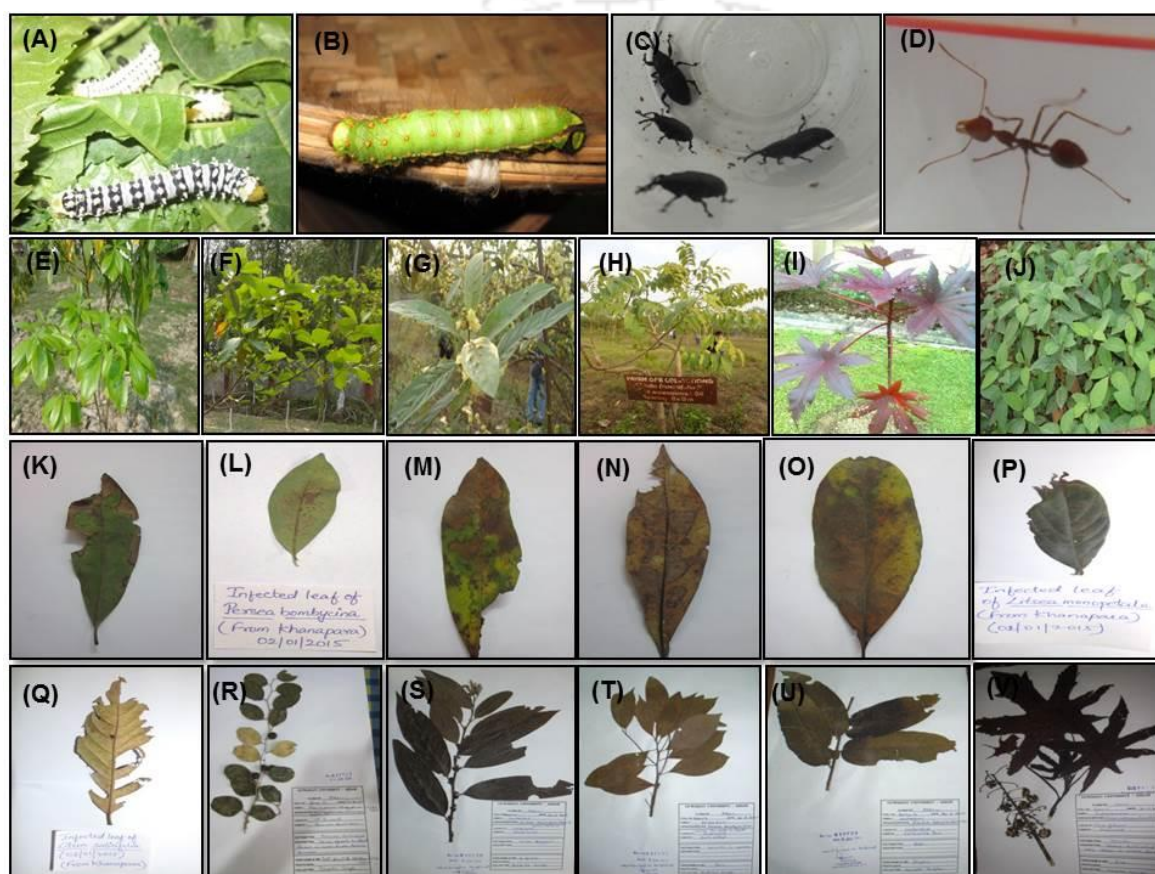


Figure 2.1 Samples collected from different locations of Assam. (A) Eri silkworm (*S. ricini*) (B) Muga silkworm (*A. assama*) (C) Weevils (D) Red ant (E) Som (*P. bombycina*) (F) Soalu (*L. monopetala*) (G) Dighloti (*L. salicifolia*) (H) Payam (*E. fraxinifolia*) (I) Red castor (*R. communis* MCB1 strain) (J) Mejankari (*L. cubeba*) (K-N) Infected *P. bombycina* leaves (O-P) Infected *L. monopetala* leaves (Q) Infected *L. salicifolia* leaf (R-V) Herbarium sheets of the collected host plants.

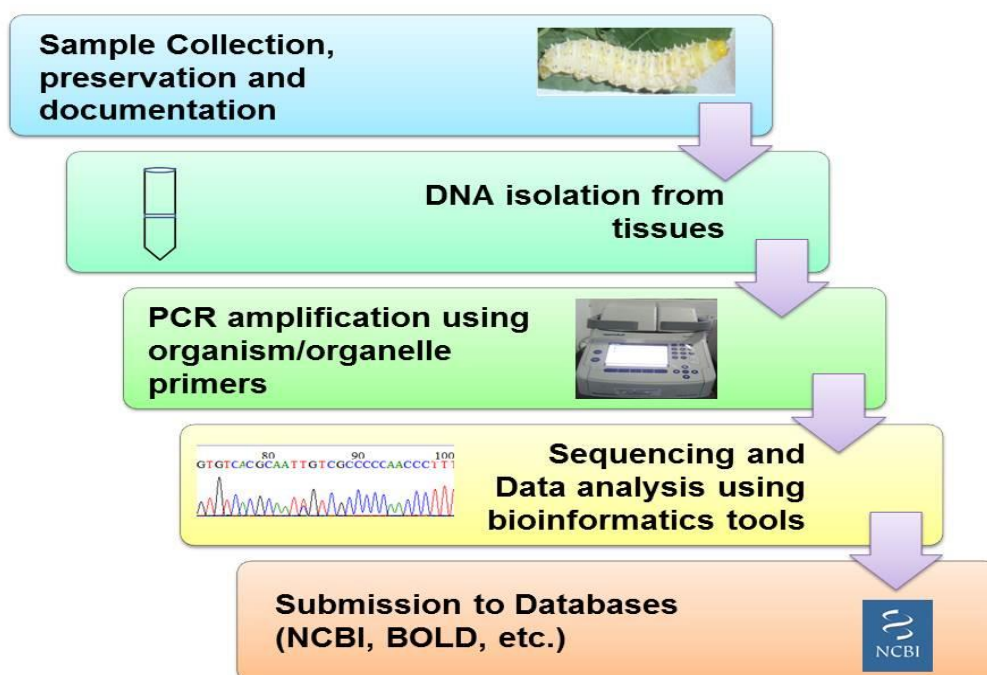


Figure 2.2 A schematic representation of the workflow for DNA barcoding.

For plants, total DNA was isolated from the leaf tissues through CTAB (cetyltrimethylammonium bromide) extraction method followed by chloroform-isoamylalcohol (C: I) extraction (Vinod, 2004). About 0.2-0.3 g of individual leaves were ground in pestle and mortar and homogenized in 5 ml of CTAB (Cetyl trimethylammonium bromide) extraction buffer [2% (w/v) CTAB, 1 M Tris (pH 8.0), 0.5 M EDTA (pH 8.0), 5 M NaCl, 1% (w/v) PVP (Polyvinylpyrrolidone), 0.6-0.8% (v/v) β -mercaptoethanol] followed by incubation at 65 °C for 1 hour in water bath with vigorous shaking. The homogenized sample was extracted twice with equal volume of chloroform: isoamylalcohol (24:1) forming two layers. The upper aqueous layer containing DNA was transferred to a fresh tube with subsequent addition of 3 M sodium acetate (or 5 M Potassium acetate) pH 5.2 (1/4th volume) and isopropanol (2 volumes). The sample was kept for overnight incubation at -20 °C and precipitated DNA was then separated by centrifugation at 12,000 rpm for 15-20 minutes at 4°C. DNA pellet thus obtained was washed with chilled 70%

(v/v) ethanol and air dried for 2 hours to overnight. The dried pellet was finally resuspended in 50-100 µl of TE buffer and incubated at 37 °C for 1 hour.

To isolate muga silkworm gut microflora, the larva was surface sterilized with 70% ethanol and dissected followed by isolation and chopping of gut under aseptic conditions. The chopped gut was then incubated in sterile ice cold 0.85% (w/v) NaCl solution at 37 °C for 30 minutes to 1 hour. The sample was serially diluted, spread on nutrient agar plates and kept for incubation at 37 °C for 24-48 hours. The bacterial colonies observed were picked and sub-cultured several times to obtain a pure culture (Ramesh *et al.*, 2010; Anand *et al.*, 2010). Single bacterial colonies thus obtained were inoculated in liquid nutrient broth for genomic DNA isolation. Similarly, infected leaves of host plants were washed several times with sterile distilled water, treated with sodium hypochlorite (2 min) and washed with distilled water. The leaves were then cut into small pieces under sterile conditions, placed on potato dextrose agar (PDA) plates and incubated at 28-30 °C for 48-72 hrs. The fungal colonies thus obtained were sub-cultured till single fungal colonies were obtained. The single isolated fungal colonies were finally grown in potato dextrose broth for genomic DNA isolation (Quintana-Obregón *et al.*, 2013). The extraction of DNA from bacterial and fungal colonies was performed using HiPurA™ Bacterial Genomic DNA Purification Kit and standard CTAB method, respectively (Lee *et al.*, 1988). The integrity of DNA extracted by the above methods was confirmed through 0.8% (w/v) agarose gel electrophoresis after staining with ethidium bromide (EtBr) prior to PCR amplification.

2.2.3 PCR amplification and sequencing of barcode regions

The DNA barcode regions were amplified from the extracted DNA using different universal primer sets (*COI*, *matK*, *rbcL*, *trnH-psbA*, ITS, ITS2 and 16S *rRNA*) as

listed in **Table 2.2**. Annealing temperatures for the primers were optimized by gradient PCR amplification at different temperature range. For silkworms, PCR reaction was performed in a 50 µl volume reaction mixture- 10 µl of 5X Phusion® HF Buffer, 1 µl of 10 mM Fusion dNTP Solution Mix (Size: 8 µmol of each, conc: 10 mM each nt), 2.5 µl each of 10 µM forward and reverse primer, 0.5 µl of Phusion® high fidelity DNA polymerase (Catalogue No. M0530S; Size: 100 units, Conc: 2,000 units/ml), 4.0 µl of total DNA and 29.5 µl of water. The PCR reaction mixture (50 µl) for other samples consisted of 25 µl of 2X Sigma ReadyMix Taq PCR Reagent Mix (1.5 units Taq DNA polymerase, 10 mM Tris-HCl, 50 mM KCl, 1.5 mM MgCl₂, 0.001% gelatin, 0.2 mM dNTP, stabilizers; Catalogue No. P4600), 1.5 µl each of 10 µM forward and reverse primer, 5 µl (about 20 ng) of total DNA and 17 µl of water. The integrity and concentration of the amplified products were checked through 1.5%- 2% (w/v) agarose gel electrophoresis and NanoDrop Spectrophotometer, respectively. The desired PCR products (amplicons) with concentration >50 ng/µl were gel purified using Sigma gel extraction kit and sent for sequencing to SciGenom Labs Private Ltd. Cochin, Kerala. The sequencing was performed through Sanger's Dideoxy Sequencing method using ABI 3130 XL DNA sequencer.

Table 2.2 List of primers used in the study.

Gene/barcode region	Primers	References
COI (650 bp fragment from 5' end)	TY-J-1460F: 5'-TACAATTTATCGCCTAAACTTCAGCC-3'	Yukuhiro <i>et al.</i> , 2011
	CI-N-2191R: 5'-CCCGGTAAAATTAAAATATAAACTTC-3'	
	LCO1490F: 5'-GGTCAACAAATCATAAAGATATTGG-3'	Szalanski <i>et al.</i> , 2010
	HCO2198R:	

	5'-TAAACTTCAGGGTGACCAAAAAATCA-3'	
matK (600-900 bp)	472F: 5'- CCCRTYCATCTGGAAATCTTGGTTC-3'	Enan et al., 2012
	1248R: 5'- GCTRTRATAATGAGAAAGATTTCTGC-3'	
	390F: 5'-CGATCTATTCAATATTTC-3'	Chen et al., 2010; Bafeel et al., 2011
	1326 R: 5'-TCTAGCACACGAAAGTCGAAGT-3'	
	2.1 F: 5'-CCTATCCATCTGGAAATCTTAG-3'	Bafeel et al., 2011
	5R: 5'-GTTCTAGCACAAGAAAGTCG-3'	
rbcl (700-800 bp)	LaF: 5'- ATGTCACCACAAACAGAGACTAAAGC-3'	Group et al., 2009; Bafeel et al., 2011
	LaR: 5'-GTAAAATCAAGTCCACCRCG-3'	
	1F: 5'-ATGTCACCACAAACAGAAAC-3'	Bafeel et al., 2011
	724R: 5'-TCGCATGTACCTGCAGTAGC-3'	
psbA-trnH (varies from 400-1000 bp)	psbAF: 5'- GTTATGCATGAACGTAATGCTC-3'	Group et al., 2009
	trnHR: 5'- CGCGCATGGTGGATTACAATCC-3'	
ITS (varies from 700-1153 bp)	1F: 5'- TCCGTAGGTGAACCTGCGG-3'	Chen et al., 2010
	4R: 5'- TCCTCCGCTTATTGATATGC-3'	
ITS2 (200-400 bp)	S2F: 5'-ATGCGATACTTGGTGTGAAT-3'	Chen et al., 2010
	S3R: 5'-GACGCTTCTCCAGACTACAAT-3'	
16S rRNA (900-1400 bp)	8F: 5'-AGAGTTTGATCCTGGCTCAG-3'	Liang et al., 2009
	1492R: 5'-TACGGTTACCTTGTTACGACTT-3'	

2.2.4 Sequence analysis and submission to database

The electropherogram (.ab1 file) obtained after sequencing for each region was trimmed using BioEdit tool (Hall, 1999). The sequences generated after trimming were further analyzed using NCBI-BLAST (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>)

to check the specificity of primers and to determine sequence homology. Multiple sequence alignment of the sequences was performed using other bioinformatical tools like Clustal Omega and Muscle (Thompson *et al.*, 2002; Edgar, 2004). NCBI-ORF Finder was used to obtain open reading frame without stop codons for protein coding genes (<https://www.ncbi.nlm.nih.gov/orffinder/>). The processed sequences were properly annotated using NCBI Sequin tool and finally submitted to NCBI GenBank database.

2.2.5 Phylogenetic analysis

For phylogenetic analysis, different barcode sequences showing similarity with the published sequences were retrieved from NCBI GenBank database (based on BLAST results) taking organisms from other order/family as outgroups. For instance, *COI* sequence of *Bombyx mori* belonging to different family of Bombycoidea superfamily (Lepidoptera order) was used as outgroup. Similarly, *trnH-psbA* spacer sequence of *Litsea verticillata* belonging to family Lauraceae was taken as outgroup in the study. Organisms considered for phylogenetic analysis in the present study are listed in **Table 2.3 (A-I)**.

Table 2.3 List of organisms retrieved from NCBI GenBank for phylogenetic analysis.

(A) Organisms selected for phylogenetic analysis of silkworms based on *COI* gene.

S. No.	GenBank Accession	Species Name	Classification
1	KJ676596.1	<i>Antheraea frithi</i>	Lepidoptera: Bombycoidea: Saturniidae
2	KJ740437.1	<i>Antheraea frithi</i>	Lepidoptera: Bombycoidea: Saturniidae
3	KC968964.1	voucher IBSDIRWSA1 <i>Antheraea mylitta</i>	Lepidoptera: Bombycoidea: Saturniidae
4	KU379695.1	<i>Antheraea assama</i>	Lepidoptera: Bombycoidea: Saturniidae
5	AB533559.1	<i>Antheraea formosana</i>	Lepidoptera: Bombycoidea:

			Saturniidae
6	HQ579941.1	<i>Antheraea paniki</i> voucher BC-ULP0432	Lepidoptera: Bombycoidea: Saturniidae
7	JN215366.1	<i>Samia cynthia ricini</i>	Lepidoptera: Bombycoidea: Saturniidae
8	AB015866.1	<i>Samia ricini</i>	Lepidoptera: Bombycoidea: Saturniidae
9	KX862929.1	<i>Samia canningi</i> voucher NIBGE MOT- 03452	Lepidoptera: Bombycoidea: Saturniidae
10	AB015865.1	<i>Samia 68ynthia pryeri</i>	Lepidoptera: Bombycoidea: Saturniidae
11	GU360733.1	<i>Samia cynthia ricini</i> strain Guangxi	Lepidoptera: Bombycoidea: Saturniidae
12	AB533594.1	<i>Samia wangi</i>	Lepidoptera: Bombycoidea: Saturniidae
13	HQ579933.1	<i>Samia abrerai</i> voucher BC-ULP0421	Lepidoptera: Bombycoidea: Saturniidae
14	GU703324.1	<i>Samia kalimantanensis</i> voucher BC-ULP0155	Lepidoptera: Bombycoidea: Saturniidae
15	HQ579928.1	<i>Samia naumanni</i> voucher BC-ULP0412	Lepidoptera: Bombycoidea: Saturniidae
16	HQ579940.1	<i>Samia kohlli</i> voucher BC-ULP0430	Lepidoptera: Bombycoidea: Saturniidae
17	AY605253.1	<i>Phylosomia ricini</i>	Lepidoptera: Bombycoidea: Saturniidae
18	HQ599816.1	<i>Samia watsoni</i> voucher SNB 1872	Lepidoptera: Bombycoidea: Saturniidae
19	AF167279.1	<i>Bombyx mori</i> strain C108	Lepidoptera: Bombycoidea: Bombycidae (Outgroup)

(B) Organisms selected for phylogenetic analysis of weevils based on *COI* gene.

S. No.	GenBank Accession	Species Name	Classification
1	KP233792.1	<i>Odoiporus longicollis</i> voucher RO_ODLO1	Coleoptera: Cucujiformia: Curculionidae
2	KU863027.1	<i>Odoiporus longicollis</i> isolate OHS01	Coleoptera: Cucujiformia: Curculionidae
3	KU863028.1	<i>Odoiporus longicollis</i> isolate OPV07	Coleoptera: Cucujiformia: Curculionidae
4	KR494261.1	<i>Rhynchophorus</i> <i>palmarum</i> isolate RP008	Coleoptera: Cucujiformia: Curculionidae
5	KR494260.1	<i>Rhynchophorus</i> <i>palmarum</i> isolate RP007	Coleoptera: Cucujiformia: Curculionidae

6	FJ867843.1	<i>Heilipodus argentinicus</i> voucher BTOLDDM0368	Coleoptera: Cucujiformia: Curculionidae
7	KF737831.1	<i>Xenohammus</i> <i>bimaculatus</i> voucher HNAU.PP 052	Coleoptera: Cucujiformia: Chrysomeloidea
8	KM443549.1	<i>Peritelus sphaeroides</i> voucher GBOL_Col_FK_0257	Coleoptera: Cucujiformia: Curculionidae
9	KY912926.1	<i>Sitophilus oryzae</i> voucher H-SC-EM	Coleoptera: Cucujiformia: Curculionidae
10	KM445312.1	<i>Peritelus sphaeroides</i> voucher GBOL01886	Coleoptera: Cucujiformia: Curculionidae
11	KY084077.1	<i>Protapion nigrirtase</i>	Coleoptera: Cucujiformia: Brentidae
12	HQ891445.1	<i>Eustylus hybridus</i>	Coleoptera: Cucujiformia: Curculionidae
13	KF309736.1	<i>Apotolamprus metallicus</i> voucher Ara.sp1.1.2	Coleoptera: Scarabaeiformia: Scarabaeidae
14	LC072650.1	<i>Lyctus brunneus</i>	Coleoptera: Bostrichiformia: Bostrichidae (Outgroup)

(C) Organisms selected for phylogenetic analysis of red ant based on *COI* gene.

S. No.	GenBank Accession	Species Name	Classification
1	JQ681064.1	<i>Oecophylla smaragdina</i> isolate GDXT0104	Hymenoptera: Vespoidea: Formicidae
2	DQ353382.1	<i>Oecophylla smaragdina</i> voucher RA0264	Hymenoptera: Vespoidea: Formicidae
3	JQ344565.1	<i>Hymenoptera</i> sp. <i>XSBNHap_9</i>	Hymenoptera
4	KC685044.1	<i>Oecophylla smaragdina</i> voucher AAI_73	Hymenoptera: Vespoidea: Formicidae
5	KC685043.1	<i>Oecophylla smaragdina</i> voucher AAI_15	Hymenoptera: Vespoidea: Formicidae
6	KM348201.1	<i>Oecophylla smaragdina</i> voucher CSM0633	Hymenoptera: Vespoidea:

7	AB019426.1	<i>Oecophylla smaragdina</i>	Formicidae Hymenoptera: Vespoidea:
8	KC335785.1	<i>Cephalotes auricomus</i> voucher MLA01	Formicidae Hymenoptera: Vespoidea:
9	KX054985.1	<i>Tetramorium bicarinatum</i> voucher sc_05959	Formicidae Hymenoptera: Vespoidea:
10	MF071555.1	<i>Cataglyphis isis</i> isolate MZR9	Formicidae Hymenoptera: Vespoidea:
11	MF071530.1	<i>Cataglyphis israelensis</i> isolate RAM1	Formicidae Hymenoptera: Vespoidea:
12	KR802618.1	<i>Eumenes crucifera</i> voucher BIOUG01048- B12	Formicidae Hymenoptera: Vespoidea: Vespidae (Outgroup)

(D) Organisms selected for phylogenetic analysis of host plants based on ITS2 region.

S. No.	GenBank Accession	Species Name	Classification
1	KX259274.1	<i>Ricinus communis</i> isolate JD0021MT003	Malpighiales: Euphorbiaceae
2	KX259273.1	<i>Ricinus communis</i> isolate JD0021MT002	Malpighiales: Euphorbiaceae
3	KR082766.1	<i>Ricinus communis</i>	Malpighiales: Euphorbiaceae
4	KF454062.1	<i>Ricinus communis</i> voucher 654021120831005	Malpighiales: Euphorbiaceae
5	KF454061.1	<i>Ricinus communis</i> voucher 654025120731005	Malpighiales: Euphorbiaceae
6	KF454059.1	<i>Ricinus communis</i> voucher 652822-120702- 113	Malpighiales: Euphorbiaceae
7	KF186448.1	<i>Ricinus communis</i> voucher KUN:GBOWS YNS0215	Malpighiales: Euphorbiaceae
8	KJ130048.1	<i>Ricinus communis</i>	Malpighiales: Euphorbiaceae
9	KJ130047.1	<i>Ricinus communis</i>	Malpighiales: Euphorbiaceae
10	KF815504.1	<i>Chrozophora tinctoria</i> voucher KSUFS388	Malpighiales: Euphorbiaceae

11	AY541602.1	<i>Viola orientalis</i>	Malpighiales: Violaceae
12	AB441755.1	<i>Pyrus communis</i>	Rosales: Rosaceae (Outgroup)

(E) Organisms selected for phylogenetic analysis of host plants based on *matK* region.

S. No.	GenBank Accession	Species Name	Classification
1	KP089335.1	<i>Tetradium daniellii</i> voucher BOP010405	Sapindales: Rutaceae
2	KR531568.1	<i>Tetradium glabrifolium</i> isolate BB0474	Sapindales: Rutaceae
3	KP089168.1	<i>Phellodendron chinense</i> voucher BOP010196	Sapindales: Rutaceae
4	EF489105.1	<i>Euodia hupehensis</i> voucher Sandro Wagen 6 (Z)	Sapindales: Rutaceae
5	AF244402.1	<i>Lindera obtusiloba</i>	Lurales: Lauraceae
6	AB924731.1	<i>Cinnamomum polyadelphum</i> voucher: KYUM<JPN>:60	Lurales: Lauraceae
7	EU153844.1	<i>Lindera benzoin</i> voucher K16947 (ANDES)	Lurales: Lauraceae
8	EU153841.1	<i>Lauraceae sp. 3 DRP38</i>	Lurales: Lauraceae
9	AB259089.1	<i>Litsea resinosa</i> voucher: MAK:XX.A. 42	Lurales: Lauraceae
10	AF244395.1	<i>Litsea umbellata</i>	Lurales: Lauraceae
11	AB259082.1	<i>Litsea grandis</i> voucher: MAK:IZ 2042	Lurales: Lauraceae
12	AB259078.1	<i>Litsea firma</i> voucher: MAK:XX.B. 171	Lurales: Lauraceae
13	AB259086.1	<i>Litsea mappacea</i> voucher: MAK:VII.C. 51B	Lurales: Lauraceae
14	HQ427403.1	<i>Litsea elongata</i>	Lurales: Lauraceae
15	AJ247171.1	<i>Litsea japonica</i> voucher Rohwer 156 (MJG)	Lurales: Lauraceae
16	AJ247154.2	<i>Cinnamomum camphora</i> voucher Rohwer 164 (MJG)	Lurales: Lauraceae
17	KT348516.1	<i>Machilus yunnanensis</i>	Lurales: Lauraceae
18	KJ510886.1	<i>Machilus kwangtungensis</i>	Lurales: Lauraceae

19	HQ427418.1	<i>Machilus pauhoi</i>	Lurales: Luraceae
20	HQ415335.1	<i>Machilus chinensis</i>	Lurales: Luraceae
21	HQ415333.1	<i>Machilus phoenicis</i>	Lurales: Luraceae
22	HQ427400.1	<i>Phoebe shearerii</i>	Lurales: Luraceae
23	AB259098.1	<i>Machilus rimosa</i> voucher: MAK:VII.C. 58D	Lurales: Luraceae
24	KM030011.1	<i>Morus alba</i>	Rosales: Moraceae (Outgroup)

(F) Organisms selected for phylogenetic analysis of host plants based on *rbcL* gene.

S. No.	GenBank Accession	Species Name	Classification
1	KT348516.1	<i>Machilus yunnanensis</i>	Lurales: Luraceae
2	KP094423.1	<i>Machilus breviflora</i> isolate SCBGP145_1	Lurales: Luraceae
3	KJ439990.1	<i>Machilus robusta</i>	Lurales: Luraceae
4	KJ439985.1	<i>Machilus kwangtungensis</i>	Lurales: Luraceae
5	KF569893.1	<i>Machilus grijsii</i>	Lurales: Luraceae
6	KP094782.1	<i>Machilus gamblei</i>	Lurales: Luraceae
7	GQ436747.1	<i>Tetradium ruticarpum</i> voucher PS1614MT01	Sapindales: Rutaceae
8	GQ436739.1	<i>Zanthoxylum bungeanum</i> voucher PS1599MT01	Sapindales: Rutaceae
9	KJ922006.1	<i>Zanthoxylum avicennae</i>	Sapindales: Rutaceae
10	KT153018.1	<i>Zanthoxylum piperitum</i>	Sapindales: Rutaceae
11	FN599466.1	<i>Tetradium meliifolium</i>	Sapindales: Rutaceae
12	AF193970.1	<i>Laurus nobilis</i>	Lurales: Luraceae
13	LJU06843	<i>Litsea japonica</i>	Lurales: Luraceae
14	KT626733.1	<i>Litsea calicaris</i>	Lurales: Luraceae
15	KF022500.1	<i>Ricinus communis</i> isolate 01	Malpighiales: Euphorbiaceae
16	AY794915.1	<i>Ricinus communis</i>	Malpighiales: Euphorbiaceae

17	KX951434.1	<i>Ricinus communis</i> clone R8	Malpighiales: Euphorbiaceae
18	KM273034.1	<i>Ricinus communis</i> voucher MOSEL 265	Malpighiales: Euphorbiaceae
19	EU128738.1	<i>Litsea monopetala</i> voucher DUH 13263	Lurales: Lauraceae
20	KJ439993.1	<i>Litsea baviensis</i>	Lurales: Lauraceae
21	KJ439986.1	<i>Litsea cubeba</i>	Lurales: Lauraceae
22	JN114823.1	<i>Morus alba</i>	Rosales: Moraceae (Outgroup)

(G) Organisms selected for phylogenetic analysis of host plants based on *trnH-psbA* region.

S. No.	GenBank Accession	Species Name	Classification
1	KU236003.1	<i>Ziziphus mauritiana</i> voucher 06AZ520151116065	Rosales: Rhamnaceae
2	KU236004.1	<i>Ziziphus mauritiana</i> voucher 06AZ820151109011	Rosales: Rhamnaceae
3	HE602473.1	<i>Ziziphus lotus</i>	Rosales: Rhamnaceae
4	HG005114.1	<i>Ziziphus cf. fungii</i> MR-2013	Rosales: Rhamnaceae
5	EU075111.1	<i>Ziziphus incurva</i>	Rosales: Rhamnaceae
6	KU236010.1	<i>Ziziphus jujuba</i> var. <i>spinosa</i> voucher 06AZ820151126372	Rosales: Rhamnaceae
7	EU075112.1	<i>Paliurus spina-christi</i>	Rosales: Rhamnaceae
8	KM895386.1	<i>Alphitonia excelsa</i> voucher 1076418180	Rosales: Rhamnaceae
9	KM895104.1	<i>Alphitonia petriei</i> voucher 1076418203	Rosales: Rhamnaceae
10	GQ248363.1	<i>Phyllica arborea</i> voucher JE Richardson 51 (K)	Rosales: Rhamnaceae
11	KU564716.1	<i>Pomaderris notata</i>	Rosales: Rhamnaceae
12	HQ415578.1	<i>Hovenia acerba</i>	Rosales: Rhamnaceae
13	HG005084.1	<i>Berchemia hirtella</i>	Rosales: Rhamnaceae
14	KU564715.1	<i>Pomaderris clivicola</i>	Rosales: Rhamnaceae

15	HQ415479.1	<i>Litsea verticillata</i>	Laurales: Lauraceae (Outgroup)
----	------------	----------------------------	--------------------------------------

(H) Organisms selected for phylogenetic analysis of bacteria based on 16S rRNA gene.

S. No.	GenBank Accession	Species Name	Classification
1	JQ659309.1	<i>Micrococcus endophyticus</i> strain L1-101	Micrococcales: Micrococcaceae
2	KF703726.1	<i>Micrococcus</i> sp. EAAG100	Micrococcales: Micrococcaceae
3	KC545358.1	<i>Micrococcus</i> sp. EF1B-B144	Micrococcales: Micrococcaceae
4	KP860533.1	<i>Micrococcus endophyticus</i> strain EAS4E14	Micrococcales: Micrococcaceae
5	MF045105.1	<i>Micrococcus</i> sp. strain 70025	Micrococcales: Micrococcaceae
6	KU922513.1	<i>Micrococcus luteus</i> strain SSW15	Micrococcales: Micrococcaceae
7	KU925163.1	<i>Micrococcus yunnanensis</i> strain AU D9	Micrococcales: Micrococcaceae
8	KT719557.1	<i>Micrococcus aloeverae</i> strain MER_TA_144	Micrococcales: Micrococcaceae
9	KM108306.1	<i>Micrococcus flavus</i> strain PS5	Micrococcales: Micrococcaceae
10	AM158920.1	<i>Micrococcus indicus</i> strain BBQ1	Micrococcales: Micrococcaceae
11	KX087370.1	<i>Bacillus</i> sp. strain AB1	Bacillales: Bacillaceae (Outgroup)

(I) Organisms selected for phylogenetic analysis of fungi based on ITS region.

S. No.	GenBank Accession	Species Name	Classification
1	KX098316.1	<i>Colletotrichum gloeosporioides</i> isolate S2_BA	Glomerellales: Glomerellaceae
2	KC411911.1	<i>Colletotrichum fragariae</i> strain BF51	Glomerellales: Glomerellaceae
3	KX098318.1	<i>Colletotrichum gloeosporioides</i> isolate S4_BA	Glomerellales: Glomerellaceae
4	LC074359.1	<i>Lasiodiplodia theobromae</i>	Botryosphaerales: Botryosphaeriaceae
5	KP721561.1	<i>Lasiodiplodia</i> sp.	Botryosphaerales:

		AQGS49	Botryosphaeriaceae
6	KP721592.1	<i>Lasiodiplodia</i> sp. AQGWD 5	Botryosphaeriales: Botryosphaeriaceae
7	KX631698.1	<i>Lasiodiplodia</i> <i>pseudotheobromae</i> strain HHL31	Botryosphaeriales: Botryosphaeriaceae
8	KU977555.1	<i>Fungal</i> sp. voucher ARIZ:PS0036	Fungi
9	KU720061.1	<i>Pestalotiopsis microspora</i> strain IRNHM-PS1	Xylariales: Sporocadaceae
10	KT377207.1	<i>Neopestalotiopsis</i> <i>clavispora</i> strain AT13	Xylariales: Sporocadaceae
11	KR183786.1	<i>Neopestalotiopsis</i> <i>foedans</i> strain Cg2	Xylariales: Sporocadaceae
12	KT936426.1	<i>Neopestalotiopsis</i> <i>protearum</i> isolate OTTR524	Xylariales: Sporocadaceae
13	KX067861.1	<i>Penicillium oxalicum</i> isolate Liaoning-Rfsb062	Eurotiales: Aspergillaceae
14	KU240500.1	<i>Penicillium oxalicum</i> strain Z3	Eurotiales: Aspergillaceae
15	KC535017.1	<i>Neopestalotiopsis</i> <i>foedans</i>	Xylariales: Sporocadaceae
16	KM520039.1	<i>Pestalotiopsis</i> sp. LS16	Xylariales: Sporocadaceae
17	KJ590132.1	<i>Pestalotiopsis clavispora</i> strain mirt2	Xylariales: Sporocadaceae
18	KX463026.1	<i>Phomopsis phyllanthicola</i>	Diaporthales: Valsaceae
19	KF498871.1	<i>Phomopsis phyllanthicola</i> strain A0658	Diaporthales: Valsaceae
20	KX816029.1	<i>Fungal endophyte</i> isolate LIES548	Fungi
21	KP413076.1	<i>Entorrhiza tenuis</i> voucher KRAM F-57376	Entorrhizales: Entorrhizaceae (Outgroup)

The respective sequences (*COI*, *matK*, *rbcL*, *trnH-psbA*, ITS, ITS2) were aligned through Clustal W and Muscle in Mega 6.0 (Tamura *et al.*, 2013). Optimal model selection was performed for all datasets using Mega 6.0 tool (**Table 2.4**). The phylogenetic tree construction was carried out in Mega 6.0 to study evolutionary relationships among the species based on maximum likelihood (ML) method.

Bootstrap analysis of 1000 replicates was employed to estimate the robustness of clades of the trees by partial deletion of codons with gaps and missing data.

Table 2.4 Models selected for various datasets under study.

Organism	Gene/Barcode region	Model selected
Silkworm	<i>COI</i>	GTR
Pest (<i>Odoiporus longicollis</i>)	<i>COI</i>	GTR+G
Pest (<i>Oecophylla smaragdina</i>)	<i>COI</i>	GTR
Host Plant	<i>trnH-psbA</i>	T92
Host Plant (<i>R. communis</i>)	ITS2	JC
Host Plant	<i>matK</i>	T92
Host Plant	<i>rbcL</i>	T92+G
Bacteria	16S rRNA	K2
Fungus	ITS	K2+G+I

Abbreviations- G: gamma distribution, GTR: General Time Reversible, I: proportion of invariable sites, JC: Jukes-Cantor, K2: Kimura 2-parameter, T92: Tamura 3-parameter.

2.3 Results and discussion

2.3.1 Genomic DNA extraction

Genomic DNA was successfully extracted from the samples- *A. assama*, *S. cynthia*, red ants and weevils using SDS based method. Similarly, genomic DNA in plants namely *R. communis*, *P. bombycina*, *L. salicifolia*, *L. cubeba* and *Z. mauritiana* was successfully extracted by CTAB method. This method was found to be inconvenient in plants such as *L. monopetala* and *E. fraxinifolia* due to the presence of phenolic compounds, polysaccharides, and other secondary metabolites which hindered the extraction of DNA. Modified methods were therefore employed for the successful isolation of DNA from such plants (Sharma *et al.*, 2010). STE/CTAB (Sucrose-Tris-EDTA/CTAB) method proposed by Shepherd & McLay (2011) with slight modifications resulted in the successful isolation of genomic DNA from *E. fraxinifolia*. Similarly, high CTAB-NaCl DNA

extraction method with simple ethanol wash resulted in the isolation of DNA of *L. monopetala* and *E. fraxinifolia* (Kundu *et al.*, 2011). Fungal colonies were isolated from the collected leaves of host plants as depicted in **Figure 2.3**. Genomic DNA of these fungal colonies and bacterial colonies were also successfully extracted. The integrity of the isolated DNA checked by agarose gel electrophoresis was found to be optimal for PCR amplification.

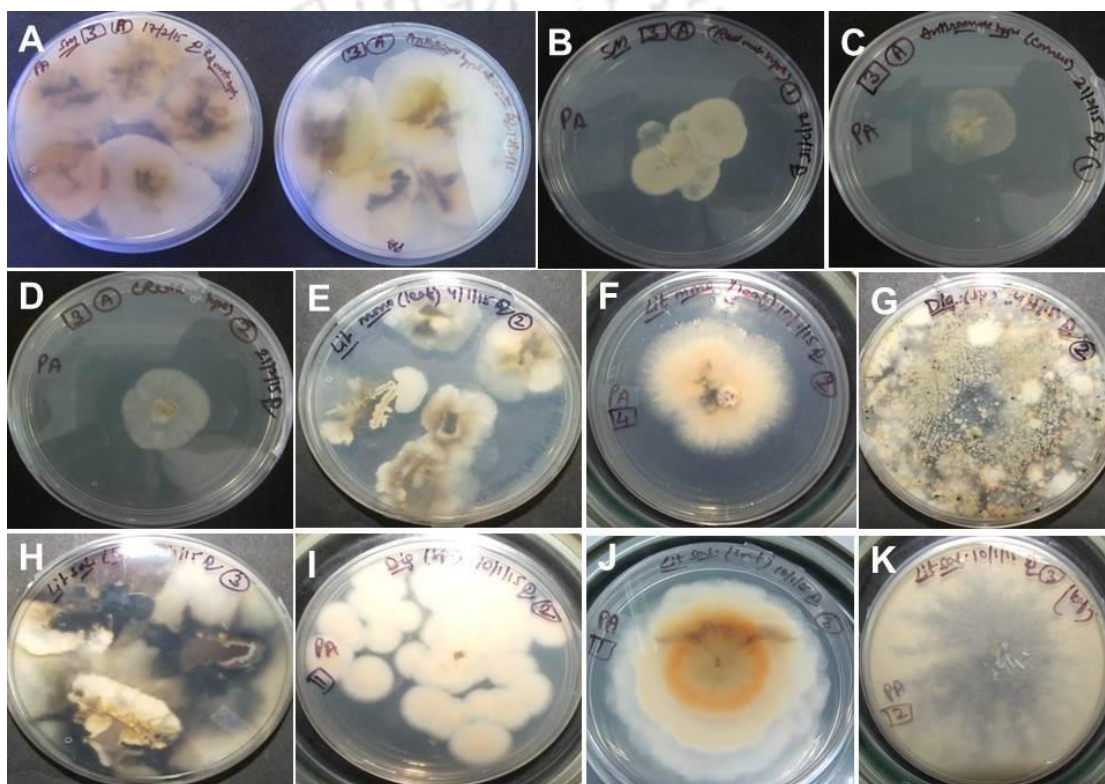


Figure 2.3 Fungus isolates from infected leaves of Muga host plants. (A-D) fungus isolated from *Persea bombycina*/*Machilus bombycina* (E&F) fungus isolated from *Litsea monopetala* (G-K) fungus isolated from *Litsea salicifolia*.

2.3.2 PCR amplification and sequencing of barcode regions

Different primers were tested for their amplification rate in the collected plant samples and the concentration of PCR products checked by NanoDrop was found to be >50 ng/μl desirable for sequencing. The optimal PCR cycling conditions for each region and for each sample is given in **Table 2.5**.

Table 2.5 Optimized PCR cycling conditions for each sample.

Species Name	Barcode region	Primer Set	Optimum PCR cycling conditions
For silkworms & pests			
<i>Antheraea assama</i> strain Nongpoh	COI	TY-J-1460F & CI-N-2191R	96 °C (2 min); 35 cycles of 96 °C (15 sec), 55 °C (15 sec), 68 °C (1 min); 68 °C (5 min)
<i>Samia ricini</i> strain Borduar	COI	TY-J-1460F & CI-N-2191R	96 °C (2 min); 35 cycles of 96 °C (15 sec), 40 °C (15 sec), 68 °C (1 min); 68 °C (5 min)
<i>Oecophylla smaragdina</i>	COI	LCO1490F & HCO2198R	94 °C (2 min); 35 cycles of 94 °C (45 sec), 45.5 °C (45 sec), 72 °C (1 min); 72 °C (5 min)
<i>Odoiporus longicollis</i>	COI	LCO1490F & HCO2198R	94 °C (2 min); 35 cycles of 94 °C (45 sec), 45.5 °C (45 sec), 72 °C (1 min); 72 °C (5 min)
Host Plants			
<i>Machilus bombycina</i>	matK	2.1F & 5R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 47.2 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Machilus bombycina</i>	rbcL	LaF & LaR	95 °C (1 min); 35 cycles of 95 °C (30 sec), 48 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Litsea salicifolia</i>	matK	2.1F & 5R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 50.1 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Litsea salicifolia</i>	rbcL	1F & 724R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 47.5 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Litsea monopetala</i>	matK	2.1F & 5R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 47.8 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Litsea monopetala</i>	rbcL	1F & 724R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 50.3 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Litsea cubeba</i>	matk	2.1F & 5R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 45 °C (30 sec), 68 °C (1 min); 68 °C (5 min)

<i>Litsea cubeba</i>	<i>rbcL</i>	1F & 724R	min) 95 °C (1 min); 35 cycles of 95 °C (30 sec), 50.3 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Ziziphus jujuba</i>	<i>psbA-trnH</i>	psbAF & trnHR	94 °C (5 min); 30 cycles of 94 °C (1 min), 54.6 °C (1 min), 72 °C (1.5 min); 72 °C (7 min)
<i>Ricinus communis</i>	<i>rbcL</i>	1F & 724R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 48 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Ricinus communis</i>	ITS2	2F & 3R	94 °C (5 min); 40 cycles of 94 °C (30 sec), 55.4 °C (30 sec), 72 °C (45 sec); 72 °C (10 min)
<i>Euodia fraxinifolia</i>	<i>rbcL</i>	1F & 724R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 48.1 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
<i>Euodia fraxinifolia</i>	<i>matK</i>	390F & 1326R	95 °C (1 min); 35 cycles of 95 °C (30 sec), 45.7 °C (30 sec), 68 °C (1 min); 68 °C (5 min)
Fungus			
<i>Colletotrichum gloeosporioides</i> strain DS01	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
<i>Colletotrichum gloeosporioides</i> strain DS02	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
<i>Lasiodiplodia theobromae</i> strain DS03	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
<i>Neopestalotiopsis clavispora</i> strain DS05	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
<i>Penicillium oxalicum</i> strain DS06	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
<i>Pestalotiopsis</i> sp. DS07	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C

(7 min)			
<i>Phomopsis phyllanthicola</i> strain DS08	ITS	ITS1 & ITS4	94 °C (5 min); 30 cycles of 94 °C (30 sec), 54.1 °C (30 sec), 72 °C (30 sec); 72 °C (7 min)
Bacteria			
<i>Micrococcus endophyticus</i> strain DS04	16S <i>rRNA</i>	8F & 1492R	94 °C (5 min); 35 cycles of 94 °C (45 sec), 49.4 °C (1 min), 72 °C (1 min); 72 °C (10 min)

COI showed successful amplification in silkworms and pest samples (700-800 bp) using the universal primer sets (TY-J-1460F & CI-N-2191R in silkworms; LCO1490F & HCO2198R in pests). Similarly, *matK*, *rbcL*, *trnH-psbA* and ITS2 primers were found to be suitable for successful amplification of *matK* (600-900 bp), *rbcL* (600-700 bp), *trnH-psbA* (~400bp) and ITS2 (200-400 bp) barcode regions in the collected plants. In addition, ITS was found to be successfully amplified in the isolated fungal species using primer set ITS1F and ITS4R (400-600 bp). Unlike COI primers, the plant specific primers did not work for all the plants. For instance, amplification of *matK* and *rbcL* regions was observed for *Persea*, *Litsea* and *Euodia* species while failed to amplify *Zizyphus* species which showed successful amplification with *trnH-psbA* primer set. The detailed agarose gel electrophoresis results of PCR products are depicted in **Figure 2.4** (silkworms, pests and host plants) and **Figure 2.5** (fungal and bacterial colonies).

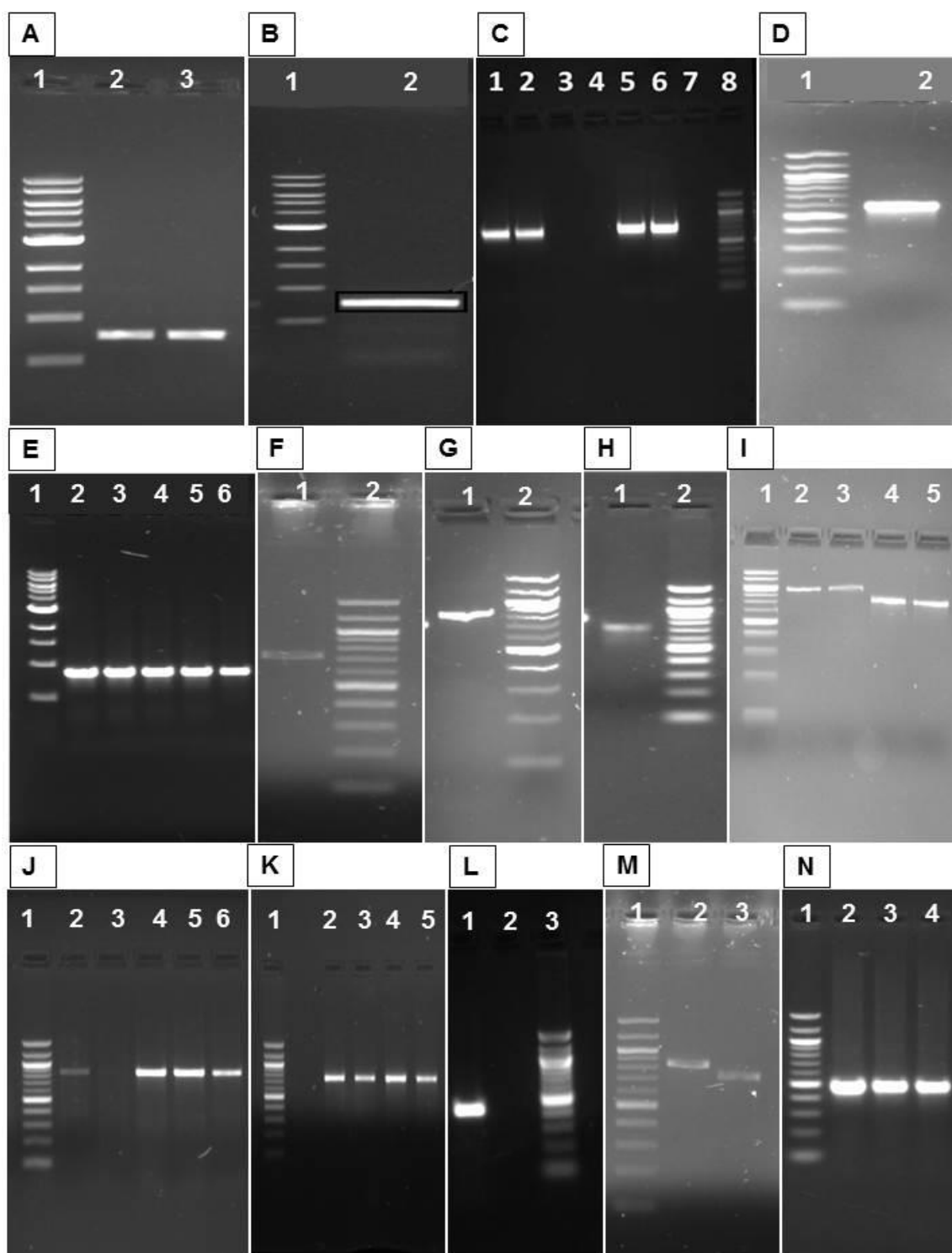


Figure 2.4 Agarose gel electrophoresis of PCR products of isolated DNA of the collected samples. (A) 1.5% agarose gel electrophoresis of PCR product of *S. ricini* DNA sample. Lane 1: 1 Kb DNA ladder, Lane 2 & 3: PCR product of *COI* gene of *S. ricini* using TY-J-1460F and CI-N-2191R primer set; (B) 1.5% agarose gel electrophoresis of PCR product of *A. assama* DNA sample. Lane 1: 1 Kb ladder, Lane 2: PCR product of *COI* gene of *A. assama* (Two wells were clubbed together to obtain good yield of DNA for sequencing) using TY-J-1460F and CI-N-

2191R primer set; (C) 1.5% agarose gel electrophoresis of amplified *COI* gene of Red ants and Weevils. Lane 1&2: Amplified *COI* gene of Red ants using LCO1490F and HCO2198R primer set, Lane 3, 4 & 7: Empty, Lane 5 & 6: Amplified *COI* gene of Weevils using same primer set, Lane 8: 100 bp DNA ladder; (D) 1% agarose gel electrophoresis of amplified *rbcL* gene products of *Persea bombycina*. Lane 1: 100 bp Ladder, Lane 2: PCR product of *rbcL* gene of *Persea bombycina* using *rbcLaF* and *rbcLaR* primer set; (E) 1% agarose gel electrophoresis of amplified *matK* gene products of *Persea bombycina*. Lane 1: 1 Kb DNA ladder, Lane 2-6: PCR product of *matK* gene using *matK* 2.1F and *matK* 5R primer set; (F) 1.5% agarose gel electrophoresis of amplified products of *Litsea monopetala*. Lane 1: PCR product of *rbcL* gene of *Litsea monopetala* using *rbcL* 1F and *rbcL* 724R primer set, Lane 2: 100 bp ladder; (G) 1.5% agarose gel electrophoresis of amplified products of *Litsea monopetala*. Lane 1: PCR product of *matK* gene of *Litsea monopetala* using *matK* 2.1F and *matK* 5R primer set, Lane 2: 100 bp ladder; (H) 1.5% agarose gel electrophoresis of amplified products of *Ricinus communis*. Lane 1: PCR product of *rbcL* gene of Red Castor using *rbcL* 1F and *rbcL* 724R primer set, Lane 2: 100 bp ladder; (I) 1.5% agarose gel electrophoresis of amplified products of *Evodia fraxinifolia*. Lane 1: 100 bp DNA ladder, Lane 2-3: PCR product of *matK* gene of *Evodia fraxinifolia* using *matK* 2.1F and *matK* 5R primer set, Lane 4-5: PCR product of *rbcL* gene of *Evodia fraxinifolia* using *rbcL* 1F, *rbcL* 724R primers; (J) 1.5% agarose gel electrophoresis of amplified products of *Litsea salicifolia*. Lane 1: 100 bp DNA ladder, Lane 2-6: PCR products of *matK* gene of *Litsea salicifolia* using *matK* 2.1F and *matK* 5R primer set; (K) 1.5% agarose gel electrophoresis of amplified products of *Litsea salicifolia*. Lane 1: 100 bp DNA ladder, Lane 2-5: PCR products of *rbcL* gene of *Litsea salicifolia* amplified using *rbcL* 1F and *rbcL* 724R primer set; (L) 1.5% agarose gel electrophoresis of amplified products of *Zizyphus mauritiana*. Lane 1: PCR products of *psbA-trnH* gene of *Z. mauritiana* amplified using *psbAF* and *trnHR* primer set, Lane 2: Empty, Lane 3: 100 bp DNA ladder; (M) 1.5% agarose gel electrophoresis of amplified products of *Litsea citrata*. Lane 1: 100 bp DNA ladder, Lane 2: PCR product of *matK* gene of *Litsea citrata* using *matK* 2.1F and *matK5R* primer set, Lane 3: PCR product of *rbcL* gene of *Litsea citrata* using *rbcL* 1F and *rbcL* 724R primer set; (N) 1.5% agarose gel electrophoresis of amplified products of *Ricinus communis*. Lane 1: 100 bp ladder, Lane 2-4: PCR product of ITS2 region of Red Castor using ITS2F and ITS3R primer set.

2.3.3 Sequence analysis and submission to database

Analysis of barcode sequences of silkworms and their pests

The COI sequences of *A. assama*, *S. ricini*, red ant and weevil were of 732 bp, 741 bp, 642 bp and 654 bp length, respectively. COI gene of *A. assama* and *S. ricini* showed highest similarity with the related species belonging to family Saturniidae (Order Lepidoptera). Similar to silkworms, COI sequences of red ant and weevil showed 99% sequence identity with *Oecophylla smaragdina* (Hymenoptera: Formicidae) and *Odoiporus longicollis* (Coleoptera: Oryophthoridae) (**Table 2.6**). Multiple sequence alignment also revealed highest similarity of *A. assama* Nongpoh with other *Antheraea* species (92.76% with *A. frithi* to 99.45% with *A. assamensis*) followed by 88-89% similarity with *Samia* species and lowest with *Bombyx mori* (86%). Similarly, *S. ricini* Borduar exhibited highest similarity with its related *Samia* species (92.40% with *S. watsoni* to 99.44% with *S. cynthia ricini*) followed by *Antheraea* species (88-89%) and lowest with *Bombyx* species (85%) (**Table 2.3A; Figure 2.6**). The multiple sequence alignment of sequenced ant and weevil species determined by Clustal Omega also showed maximum similarity with *Oecophylla* and *Odoiporus* species than other species (**Table 2.3 B & C and Figure 2.7**).

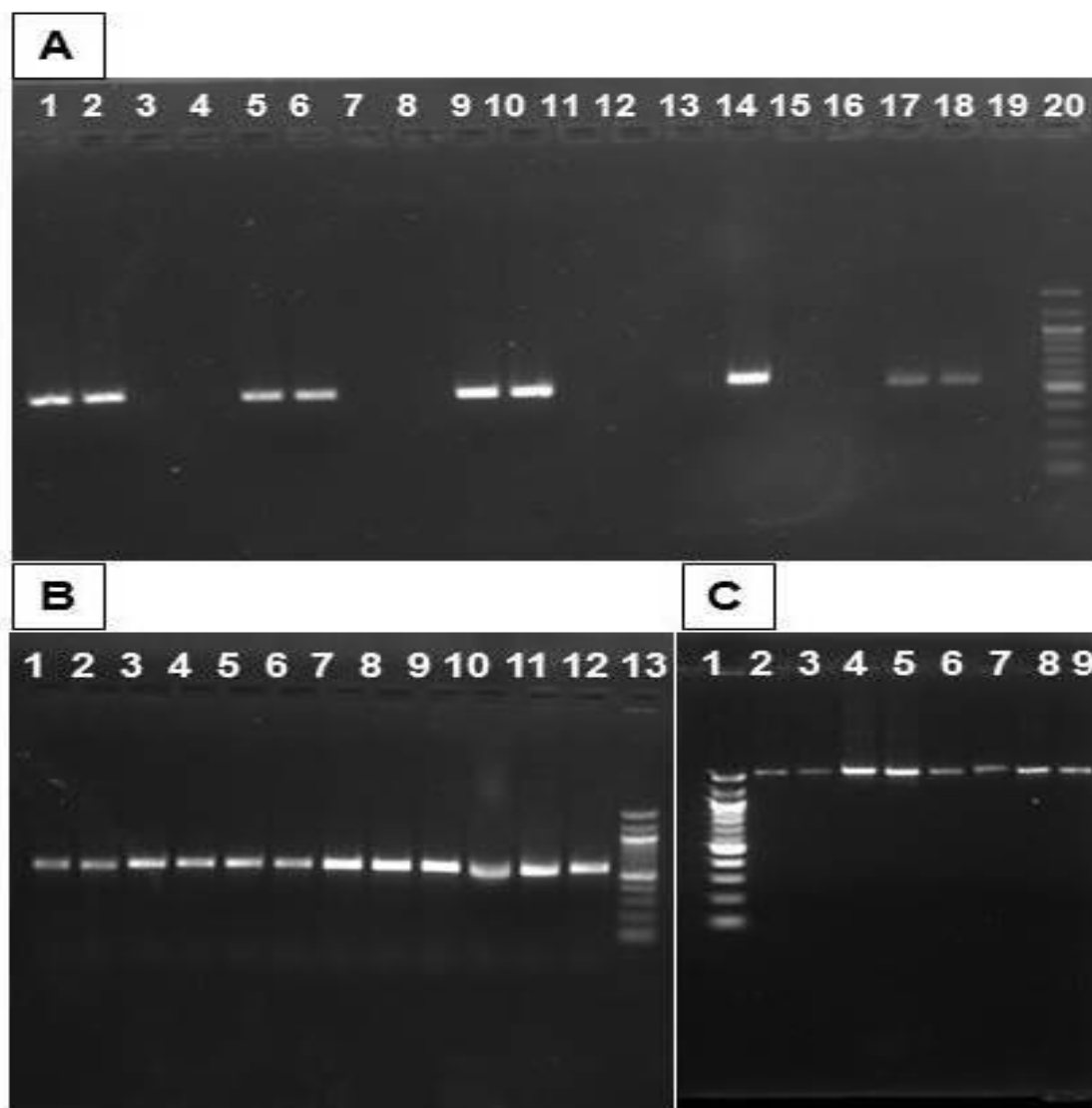


Figure 2.5 Agarose gel electrophoresis of PCR products of isolated DNA of fungal and bacterial colonies. (A) 1.5% agarose gel electrophoresis of amplified products of fungal colonies isolated from *L. monopetala* and *L. salicifolia* leaves using ITS1 and ITS4 primer set. Lane 1-2: PCR products of ITS region of fungus isolated from *L. monopetala* leaf, Lane 3, 4, 7, 8, 11, 12, 13, 15, 16, 19: Empty, Lane 5, 6, 9, 10, 14, 17 and 18: PCR products of ITS region of fungus isolated from *L. salicifolia* leaf, Lane 20: 100 bp DNA ladder; (B) 1.5% agarose gel electrophoresis of amplified products of fungal colonies isolated from *P. bombycina* leaves using ITS1 and ITS4 primer set. Lane 1-12: PCR products of ITS region of fungus isolated from *P. bombycina* leaf, Lane 13: 100 bp DNA ladder; (C) 1% agarose gel electrophoresis of amplified products of bacterial colonies isolated from gut of muga silkworm using 8F and 1492R primer set. Lane 1: 1 Kb DNA ladder, Lane 2-9: PCR products of 16S *rRNA* gene of bacterial colonies.

Table 2.6 BLAST similarity results for the sequenced species in the present study.

Species Name (Present study)	Barcode region	NCBI Percentage Similarity	Coverage	Similarity with species in NCBI GenBank
Silkworms & pests				
<i>Antheraea assama</i> strain Nongpoh	COI	99%	99%	<i>Antheraea assama</i>
<i>Samia ricini</i> strain Borduar	COI	100%	93%	<i>Samia ricini</i>
<i>Oecophylla</i> <i>smaragdina</i>	COI	99%	97%	<i>Oecophylla</i> <i>smaragdina</i>
<i>Odoiporus</i> <i>longicollis</i>	COI	99%	96%	<i>Odoiporus</i> <i>longicollis</i>
Host plants				
<i>Machilus</i> <i>bombycina</i>	matK	99%	99%	<i>Machilus</i> sp.
<i>Machilus</i> <i>bombycina</i>	rbcL	100%	100%	<i>Machilus</i> sp.
<i>Litsea salicifolia</i>	matK	99%	100%	<i>Litsea</i> sp.
<i>Litsea salicifolia</i>	rbcL	100%	100%	<i>Litsea salicifolia</i>
<i>Litsea monopetala</i>	matK	100%	100%	<i>Litsea</i> sp.
<i>Litsea monopetala</i>	rbcL	99%	99%	<i>Litsea monopetala</i>
<i>Litsea cubeba</i>	matK	99%	100%	<i>Litsea cubeba</i>
<i>Litsea cubeba</i>	matK	100%	100%	<i>Lindera obtusiloba</i>
<i>Litsea cubeba</i>	rbcL	100%	100%	<i>Litsea cubeba</i>
<i>Ziziphus jujuba</i>	psbA- trnH	99%	85%	<i>Zizyphus</i> <i>mauritiana</i>
<i>Ricinus communis</i>	rbcL	99%	100%	<i>Ricinus communis</i>
<i>Ricinus communis</i>	ITS2	100%	100%	<i>Ricinus communis</i>
<i>Euodia fraxinifolia</i>	rbcL	99%	100%	<i>Phellodendron</i> <i>amurense</i> & <i>Zanthoxylum</i> <i>bungeanum</i>
<i>Euodia fraxinifolia</i>	matK	99%	100%	<i>Euodia hupehensis</i>
<i>Euodia fraxinifolia</i>	matK	100%	99%	<i>Tetradium daniellii</i>
Fungus				
<i>Colletotrichum</i> <i>gloeosporioides</i> strain DS01	ITS	99%	100%	<i>Colletotrichum</i> <i>gloeosporioides</i>
<i>Colletotrichum</i> <i>gloeosporioides</i> strain DS02	ITS	99%	100%	<i>Colletotrichum</i> <i>gloeosporioides</i>
<i>Lasiodiplodia</i> <i>theobromae</i> strain DS03	ITS	99%	100%	<i>Lasiodiplodia</i> <i>theobromae</i>

<i>Neopestalotiopsis clavispora</i> strain DS05	ITS	99%	100%	<i>Fungal sp., Pestalotiopsis, Neopestalotiopsis</i>
<i>Penicillium oxalicum</i> strain DS06	ITS	100%	100%	<i>Penicillium oxalicum</i>
<i>Pestalotiopsis</i> sp. DS07	ITS	100%	100%	<i>Pestalotiopsis</i> sp.
<i>Phomopsis phyllanthicola</i> strain DS08	ITS	99%	100%	<i>Phomopsis phyllanthicola</i>
Bacteria				
<i>Micrococcus endophyticus</i> strain DS04	16S rRNA	100%	100%	<i>Micrococcus endophyticus</i>

Oecophylla smaragdina (red ant) is one of the major pests/predators infesting *A. assama* larvae throughout the year and the frequency is high in summer season (May-June, July-August) (Singh & Maheshwari, 2002; Subharani & Jayaprakash, 2015). But, no reports are available for identification of weevil *Odoiporus longicollis* as a pest of silkworm or its host plants. However, a study described genetic variation in avocado stem weevils *Copturus aguacatae* (Coleoptera: Curculionidae) which are major pests of avocado trees in Mexico (Engstrand et al., 2010). Similarly, weevil (*Myloccerus viridanus* Fab.) has been reported as a major pest of tropical tasar host plants (<http://silks.csb.gov.in/jhansi/diseases-and-pests-of-food-plants/>). Thus, COI based barcoding is successful in the identification of new weevil, *O. longicollis* collected from muga plantation.

Analysis of barcode sequences of host plants

Machilus bombycina showed highest similarity with its related *Machilus* species (Laurales: Lauraceae) based on both *matK* and *rbcL* genes. Likewise, *L. monopetala*, *L. salicifolia* and *L. cubeba* exhibited highest similarity with other related species belonging to the same family. Further, we found highest similarity of *matK* and *rbcL* sequences of *E. fraxinifolia* with its related members (*Euodia*

hupehensis, *Tetradium daniellii*, *Phellodendron amurense* and *Zanthoxylum bungeanum*) of family Rutaceae belonging to order Sapindales (Table 2.6). *M. bombycina* and *L. monopetala* are the primary host plants, *L. salicifolia* and *L. cubeba* are secondary host plants and *Z. mauritiana* is the tertiary host plant of muga silkworm. Similarly, *R. communis* and *E. fraxinifolia* are primary and secondary host plants of eri silkworm, respectively. The close relationship among the species based on different markers was further confirmed through phylogenetic tree analysis.

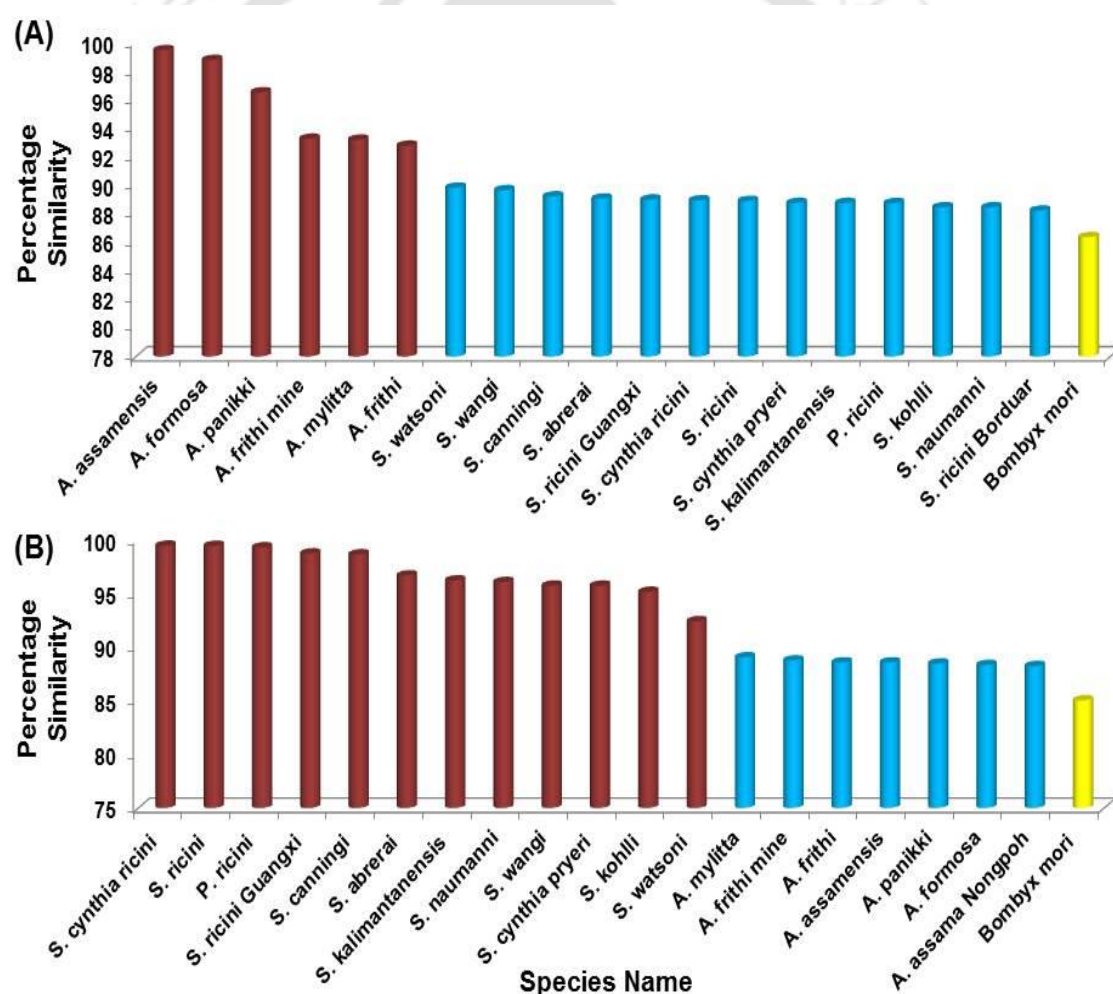


Figure 2.6 Sequence similarity of COI gene of silkworms. (A) *A. assama* Nongpoh with selected species (B) *S. ricini* Nongpoh with selected species. *Bombyx mori* was taken as outgroup in the study.

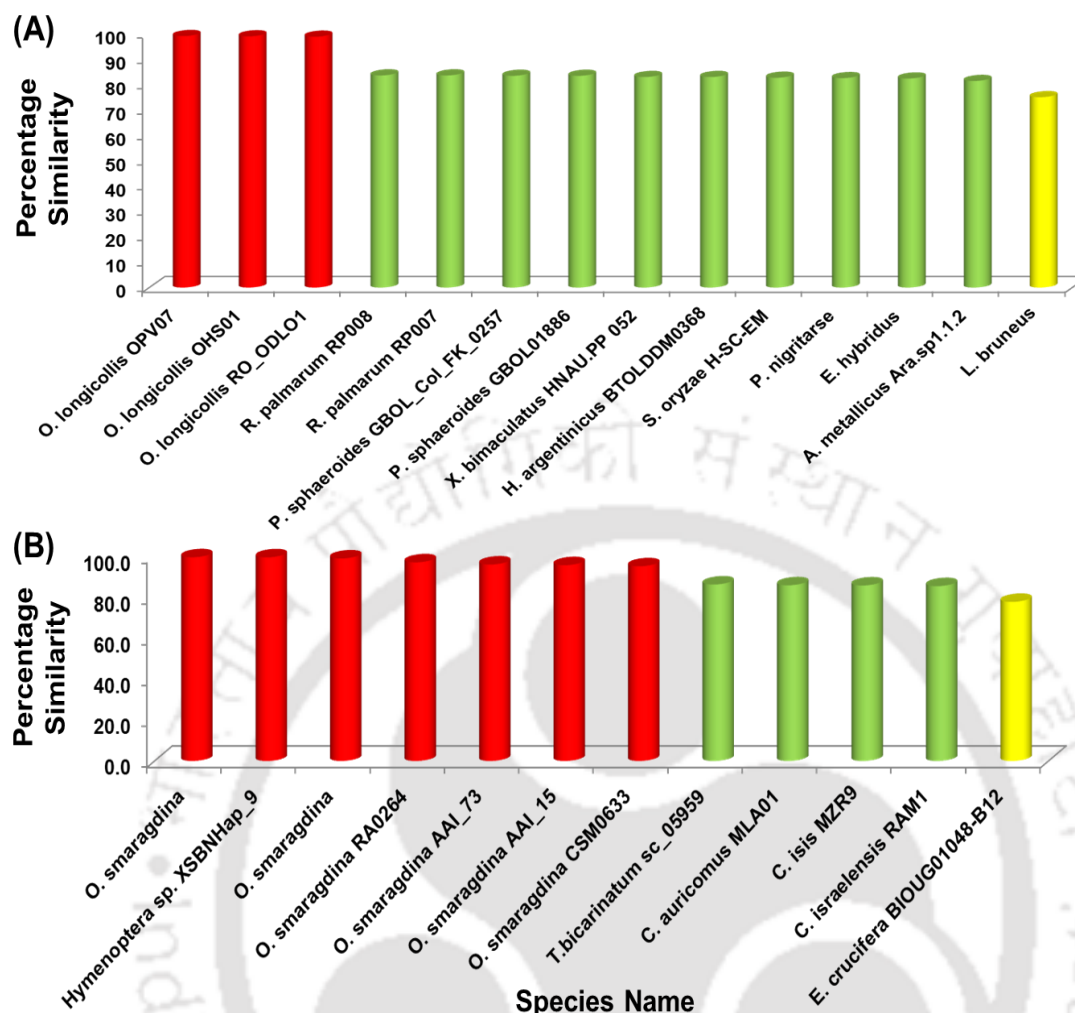


Figure 2.7 Sequence similarities of COI gene of weevils and ants with other related species. (A) *Odoiporus longicollis* with other related species taking *Lyctus brunneus* as outgroup (B) *Oecophylla smaragdina* with other related species taking *Eumenis crucifera* as outgroup.

Analysis of barcode sequences of fungal species

Morphologically different fungal colonies were successfully isolated from infected leaves of *M. bombycina*, *L. monopetala*, *L. salicifolia* and identified using ITS specific primers. BLAST similarity resulted in 99-100% sequence identity with fungal species- *Colletotrichum gloeosporioides*, *Lasiodiplodia theobromae* (*Botryosphaeria theobromae*), *Penicillium oxalicum*, *Pestalotiopsis* sp., *Neopestalotiopsis clavispora* and *Phomopsis phyllanthicola* (**Table 2.6**). Seven different fungal species were thus identified based on sequence similarity. The

species were named as *Colletotrichum gloeosporioides* DS01, *C. gloeosporioides* DS02, *Lasiodiplodia theobromae* DS03, *Neopestalotiopsis clavispora* DS05, *Penicillium oxalicum* DS06, *Pestalotiopsis* sp. DS07 and *Phomopsis phyllanthicola* DS08. The pairwise sequence alignment of *Colletotrichum* species (DS01 and DS02) revealed 98% identity between the two strains isolated from two different host plants (Som and Soalu).

The phyllosphere of plants provide shelter to microorganisms (bacteria and/or fungus) which either have negative impact on the plant by acting as pathogens or have positive impact by providing disease resistance, tolerance to different environmental conditions, nitrogen fixation, etc. Based on existing literature, *C. gloeosporioides* isolated from *M. bombycina* leaf causes **anthracnose** disease, the symptoms of which include brown/ash colored leaves with round or oval spots irregularly spread over the leaf. Brownish lesions or streaks appear at the later stage resulting in premature leaf fall (Das *et al.*, 2005; Handique *et al.*, 2008; Chowdappa *et al.*, 2009). Similarly, **Brown blight disease** is caused by *C. gloeosporioides* in *L. monopetala*, which is a serious disease that occurs throughout the year. The symptoms of this disease include ash colored round and irregular spots appearing on the young and mature leaves which then spread to the entire leaf. The spots turn grey in color with the progression of disease and thus appear blight. Studies reported around 13.8% -33.5% leaf yield loss and highest infection rate in the month of April and May (Das *et al.*, 2014).

The occurrence of grey blight disease due to *Pestalotiopsis* species has been reported in *L. monopetala* while in *M. bombycina*, it causes grey blight disease, anthracnose and stem end rot disease (Korsten *et al.*, 1994; Das *et al.*, 2010). However, we found no reports for *L. salicifolia* (Dighloti) plant. Similarly, no

information is available for other fungi *P. phyllanthicola* (Som), *N. clavispora* (Som) and *L. theobromae* (Dighloti) identified from the host plants of our study. *Phomopsis* sp. and *L. theobromae* have been reported to cause leaf spot disease in many plants like brinjal, tea, canes, etc. (Holliday, 1995; Akhtar, 2006; Tongsri *et al.*, 2016). Apart from the above fungal isolates, we found *P. oxalicum* from *L. salicifolia* leaf, which is not a pathogen but a fungus reported as a phyllosphere member of *P. bombycina* along with *Aspergillus* and *Fusarium* (Bhuyan *et al.*, 2013). Further sampling and studies on the disease pathogens of muga host plants is required for prior recognition of the infection which will help in the protection of host plants from diseases.

Analysis of barcode sequences of gut bacteria

The bacterium identified from the gut of *A. assama* is *Micrococcus endophyticus* strain DS04 which showed 100% similarity with *Micrococcus endophyticus* based on BLAST sequence similarity results (**Table 2.6**). Multiple sequence alignment (Clustal Omega) of the strain DS04 also resulted in 99.9% sequence similarity with *M. endophyticus* while 97 to 98% similarity with other *Micrococcus* species and lowest with *Bacillus* species (81.2%) which was taken as outgroup (**Figure 2.8**). Bacterial species belonging to genera *Micrococcus*, *Bacillus*, *Lactobacillus*, *Pseudomonas*, *Escherichia*, *Satphylococcus*, etc. have been reported to inhabit the intestine/gut of silkworms which help the silkworm in utilizing host plants and also in preventing diseases (Yuan *et al.*, 2006; Liang *et al.*, 2014; Haloi *et al.*, 2016). All the sequences barcoded in the present study have been submitted to NCBI GenBank as presented in **Table 2.7**.

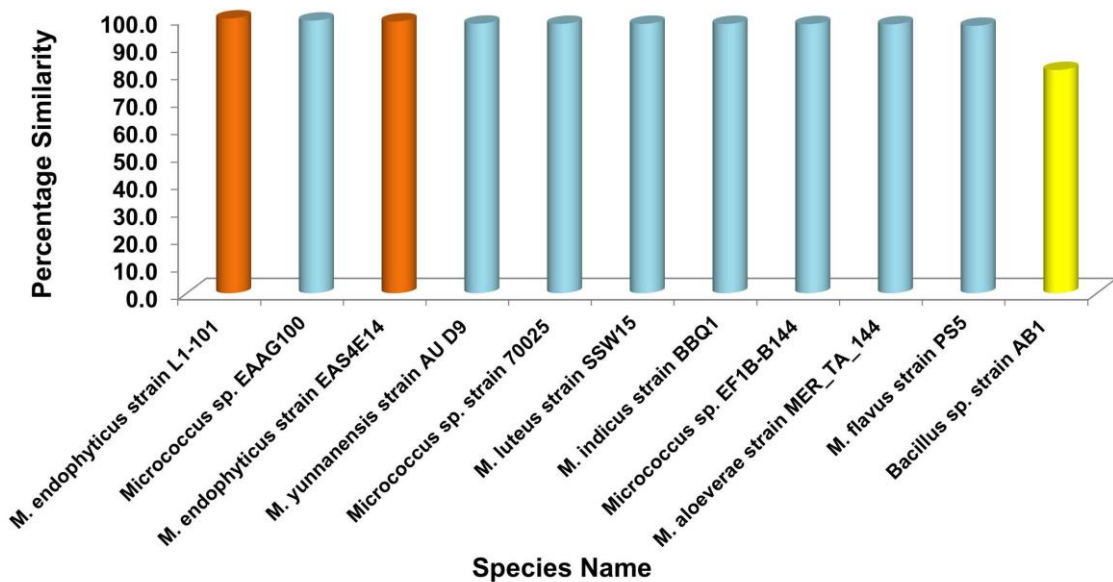


Figure 2.8 Sequence homology of 16S *r*RNA gene of *Micrococcus endophyticus* strain DS04 with other related species. *Bacillus* sp. strain AB1 here denotes an outgroup species.

Table 2.7 Barcode sequences submitted to NCBI GenBank database.

Species Name	Sample type	Gene/ Barcode region	GenBank Accession
Silkworms & pests			
<i>Antheraea assama</i> strain Nongpoh	Lepidoptera: Saturniidae	<i>COI</i>	KC248388.1
<i>Samia ricini</i> strain Borduar	Lepidoptera: Saturniidae	<i>COI</i>	KC195912.1
<i>Oecophylla smaragdina</i>	Vespoidea: Formicidae	<i>COI</i>	MF278656
<i>Odoiporus longicollis</i>	Coleoptera: Curculionidae	<i>COI</i>	MF278655
Host Plants			
<i>Machilus bombycina</i>	Laurales: Lauraceae	<i>matK</i>	KC771284.1
<i>Machilus bombycina</i>	Laurales: Lauraceae	<i>rbcL</i>	KJ499539.1
<i>Litsea salicifolia</i>	Laurales: Lauraceae	<i>matK</i>	KF523364.1
<i>Litsea salicifolia</i>	Laurales: Lauraceae	<i>rbcL</i>	KF523365.1
<i>Litsea monopetala</i>	Laurales: Lauraceae	<i>matK</i>	KF912875.1
<i>Litsea monopetala</i>	Laurales: Lauraceae	<i>rbcL</i>	KF912876.1

<i>Litsea cubeba</i>	Laurales:	<i>matk</i>	KF912877.1
	Lauraceae		
<i>Litsea cubeba</i>	Laurales:	<i>rbcL</i>	KF912878.1
	Lauraceae		
<i>Ziziphus jujuba</i>	Rosales:	<i>psbA-trnH</i>	KT342880.1
	Rhamnaceae		
<i>Ricinus communis</i>	Malpighiales:	<i>rbcL</i>	KF912880.1
	Euphorbiaceae		
<i>Ricinus communis</i>	Malpighiales:	ITS2	KP793207.1
	Euphorbiaceae		
<i>Euodia fraxinifolia</i>	Sapindales:	<i>rbcL</i>	KF912881.1
	Rutaceae		
<i>Euodia fraxinifolia</i>	Sapindales:	<i>matK</i>	KP793206.1
	Rutaceae		
Fungus			
<i>Colletotrichum gloeosporioides</i> strain DS01	Glomerellales:	ITS	KT342872.1
	Glomerellaceae		
<i>Colletotrichum gloeosporioides</i> strain DS02	Glomerellales:	ITS	KT342873.1
	Glomerellaceae		
<i>Lasiodiplodia theobromae</i> strain DS03	Botryosphaerales : Botryosphaeriaceae	ITS	KT342874.1
<i>Neopestalotiopsis clavispora</i> strain DS05	Xylariales:	ITS	KT342876.1
	Sporocadaceae		
<i>Penicillium oxalicum</i> strain DS06	Eurotiales:	ITS	KT342877.1
	Aspergillaceae		
<i>Pestalotiopsis sp.</i> strain DS07	Xylariales:	ITS	KT342878.1
	Sporocadaceae		
<i>Phomopsis phyllanthicola</i> strain DS08	Diaporthales:	ITS	KT342879.1
	Valsaceae		
Bacteria			
<i>Micrococcus endophyticus</i> strain DS04	Micrococcales:	16S rRNA	KT342875.1
	Micrococcaceae		

2.3.4 Phylogenetic analysis

Optimal model selection was performed for various datasets prior to phylogenetic analysis which are given in **Table 2.4**. The phylogenetic tree inferred using *COI* dataset was found to be consistent with traditional classification system. *A. assama* Nongpoh and *S. ricini* Borduar grouped under Saturniidae family of Bombycoidea superfamily of the order Lepidoptera. *S. ricini* Borduar formed single

cluster with its closely related species *S. cynthia ricini* while *A. assama* Nongpoh was close to its related species *A. frithi*, *A. assama* and *A. formosana*. These organisms showed distant relationship with its next related family member *B. mori* (Bombycoidea: Bombycidae) which was taken as outgroup in the study (**Figure 2.9**).

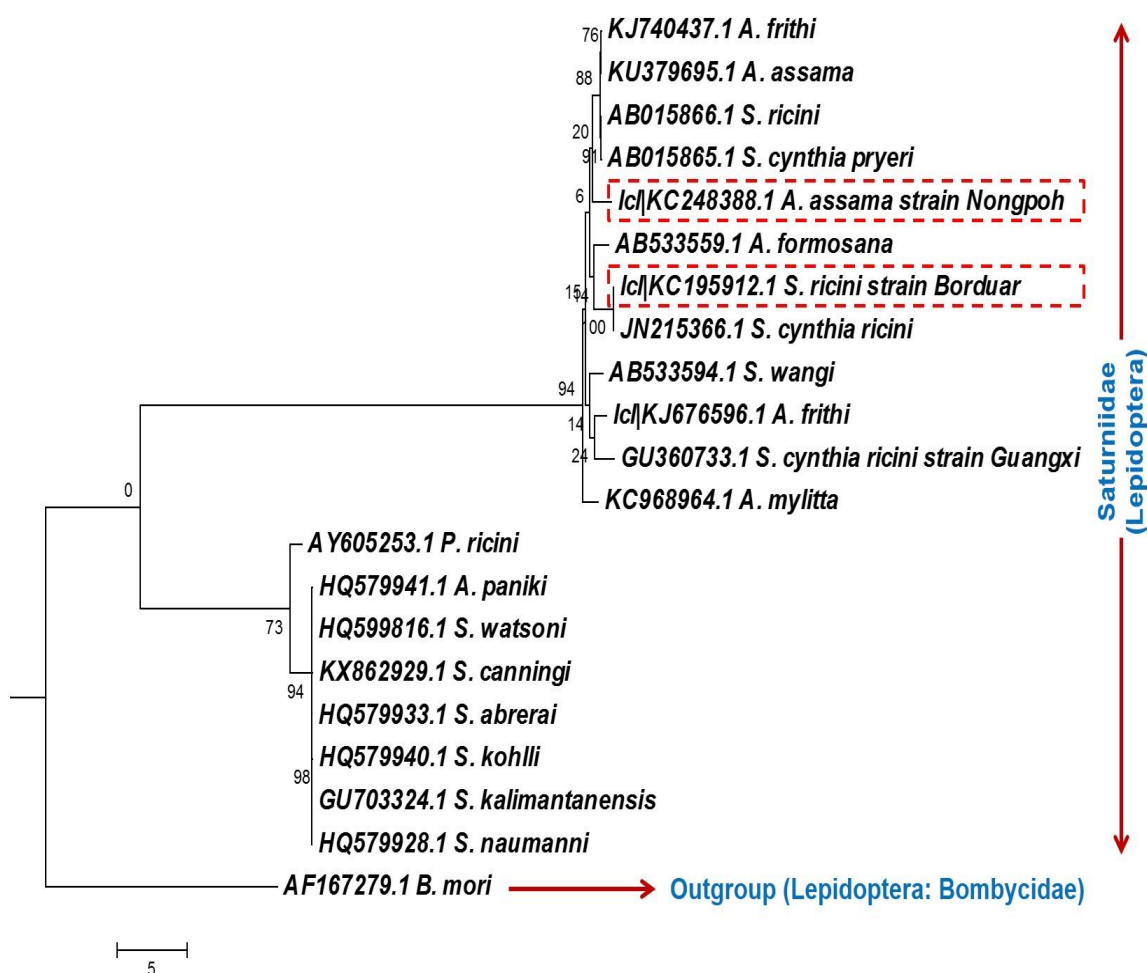


Figure 2.9 Phylogenetic tree inferred using COI sequences of *Antheraea assama* Nongpoh and *Samia ricini* Borduar with other selected silkworms through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

The phylogenetic tree inferred using ML method in pests also showed that *O. longicollis* was close to its related *O. longicollis* OPV07 while *O. smaragdina* was close to *O. smaragdina* species. Thus, phylogenetic tree analysis was in

accordance with classical traditional system- *O. longicollis* belonging to Curculionidae family of order Coleoptera while *O. smaragdina* belonging to Formicidae family of order Hymenoptera (**Figure 2.10**).

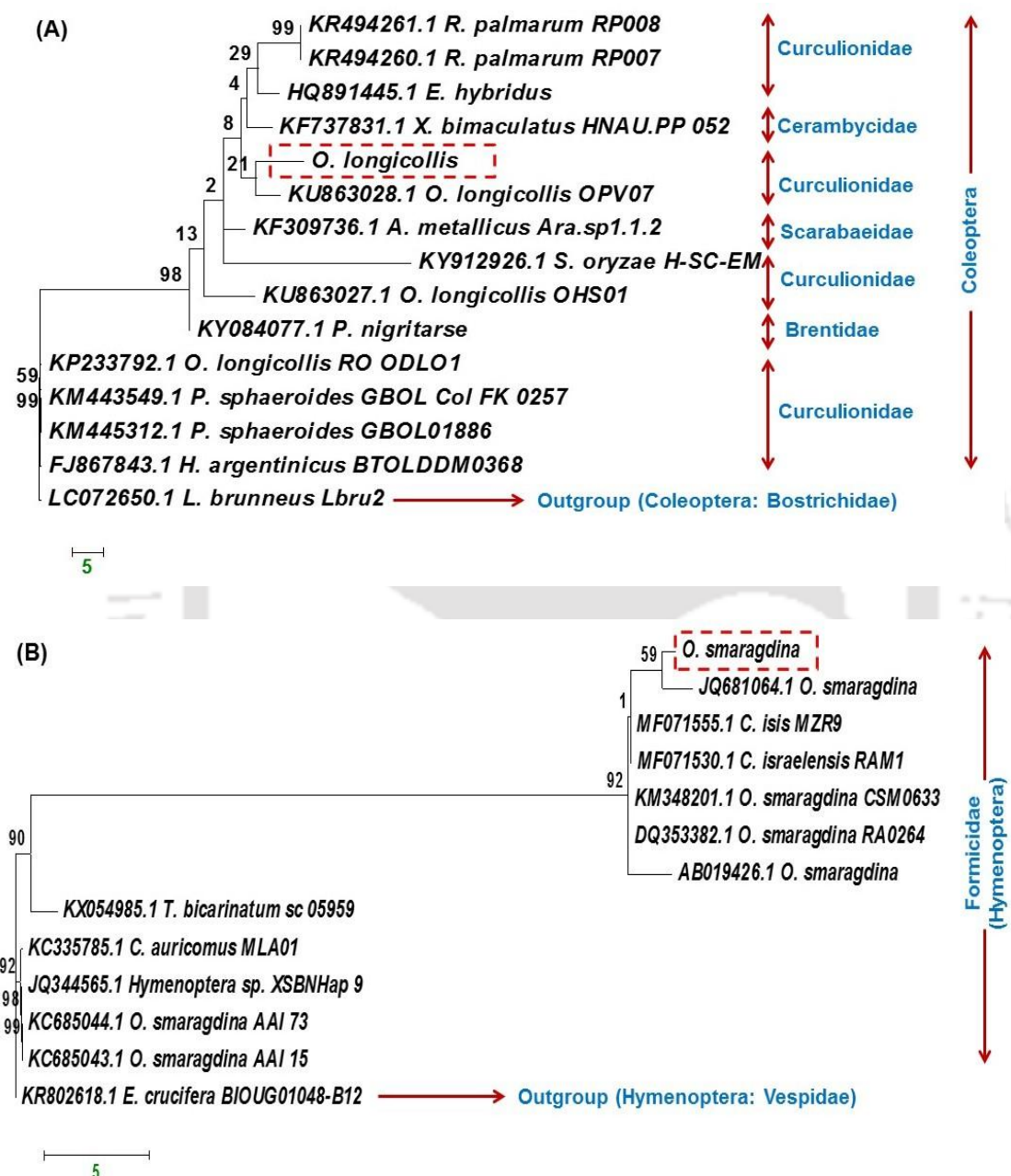


Figure 2.10 Phylogenetic tree inferred using COI sequences of sequenced pests through ML method in Mega 6.0 tool (1000 bootstrap replicates). (A) *Odoiporus longicollis* (weevil) with other selected organisms (B) *Oecophylla smaragdina* (red ant) with other selected organisms. Tree was drawn to scale with bootstrap values above the line.

Further, *matK* and *rbcl* gene sets were found to be efficient in identifying the host plants under study. The host plants belonging to Lauraceae family are *M. bombycina* clustered with other *Machilus* species; *L. monopetala*, *L. cubeba* and *L. salicifolia* showed close relationship with other related species (Order: Laurales, Family: Lauraceae). *E. fraxinifolia* grouped with other *Euodia* species (Order: Sapindales, Family: Rutaceae) while *R. communis* grouped with other *Ricinus* species (Order: Malpighiales, Family: Euphorbiaceae) (**Figure 2.11 & 2.12**). Similarly, *Z. jujuba* formed single clade with other related *Z. mauritiana* species (Order: Rosales, Family: Rhamnaceae) using *trnH-psbA* region (**Figure 2.13A**). ITS2 region was also successful in classifying *R. communis* which clustered with other related *Ricinus* species (Order: Malpighiales, Family: Euphorbiaceae) in a single clade using ML method (**Figure 2.13B**).

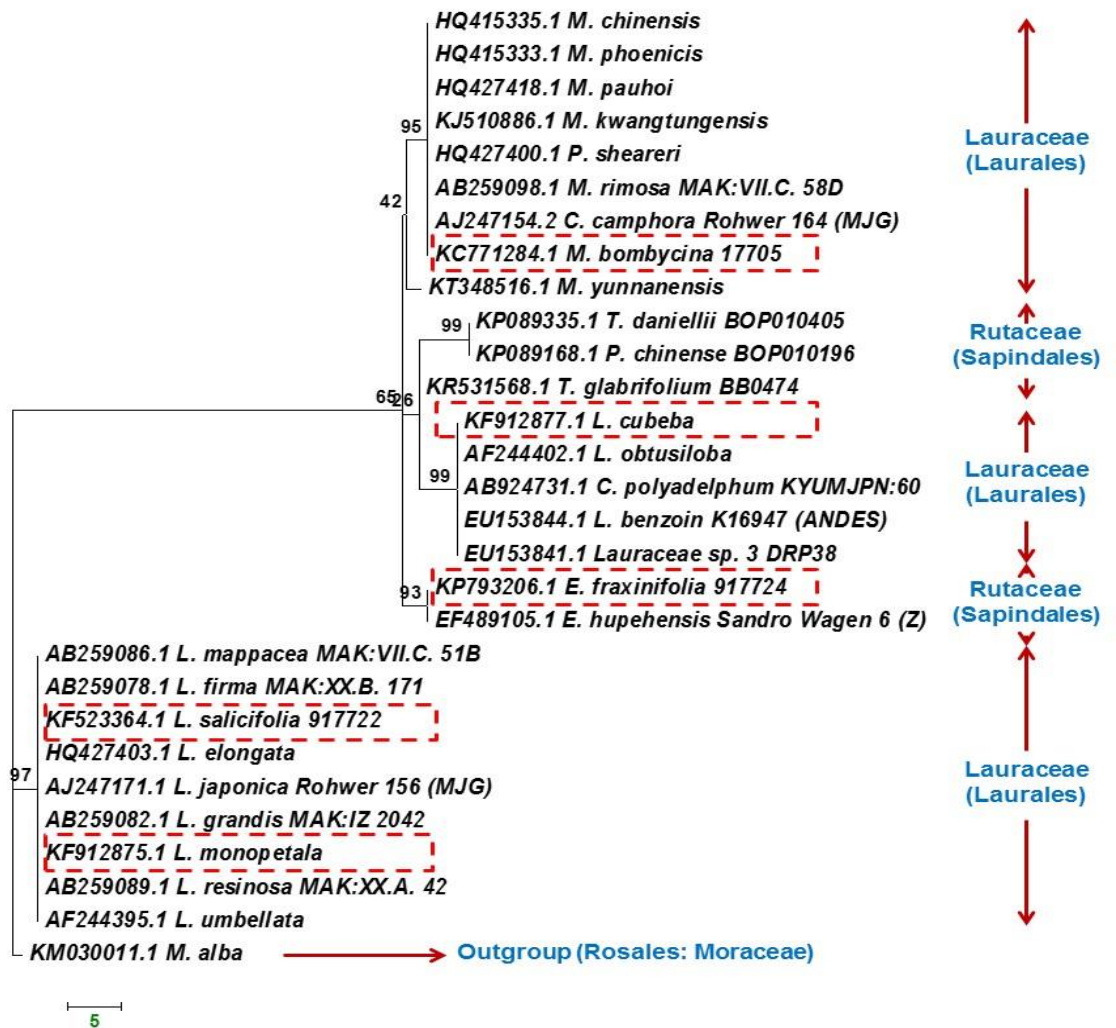


Figure 2.11 Phylogenetic tree inferred using *matK* sequences of sequenced host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

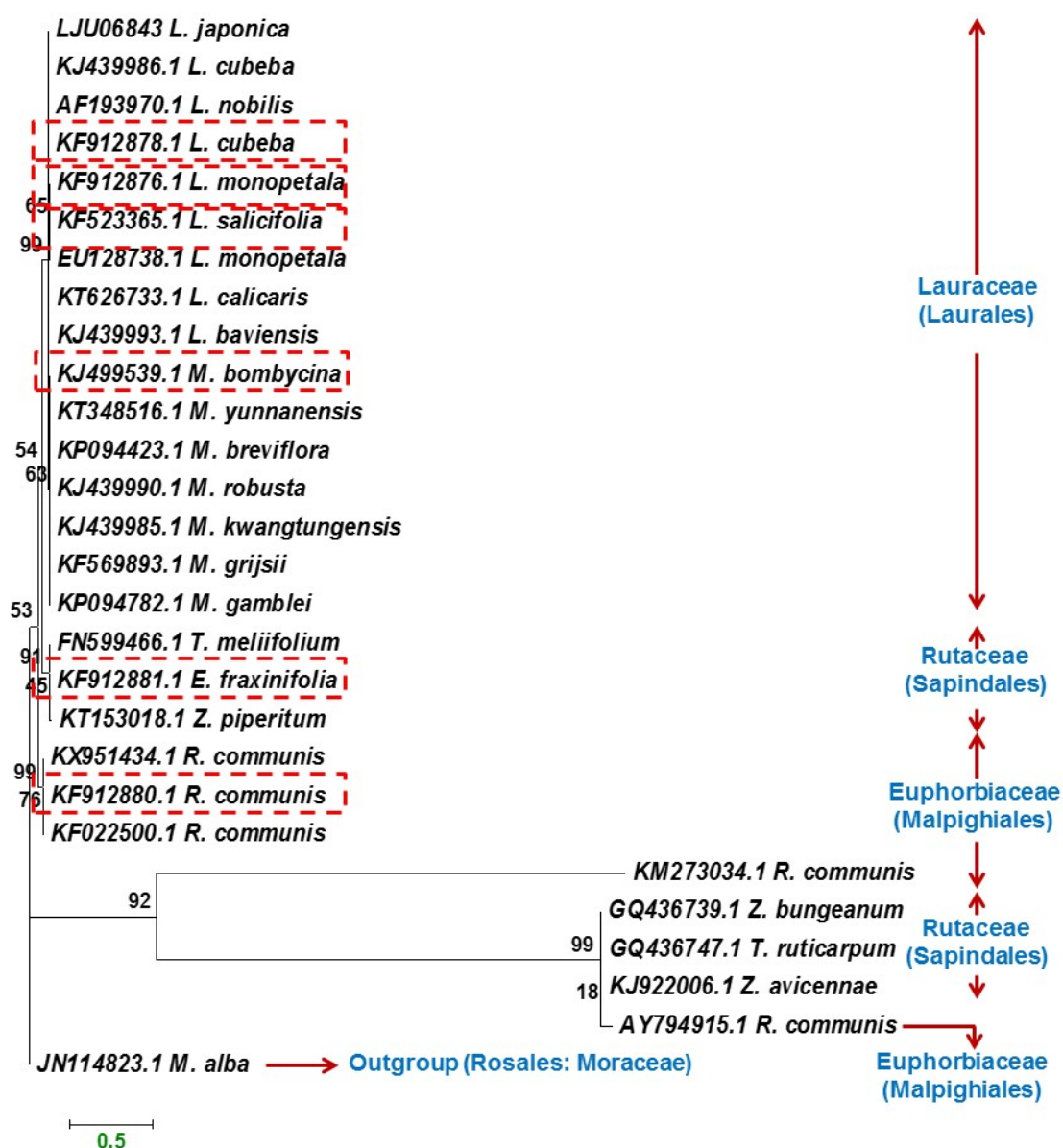


Figure 2.12 Phylogenetic tree inferred using *rbcL* sequences of sequenced host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

Phylogenetic trees inferred using ITS region for fungal isolates clustered the isolates according to their respective genus based on ML method (**Figure 2.14**). *C. gloeosporioides* DS01 and DS02 formed cluster with other *Colletotrichum* species (Order: Glomerellales, Family: Glomerellaceae), *L. theobromae* DS03 grouped with other *Lasiodiplodia* species (Order: Botryosphaeriales, Family: Botryosphaeriaceae).

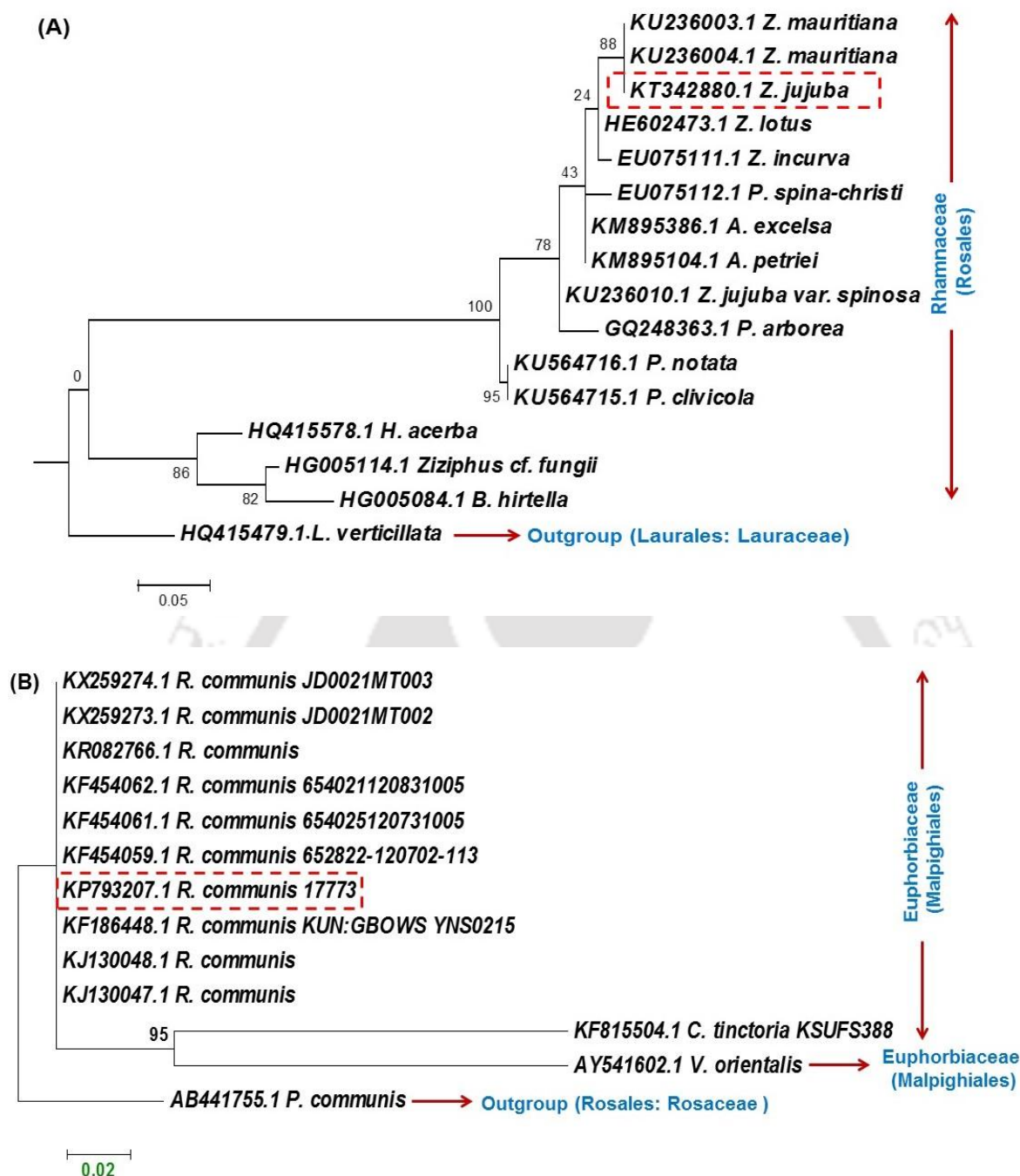


Figure 2.13 Phylogenetic tree inferred using *trnH-psbA* and ITS2 regions of host plants through ML method in Mega 6.0 tool (1000 bootstrap replicates). (A) Tree based on *trnH-psbA* region of *Z. jujuba* with other selected plants (B) Tree based on ITS2 region of *R. communis* with other selected plants. Tree was drawn to scale with bootstrap values above the line.

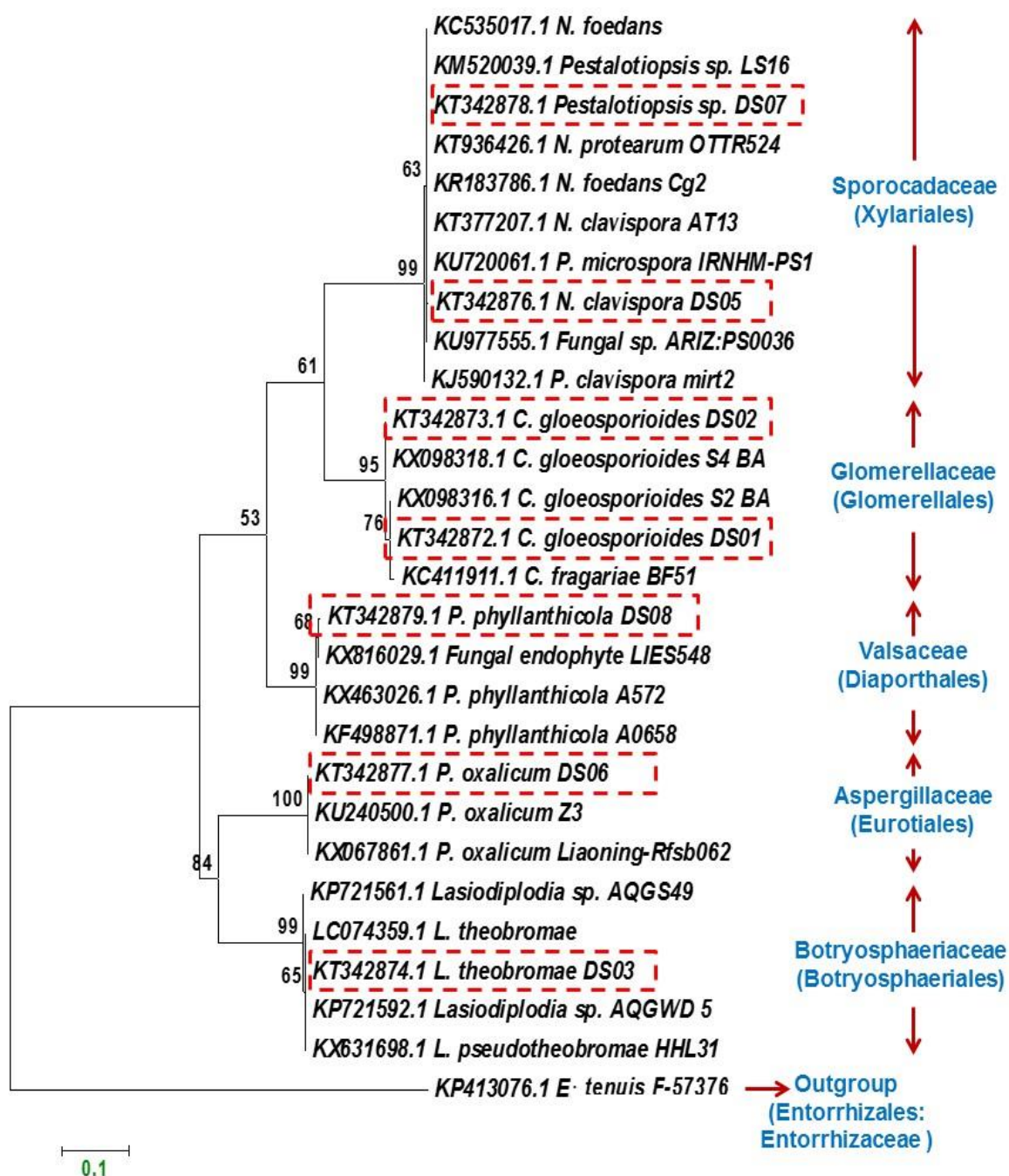


Figure 2.14 Phylogenetic tree inferred using ITS sequences of the sequenced fungal isolates and other selected species through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

Similarly, *N. clavispora* DS05 showed close relationship with its related *Neopestalotiopsis* and *Pestalotiopsis* species (Order: Xylariales, Family: Sporocadaceae), *P. oxalicum* DS06 formed single clade with other *Penicillium*

species (Order: Eurotiales, Family: Aspergillaceae) and *Pestalotiopsis* sp. DS07 clustered with other related *Pestalotiopsis* species (Order: Xylariales, Family: Sporocadaceae). *P. phyllanthicola* DS08 formed a clade with fungal endophyte LIES548 which was nested with another clade of its closely related *P. phyllanthicola* species (Order: Diaporthales, Family: Valsaceae).

Moreover, the phylogenetic tree inferred based on ML method using 16S rRNA sequences showed that *M. endophyticus* DS04 formed cluster with its closely related *M. endophyticus* L1-101 (Order: Micrococcales, Family: Micrococcaceae) (**Figure 2.15**). Our study thus suggests that the selected barcode regions and their amplification using selected primer sets were successful in the identification of selected samples.

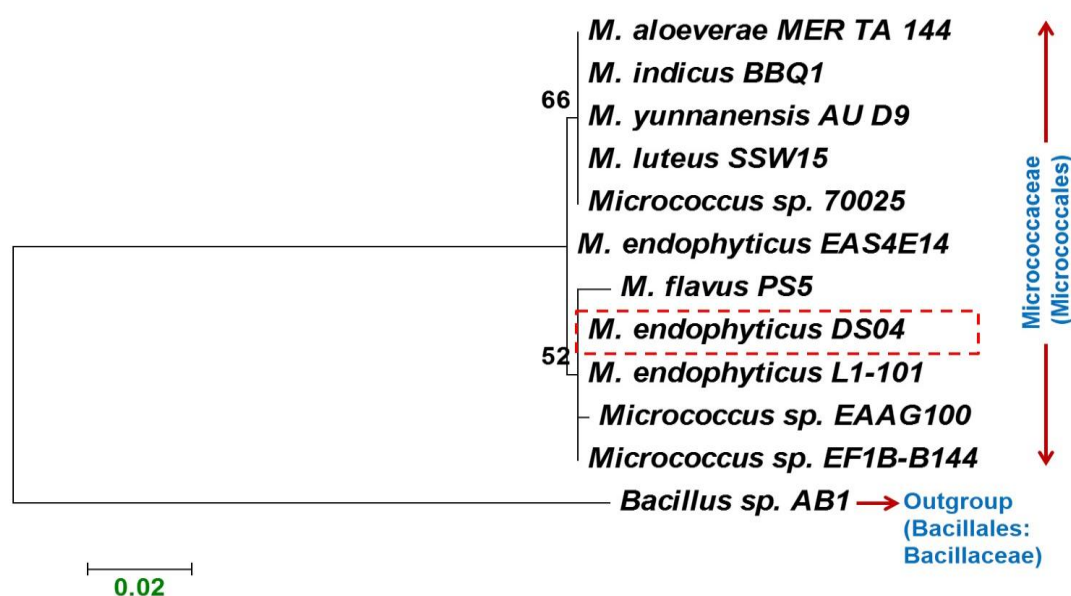
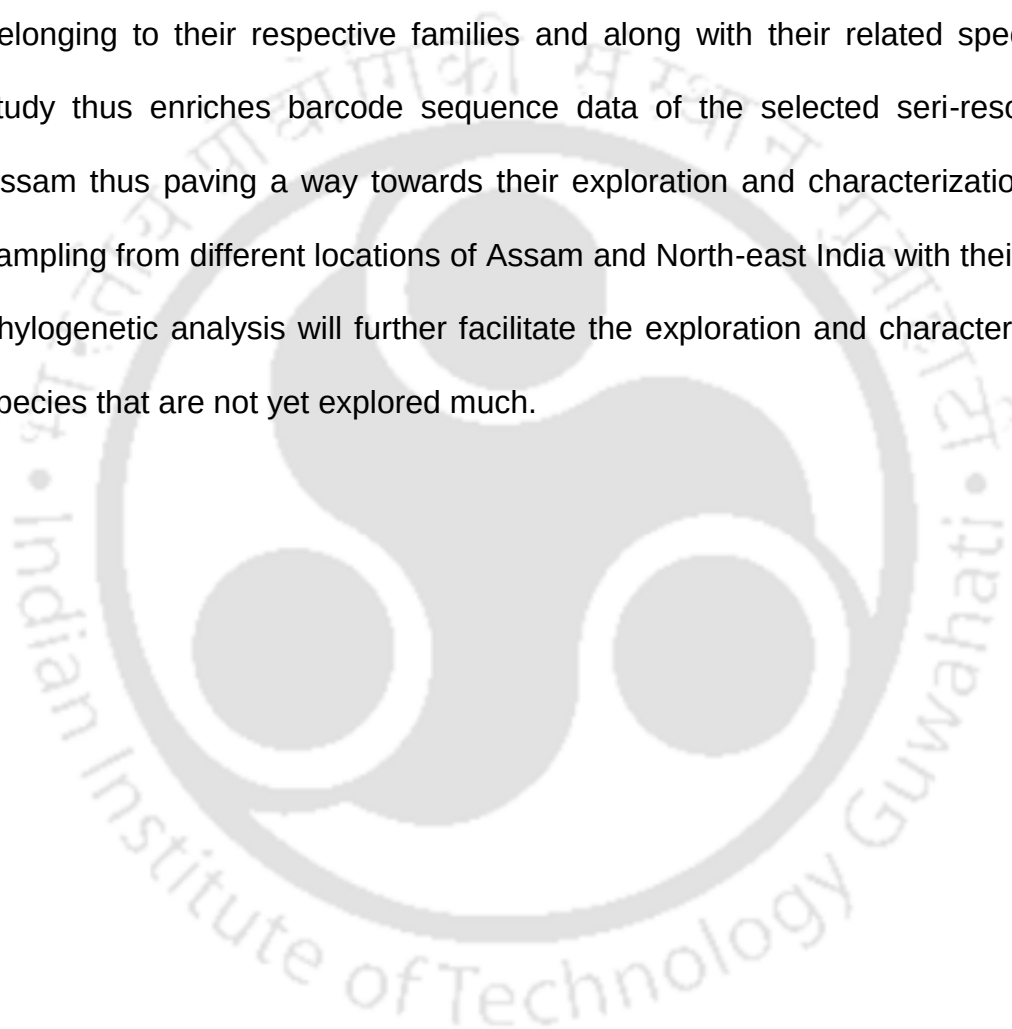


Figure 2.15 Phylogenetic tree using 16S rRNA gene sequence of sequenced bacterial isolate through ML method in Mega 6.0 tool (1000 bootstrap replicates). Tree was drawn to scale with bootstrap values above the line.

2.4 Conclusions

The present study aimed at sequencing barcode regions of selected ser-resources of Assam for proper identification and characterization of the species. A

total of 25 barcode sequences of *A. assama* Nongpoh, *S. ricini* Borduar, muga and eri host plants, fungal isolates, bacteria, ant and weevil have been generated for the first time. Sequence homology revealed higher similarity with its related species than distant species. Phylogenetic trees inferred using barcode sequences based on ML method clustered the species into distinctive groups belonging to their respective families and along with their related species. Our study thus enriches barcode sequence data of the selected seri-resources of Assam thus paving a way towards their exploration and characterization. Taxon sampling from different locations of Assam and North-east India with their detailed phylogenetic analysis will further facilitate the exploration and characterization of species that are not yet explored much.





Chapter 3

**Mitochondrial genome
sequencing of *Antheraea
assamensis*, a semi-
domesticated silkworm endemic
to Assam and North-eastern
part of India**

CHAPTER 3

Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkmoth endemic to Assam and North-eastern part of India

3.1 Introduction

Although small in size, the mitogenome governs maternal inheritance and has several unique features like faster evolution rate, low or absence of homologous recombination, evolutionary conserved gene products and richness in genetic polymorphism. This makes it a potential marker for barcoding, phylogeographic and phylogenetic studies (Cameron, 2014; Timmermans *et al.*, 2014). It plays a potential role in molecular evolutionary studies by elucidating evolutionary models and substitution patterns that vary timely and across sequences. Compared to individual genes, whole mitogenomes are more informative phylogenetic models due to its multiple genome level features like gene position, content, secondary structures of RNA and control region.

Over the past few decades, animal mitogenomes, particularly insects (~80% of the sequenced arthropods), have been widely studied for comparative genomics and molecular systematics (Boore, 2006; Cameron, 2014). More than 250 Lepidopteran insect mitogenomes have been sequenced till now using Sanger sequencing and next generation sequencing (NGS) technologies as found in GenBank (<https://www.ncbi.nlm.nih.gov/>). The mitogenome ranges from 14-16 kilo basepairs (kb) in majority of the lepidopterans and consists of 37 genes (13 PCGs, 2 rRNA genes, 22 tRNA genes) along with a control region. The PCGs

encode 2 subunits of ATPase (ATP6 and ATP8), 3 subunits of cytochrome c oxidase (COI, COII and COIII), 1 subunit of cytochrome *b* (CYTB) and 7 subunits of NADH dehydrogenase (ND1, ND2, ND3, ND4, ND4L, ND5 and ND6). These proteins are responsible for oxidative phosphorylation (OXPHOS) as they form essential mitochondrial membrane-associated protein complex systems (Cameron, 2014). The control region plays a role in the replication and transcription of mitogenome. The utilities and potential of mitochondrial PCGs (*cox1* and *cox2*) as barcode markers have been well demonstrated in the order Lepidoptera (Mandal *et al.*, 2014). The comparative mitogenome analysis also elucidated sequence divergence patterns among domesticated and non-domesticated lepidopteran mitogenomes (Arunkumar *et al.*, 2006; Liu *et al.*, 2012). Although frequency of mitogenome sequencing of lepidopterans has increased, the evolutionary relationships among many family members of the same order have been seldom investigated.

Saturniidae is the most diverse family of wild silkmoths which include giant moths, royal moths and emperor moths. Most of these silkmoths are unexplored and might have potential significance in sericulture area (Nässig *et al.*, 1996). Till date, mitogenome data of eleven species have been sequenced and made available in GenBank. However, individual gene sequences of various wild species are also available without their complete mitogenome sequences. *Antheraea assamensis*, muga silkmoth is semi-domesticated and one of the economically important moths of the same family. It is endemic to Assam and its adjoining hilly areas located in the North- Eastern part of India. It is a multivoltine and polyphagous moth that primarily thrives on two host plants *Persea bombycina* and *Litsea monopetala* (Tikader *et al.*, 2013). Its silk has wide applications in textile industry and has

great potential as a biomaterial due to its unique biophysical properties like golden lustre, tenacity and UV radiation absorption (Kasoju *et al.*, 2009). Like other *Antheraea* species, it produces reelable silk which is the most expensive silks of the world. However, its semi-domesticated nature and extraction of fibroin directly from cocoon fibers limits its extensive rearing and prospects of global applicability as a biomaterial. The whole genome of *A. assamensis* is not yet available. However, *de novo* transcriptome data from our laboratory is available in GenBank (Accession Number- SRX1293136, SRX1293137 and SRX1293138).

In the present study, we report the whole mitogenome sequence of *A. assamensis* using NGS and comparative analysis of its sequences and genome architectures with that of the other lepidopterans. The comparative study was based on several characteristics such as genome arrangement, PCGs, tRNAs, rRNAs, nucleotide composition, codon usage, evolutionary rates, gene divergence, conserved regions in control region, etc. Furthermore, phylogenetic trees inferred using datasets like nucleotide sequences of 13 PCGs and 13 PCGs+2 rRNAs were analyzed to elucidate the relationships among lepidopteran insects. This study will thus facilitate better understanding of comparative and evolutionary biology of *A. assamensis* with the other lepidopteran insects.

3.2 Materials and methods

3.2.1 Sample processing, DNA sequencing and assembly

The larvae of *A. assamensis* were reared on *Persea bombycina* in the experimental field of Central Muga Eri Research and Training Institute, Lahdoigarh, Jorhat, Assam, India following recommended package of practices (Chakravorty *et al.*, 2005). The fifth instar larval stage of *A. assamensis* was used

for the present study (Sample ID: CMERI-Aa-001). The larvae were sterilized by washing with 70% ethanol before dissecting and chopping for mitogenome studies. The larvae were stored in 95% absolute ethanol at -80°C for future use. The total DNA was extracted using CTAB (Cetyl trimethylammonium bromide) based buffer and silica column by the service provider (Genotypic Technology Pvt. Ltd. Bengaluru, India). Subsequently, mitochondrial DNA was enriched from total DNA extracted once the integrity, quantity and purity of extracted DNA was confirmed by agarose gel electrophoresis, light absorbance and fluorescence method. The complete overview of sequencing and analysis of mitochondrial genome of *A. assamensis* is represented in **Figure 3.1**.

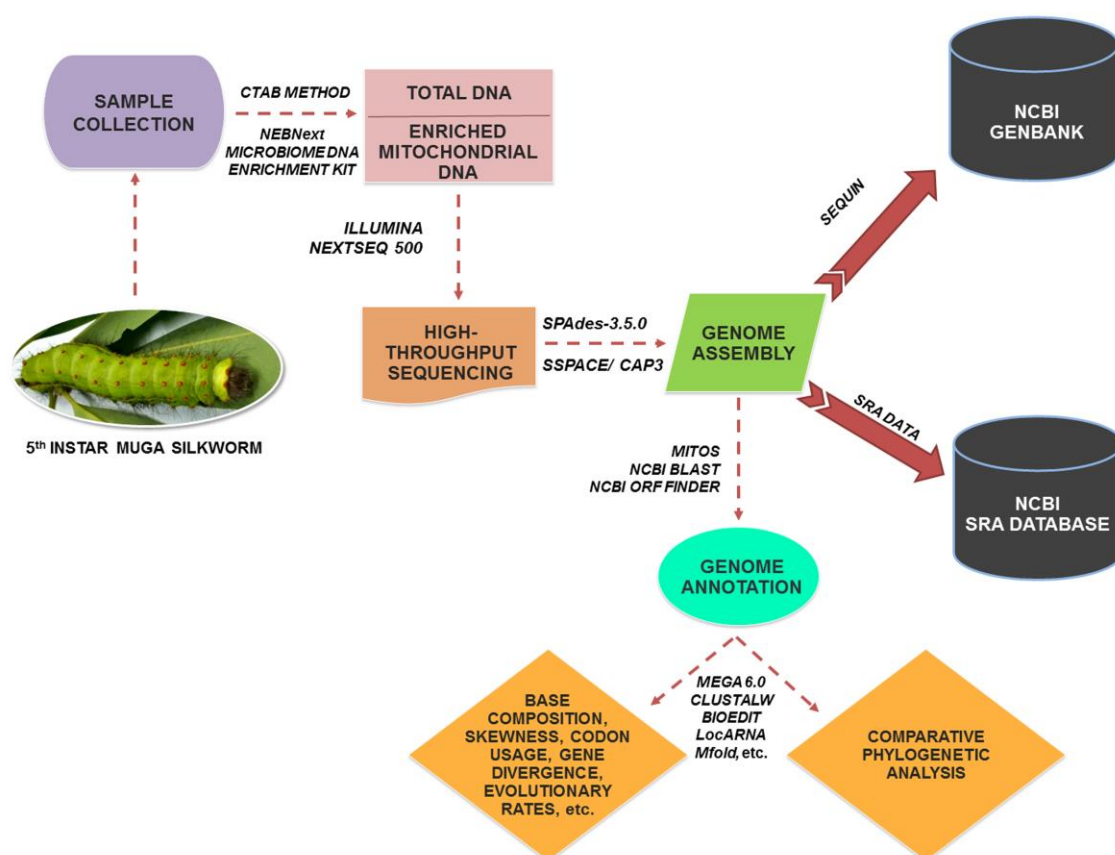


Figure 3.1 A pictorial overview of the methodologies used for sequencing and analysis of *A. assamensis* mitogenome.

The preferential enrichment of mitochondrial DNA was carried out with NEBNext microbiome DNA enrichment kit (New England Biolabs, USA) which selectively removes CpG-methylated eukaryotic nuclear DNA. The enriched mitochondrial DNA was subjected to DNA sequencing at Genotypic Technology's Genomics facility where the enriched DNA was acoustically sheared to 300-500 bp using a specially designed Covaris microTube (for focused ultra-sonication). The fragmented mitochondrial DNA was cleaned up using HighPrep beads (MagBio Genomics, Inc, Gaithersburg, Maryland) and subjected to end-repair, A-tailing and ligation with multiplex adaptors using the NEXTFlex DNA Sequencing kit (Catalogue # 5140-02, Bioo Scientific). The ligated DNA was cleaned with HighPrep beads and subjected to amplification via PCR as follows: initial denaturation at 98°C for 2 min; 10 cycles of denaturation at 98°C for 30 sec, annealing at 65°C for 30 sec and extension at 72°C for 60 sec; and a final extension at 72°C for 4 min using the primers provided by NEXTFlex DNA Sequencing kit. Finally, the PCR product was purified with HighPrep beads, quantified and fragment range was assessed using Qubit fluorometer and the Agilent D1000 Tape (Agilent Technologies), respectively prior to sequencing. The sequencing was carried out in Illumina NextSeq 500 sequencer (Illumina Inc, Sandiego, USA) through 2 × 150 bp paired-end chemistry. The raw paired end data obtained was de-multiplexed using Bcl2fastQ, assessed with FastQC tool (Andrews, 2010) to remove low quality bases (Q<30) and preprocessed using ABLT-Scripts (Genotypic technology, Bangalore India). The SPAdes-3.5.0 was used for sequence assembly followed by the filling of gaps. Finally, scaffolding of assembled contigs and clustering were carried out with SSPACE and CAP3 programs, respectively (Huang & Madan, 1999; Bankevich *et al.*, 2012).

3.2.2 Genome annotation, visualization and comparative analysis

The assembled scaffolds were annotated using MITOS WebServer to predict the location of protein coding regions/genes, tRNAs, rRNAs and secondary structures of tRNAs. MITOS is a widely used webserver for the annotation of metazoan mitochondrial genomes due to its advanced annotation methodology (Bernt *et al.*, 2013). The location of PCGs, tRNAs and rRNAs was further confirmed using tools like NCBI (National Center for Biotechnology Information) – BLAST, BioEdit and Clustal Omega by comparing the sequences of *A. assamensis* with the respective sequences published for other lepidopteran insects (Hall, 1999; Thompson *et al.*, 2002). The boundaries of PCGs i.e. initiation and termination codons were determined using NCBI ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>) by specifying the invertebrate mitochondrial genetic code. Further, the number of overlapping or spacer regions between the genes were visualized and calculated manually. The control region was validated by comparing with the available sequences in GenBank and the tandem repeats in this region were searched through Tandem Repeat Finder (Benson, 1999). Then, the whole mito-map was constructed and visualized using Blast Ring Image Generator (BRIG) tool (Alikhan *et al.*, 2011). Finally, the complete annotated file of mitochondrial genome was prepared using NCBI Sequin tool (<http://www.ncbi.nlm.nih.gov/Sequin/>) and the sequin file along with SRA data were submitted to NCBI GenBank.

For understanding evolutionary relationships of *A. assamensis* with various lepidopterans as listed in **Table 3.1**, the sequences of whole mitogenome, coding regions, tRNAs, rRNAs and control region were retrieved from NCBI GenBank database and a comparative analysis was performed.

Table 3.1 List of species considered for comparative mitogenome study. (Highlighted ones are the organisms of Bombycoidea superfamily used for comparative mitogenome analysis with respect to *A. assamensis*).

Superfamily	Family	Species	Size (bp)	Accession No.	References
Bombycoidea	Saturniidae	<i>Antheraea assamensis</i>	15,272	KU379695	This study
		<i>Attacus atlas</i>	15,282	KF006326	Chen <i>et al.</i> , 2014
		<i>Samia cynthia ricini</i>	15,384	JN215366	Kim <i>et al.</i> , 2012
		<i>Antheraea pernyi</i>	15,575	AY242996	Liu <i>et al.</i> , 2008
		<i>Antheraea yamamai</i>	15,338	EU726630	Kim <i>et al.</i> , 2009
		<i>Antheraea frithi</i>	15,338	NC_027071	GenBank
		<i>Actias selene</i>	15,236	NC_018133	Liu <i>et al.</i> , 2012
		<i>Saturnia boisduvalii</i>	15,360	NC_010613	Hong <i>et al.</i> , 2008
		<i>Eriogyna pyretorum</i>	15,327	FJ685653	Jiang <i>et al.</i> , 2009
	Bombycidae	<i>Bombyx mandarina</i>	15,682	AY301620	Pan <i>et al.</i> , 2008
		<i>Bombyx mori</i>	15,656	AB070264	Yukuhiro <i>et al.</i> , 2002
		<i>Rondotia menciaana</i>	15,301	KC881286	Kong & Yang, 2015
	Sphingidae	<i>Manduca sexta</i>	15,516	EU286785	Cameron & Whiting, 2008
		<i>Sphinx morio</i>	15,299	KC470083	Kim <i>et al.</i> , 2013
Geometroidea	Geometridae	<i>Biston panterinaria</i>	15,517	NC_020004	Yang <i>et al.</i> , 2013
		<i>Phthonandria atrilineata</i>	15,499	NC_010522	Yang <i>et al.</i> , 2009
Cossoidea	Cossidae	<i>Eogystia hippophaecolus</i>	15,431	KC831443	Gong <i>et al.</i> , 2014
Papilionoidea	Riodinidae	<i>Abisara fylloides</i>	15,301	HQ259069	Zhao <i>et al.</i> , 2013
	Lycaenidae	<i>Spindasis takanonis</i>	15,349	HQ184266	Kim <i>et al.</i> , 2011
		<i>Protantigius superans</i>	15,248	HQ184265	Kim <i>et al.</i> , 2011
	Papilionidae	<i>Papilio maraho</i>	16,094	FJ810212	Wu <i>et al.</i> , 2010
		<i>Parnassius bremeri</i>	15,389	NC_014053	Kim <i>et al.</i> , 2009
	Nymphalidae	<i>Danaus plexippus</i>	15,314	NC_021452	GenBank
		<i>Eumenis autonoe</i>	15,489	GQ868707	Kim <i>et al.</i> , 2010
Noctuoidea	Lymantriidae	<i>Gynaephora menyuanensis</i>	15,770	KC185412	Yuan & Zhang, 2013
		<i>Lymantria dispar</i>	15,569	NC_012893	GenBank
		<i>Plutella xylostella</i>	16,014	KM023645	Dai <i>et al.</i> , 2016
Yponomeutoidea	Plutellidae				
Tortricoidea	Tortricidae	<i>Adoxophyes honmai</i>	15,680	DQ073916	Lee <i>et al.</i> , 2006
		<i>Cydia pomonella</i>	15,253	JX407107	Shi <i>et al.</i> , 2013
Pyraloidea	Crambidae	<i>Ostrinia furnacalis</i>	14,536	NC_003368	Coates <i>et al.</i> , 2005
		<i>Ostrinia nubilalis</i>	14,535	NC_003367	Coates <i>et al.</i> , 2005
Diptera		<i>Drosophila melanogaster</i>	19,517	NC_001709	deBruijn, 1983
(Outgroup)		<i>Drosophila yakuba</i>	16,019	NC_001322	Clary & Wolstenholme, 1985
		<i>Anopheles gambiae</i>	15,363	NC_002084	Beard <i>et al.</i> , 1993

3.2.2.1 Comparative mitogenome analysis

Comparative mitogenome analysis was carried out in order to find out similarities and differences of *A. assamensis* with other insects in terms of characteristics like genome size, organization, structure, gene content and rearrangement. The sequence homology of *A. assamensis* mitogenome with selected insects was also performed using Clustal Omega.

3.2.2.2 Comparative analysis among protein coding genes (PCGs)

The PCGs of *A. assamensis* were compared with selected organisms based on their number, length, initiation and termination codons. The sequence similarity among the PCGs was determined by aligning the sequences of *A. assamensis* with selected Bombycoids using Clustal Omega. MEGA 6.0 tool was used for determining the gene-by-gene divergences in 13 PCGs in terms of phylogenetically informative sites, conserved sites as well as variable sites (Tamura *et al.*, 2013). The same tool was used for estimating nucleotide substitution and evolutionary rates among the genes in terms of transitions (ts), transversions (tv), rate of synonymous substitutions or synonymous substitution per synonymous site (Ks), rate of non-synonymous substitutions or non-synonymous substitution per non-synonymous site (Ka) and their ratios. Further, the correlation between GC content and Ka/Ks ratio was studied in order to predict the effect of GC content on evolutionary rates of PCGs.

3.2.2.3 Comparative analysis among transfer RNAs (tRNAs)

The length, arrangement, secondary structures and variation in the structures were studied among the selected organisms. Homologous sites in their secondary structures were identified by aligning the tRNA sequences using LocARNA tool and the type/number of mismatches were calculated (Smith *et al.*, 2010).

3.2.2.4 Comparative analysis among ribosomal RNAs (rRNAs)

The number, type, location and length of rRNAs in *A. assamensis* were compared with the selected organisms in order to study the conservation pattern. In addition, sequence homology of rRNAs of *A. assamensis* with the selected organisms was determined and gene-by-gene divergences among the rRNAs were studied using MEGA 6.0 tool (Tamura *et al.*, 2013).

3.2.2.5 Comparative analysis among control region (A+T rich region)

The control region of *A. assamensis* was compared with selected organisms based on the length and location. Multiple sequence alignment of control region was then carried out using Clustal omega and their conserved regions, repeats and indels were visualized using BioEdit tool (Hall, 1999).

3.2.2.6 Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions

The OS and IGS regions of *A. assamensis* mitogenome were compared with selected organisms (**Table 3.1**) in terms of number, length and location. The sequence homology of OS and IGS was determined through sequence alignment using Clustal Omega and BioEdit tools to find conserved motifs.

3.2.2.7 Comparative analysis with respect to nucleotide composition, skewness and codon usage

The nucleotide composition of sequences of whole mitochondrial genome, concatenated and individual PCGs, tRNAs, rRNAs, spacers and control region was calculated using MEGA 6.0 software (Tamura *et al.*, 2013). The base composition skewness was also calculated for all the regions of mitogenome using the formula (E3.1 and E3.2) described by Junqueira *et al.*, 2004.

$$\text{AT skew} = (A-T)/(A+T) \dots\dots\dots \text{E3.1}$$

$$\text{GC skew} = (G-C)/(G+C) \dots\dots\dots \text{E3.2}$$

where A, T, G and C denote the frequencies of respective bases.

The codon usage and relative synonymous codon usage (RSCU) values of PCGs were determined using MEGA 6.0.

3.2.3 Comparative phylogenetic analysis

About 34 organisms representing 13 different families of 8 super families within the order Lepidoptera were considered for phylogenetic analysis along with three organisms belonging to the order Diptera as outgroups (**Table 3.1**). The phylogenetic relationship of *A. assamensis* was elucidated using two criteria's: (i) concatenated nucleotide sequences of 13 PCGs and (ii) concatenated nucleotide sequences of 13 PCGs and 2 rRNAs.

The nucleotide sequences of each PCG were translated via amino acid alignment using MAFFT algorithm in TranslatorX server which were then back translated into nucleotide alignment using GBlocks (Abascal *et al.*, 2010). The individual nucleotide sequences obtained from GBlocks and rRNA sequences were then concatenated for further phylogenetic analysis. Substitution model optimization for each dataset was performed in jModeltest 2.1.7 (Guindon & Gascuel, 2003; Darriba *et al.*, 2012). The GTR+G+I model was selected as the best substitution model for both the datasets to conduct maximum likelihood (ML) and bayesian inference (BI) analysis. Phylogenetic trees for each criterion were inferred using maximum likelihood (ML) method with 1000 bootstrap replicates implemented in RaxmlGUI v1.3 (Silvestro & Michalak, 2012). Similarly, BI analysis was performed in MrBayes v3.2.6 using Markov chain Monte Carlo (MCMC) method for both the datasets (Ronquist *et al.*, 2012). Trees were sampled every 1000 generations and

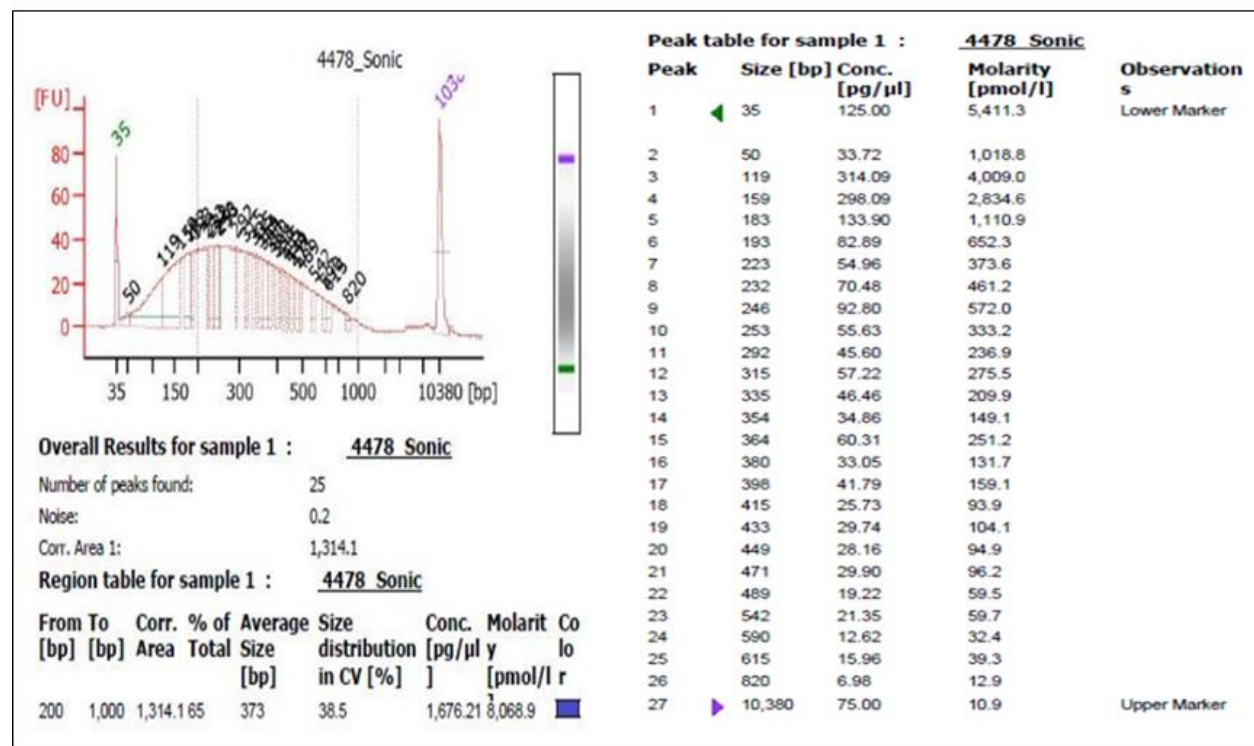
the first 25% were discarded as burn-in. The stationary was considered to reach when average standard deviation of split frequency reached below 0.01 with additional parameters like effective sample size (ESS) and potential scale reduction factor (PSRF) measured using Tracer v1.6 (Rambaut *et al.*, 2014). Finally, the tree topologies were visualized using iTOL v3 (Interactive tree of life) tool (Letunic & Bork, 2006).

3.3 Results and discussion

3.3.1 Sample processing, DNA sequencing and assembly

Total DNA extracted from fifth instar larvae of *A. assamensis* was found to be optimal (1363.23 ng/μl concentration by NanoDrop spectrophotometer and 234.36 ng/μl concentration by Qubit fluorometer) for mitochondrial DNA enrichment. The mitochondrial DNA was enriched from total DNA for library preparation. The library profile showed that size of mitogenome fragments were in the range of 180 to 880 bp. However, total insert size distribution was from 300 to 1000 bp as it involved ~120 bp adapters along with mitogenome fragments (**Figure 3.2**). In addition to the proper distribution of fragments, their concentrations (~27.2 ng/μl) were also found to be adequate for sequencing. The sequencing resulted in 32,12,599 total number of raw reads, out of which around 28,69,153 high quality reads were processed for scaffold preparation. These reads were used for mitogenome scaffold preparation through *de novo* assembly of sequenced contigs using SPAdes-3.5.0, SSPACE and CAP3 program. Finally, a scaffold of 15,272 bp length was obtained which represents the whole mitogenome sequence of *A. assamensis*.

Bioanalyzer profile after sonication of sample



Bioanalyzer profiles of the libraries

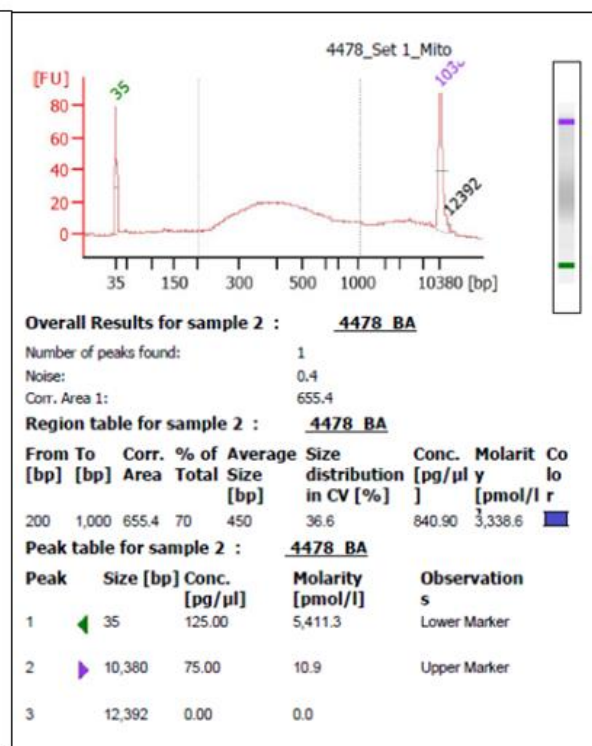


Figure 3.2 Bioanalyzer profiles after sonication of enriched mitochondrial DNA sample and libraries.

3.3.2 Genome annotation, visualization and comparative analysis

The annotated mitochondrial genome of *A. assamensis* appeared to be closed circular structure comprising of 37 genes (13 PCGs, 22 tRNAs and 2 rRNAs) along with a non-coding control region over its 15,272 bp length. 23 genes were found to be encoded by the majority strand or J-strand (+) and the remaining 14 genes by the minority strand or N-strand (-). The whole mito-map constructed using BRIG and its characteristics like gene location and arrangement are shown in **Figure 3.3**. The full annotated mitogenome sequence and SRA data of *A. assamensis* submitted to NCBI GenBank are available under the accession numbers KU379695 and SRR3948351, respectively.

3.3.2.1 Comparative mitogenome analysis

In order to study the conservation pattern, features like genome size, gene content, genome organization, structure, rearrangement and sequence of *A. assamensis* mitogenome were compared with other lepidopterans belonging to various levels of taxonomic classification (**Table 3.1**).

We found that the length of *A. assamensis* mitogenome was smaller than that of other sequenced Bombycoids (**Table 3.2**) and just bigger than that of *Actias selene* (15,236 bp) and *A. artemis* (15,243 bp) (Liu *et al.*, 2012; Park *et al.*, 2016). Notably, it lies within the characteristic size of most of the insects (14 to 20 kb). In insects, mitogenome size variation may be attributed to variation in non-coding regions especially the control region that shows great differences in length as well as pattern. It collectively leads to higher degree of gene rearrangements. In contrast, PCGs are quite stable in the mitogenome (Kim *et al.*, 2012; Chen *et al.*,

2014). Smaller mitogenome size in *A. assamensis* may be favored to reduce accumulation of slightly deleterious mutations during its evolution.

Table 3.2 Comparative gene lengths (in bp) in Bombycoids.

Genes	<i>A. assamensis</i>	<i>S. ricini</i>	<i>A. pernyi</i>	<i>A. yamamai</i>	<i>B. mandarina</i>	<i>B. mori</i>	<i>M. sexta</i>
<i>tRNA^{Met}</i>	68	68	67	67	68	68	69
<i>tRNA^{Ile}</i>	65	65	64	64	66	66	65
<i>tRNA^{Gln}</i>	69	69	69	69	69	69	69
<i>nad2</i>	1014	1014	1014	1014	1023	1023	1015
<i>tRNA^{Trp}</i>	68	68	68	68	70	70	67
<i>tRNA^{Cys}</i>	64	62	61	63	67	67	64
<i>tRNA^{Tyr}</i>	67	65	66	67	66	66	66
<i>cox1</i>	1531	1540	1531	1540	1531	1531	1531
<i>tRNA^{Leu2}</i>	68	68	66	67	67	67	67
<i>cox2</i>	682	685	682	682	682	682	682
<i>tRNA^{Lys}</i>	71	71	71	71	71	71	71
<i>tRNA^{Asp}</i>	67	68	73	69	67	67	72
<i>atp8</i>	165	165	168	186	162	162	165
<i>atp6</i>	678	678	677	677	678	678	678
<i>cox3</i>	789	789	789	789	789	789	792
<i>tRNA^{Gly}</i>	66	66	66	67	69	66	66
<i>nad3</i>	357	357	354	357	351	351	354
<i>tRNA^{Ala}</i>	65	66	66	66	68	68	70
<i>tRNA^{Arg}</i>	65	64	65	65	64	64	68
<i>tRNA^{Asn}</i>	65	65	65	65	67	65	66
<i>tRNA^{Ser1}</i>	66	68	66	66	67	67	66
<i>tRNA^{Glu}</i>	68	66	66	66	63	65	70
<i>tRNA^{Phe}</i>	66	68	71	69	67	67	68
<i>nad5</i>	1741	1749	1741	1744	1719	1719	1723
<i>tRNA^{His}</i>	65	66	69	66	69	67	69
<i>nad4</i>	1341	1341	1341	1341	1344	1341	1335
<i>nad4l</i>	291	291	291	292	291	291	288
<i>tRNA^{Thr}</i>	65	65	65	65	66	65	67
<i>tRNA^{Pro}</i>	65	65	65	65	69	66	65
<i>nad6</i>	537	531	537	535	531	531	531

cytb	1149	1152	1149	1155	1158	1152	1155
tRNA^{Ser2}	66	66	68	69	66	66	66
nad1	933	939	939	933	945	945	936
tRNA^{Leu1}	68	68	68	71	71	70	68
rrnL	1344	1358	1369	1380	1350	1378	1391
tRNA^{Val}	68	66	67	68	71	69	65
rrnS	779	779	775	776	784	783	775
Control region	328	361	552	334	484	494	326
Whole mito-genome	15,272	15,384	15,575	15,338	15,682	15,656	15,516
Total PCGs	11,208	11,231	11,213	11,245	11,204	11,195	11,185
Total tRNAs	1,465	1,463	1,472	1,473	1,488	1,476	1,484
Total rRNAs	2,123	2,137	2,144	2,156	2,134	2,161	2,166

The existence of a typical gene content i.e. 37 genes and a control region was evident in *A. assamensis* mitogenome as observed in other sequenced mitogenomes of lepidopteran insects. The arrangement of mitochondrial genes was found to be in the order identical to other Bombycoidea insects (**Figure 3.4**). Three typical rearrangements were observed in the mitogenome of *A. assamensis* similar to most of the other lepidopteran insects (Lee *et al.*, 2006; Yang *et al.*, 2013; Dai *et al.*, 2016; Gong *et al.*, 2014) as shown in **Figure 3.4**. These were (i) *tRNA^{Met-Ile-Gln}* (*trnM/trnI/trnQ* or *MIQ*) cluster, (ii) *tRNA^{Lys-Asp}* (*K-D*) cluster and (iii) *tRNA^{Ala-Arg-Asn-Ser1-Glu-Phe}* (*ARNS1EF*) cluster. However, atypical rearrangements have also been observed in some ancestral insects (first cluster) and some lepidopteran insects (third cluster) (Cao *et al.*, 2012; Liu *et al.*, 2017).



Figure 3.4 Mitogenome arrangement in selected species. Ancestor here denotes *Thitarodes renzhiensis*.

The typical MIQ cluster rearrangement located between the control region and *nad2* gene differed from the ancestral pattern *tRNA^{Ile-Gln-Met}* (*trnI/trnQ/trnM* or IQM) as reported in two ghost moths (**Figure 3.4**) belonging to Hepialidae family - a non-ditrysian lineage (Lepidoptera: Exoporia: Hepialoidea) (Dowton & Austin, 1999; Cao *et al.*, 2012). As tRNAs gene rearrangements like inversion, translocation, duplication, etc., generally occur in insects, this typical rearrangement of tRNA cluster may be attributed to the translocation of tRNAs as also found in other sequenced lepidopterans. This suggests that *A. assamensis* along with other lepidopteran insects have evolved with a typical gene arrangement after splitting from its stem lineage during the course of time.

Further, sequence similarity performed among selected Bombycoids showed that *A. assamensis* shared highest similarity with *Antheraea* species (*A. yamamai* - 92.1%, *A. pernyi* - 91.7%) followed by the other Bombycoids like *S. ricini* (88.9%), *M. sexta* (85.5%), *B. mori* (83.4%) and *B. mandarina* (82.9%) indicating a significant level of homology among Bombycoidea superfamily of Lepidoptera.

3.3.2.2 Comparative analysis among protein coding genes (PCGs)

The mitogenome of *A. assamensis* comprised of a total of 13 PCGs which spanned 11,208 bp constituting around 73.38% of the total mitogenome. The total size of PCGs and their individual sizes were found to be lying within the range of other Bombycoid species (**Table 3.2**). The PCGs were distributed over both the strands of double stranded mitogenome, 9 genes in J-strand and 4 genes in N-strand.

The ATN sequence was found to be the start codon for all the genes in *A. assamensis* similar to other Bombycoid insects (Yukuhiro *et al.*, 2002; Cameron & Whiting, 2008; Liu *et al.*, 2008; Kim *et al.*, 2009; Kim *et al.*, 2012). These codons include ATT (*nad2*, *atp8*, *nad5*), ATA (*nad3*, *nad6*, *cytb*) and ATG (*atp6*, *cox3*, *nad4*, *nad4l*, *nad1*) (**Table 3.3**). Interestingly, start codons of *cox1* and *cox2* genes in *A. assamensis* were found to deviate from the canonical ATN initiation codon pattern. Similarly, *A. pernyi* (Liu *et al.*, 2008), *A. yamamai* (Kim *et al.*, 2009), *S. boisduvalii* (Hong *et al.*, 2008), *B. mandarina* (Pan *et al.*, 2008), *B. mori* (Yukuhiro *et al.*, 2002) and *M. sexta* (Cameron & Whiting, 2008) were also reported to diverge from this pattern. In addition to the gene sequence similarity analysis, comparative amino acid sequence alignment of the reported lepidopteran insects confirmed CGA and GTG as the initiation codons for both *cox1* and *cox2* genes of

A. assamensis, respectively. CGA is also reported as start codon for *cox1* gene in some of the members of Lepidoptera such as *A. atlas* (Chen *et al.*, 2014), *E. pyretorum* (Jiang *et al.*, 2009), *M. sexta* (Cameron & Whiting, 2008), *A. honmai* (Lee *et al.*, 2006), *Maruca vitrata* (Margam *et al.*, 2011) and dipteran insect such as *Drosophila melanogaster* (Stewart & Beckenbach, 2009). On the other hand, GTG is the start codon for *cox2* gene in *S. ricini* (Kim *et al.*, 2012), *S. boisduvalii* (Hong *et al.*, 2008) and *E. pyretorum* (Jiang *et al.*, 2009).

Further, three genes *cox1*, *cox2* and *nad5* showed incomplete termination codon (only T) instead of canonical termination codon pattern TAA or TAG (**Figure 3.3B & Table 3.3**). This is similar to *A. pernyi* which also shows incomplete termination codon in *cox1*, *cox2*, *nad5* and *nad3* genes (Liu *et al.*, 2008). The existence of incomplete stop codon has been observed as a common phenomenon of mitochondrial genes of metazoans like lepidopterans in order to minimize IGS and OS (Clary & Wolstenholme, 1985; Yukuhiro *et al.*, 2002; Liu *et al.*, 2008; Cameron & Whiting, 2008; Kim *et al.*, 2009). Several studies reveal that it could be a recognition site for an endonuclease that truncates the polycistronic pre-mRNA. These incomplete codons are rectified by the post-transcriptional polyadenylation to yield functional stop codon with TAA termini (Cameron & Whiting, 2008).

Table 3.3 Initiation and termination codons of PCGs in *A. assamensis* with respect to selected members of Bombycoidea superfamily.

Species Name	<i>A. assamensis</i>		<i>S. ricini</i>		<i>A. pernyi</i>		<i>A. yamamai</i>		<i>B. mandarina</i>		<i>B. mori</i>		<i>M. sexta</i>	
Gene	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n	Start Codo n	Stop Codo n
<i>nad2</i>	ATT	TAA	ATT	TAA	ATT	TAA	ATT	TAA	ATA	TAA	ATA	TAA	ATT	T
<i>cox1</i>	CGA	T	ATT	T	TTAG	T	TTAG	TAA	TTAG	T	TTAG	T	CGA	T
<i>cox2</i>	GTG	T	GTG	T	ATG	T	ATG	T	ATG	T	ATG	T	ATG	T
<i>atp8</i>	ATT	TAA	ATC	TAA	ATT	TAA	ATT	TAA	ATA	TAA	ATA	TAA	ATC	TAA
<i>atp6</i>	ATG	TAA	ATG	TAA	ATG	TA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA
<i>cox3</i>	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA
<i>nad3</i>	ATA	TAA	ATA	TAG	ATT	T	ATA	TAG	ATA	TAA	ATA	TAA	ATT	TAA
<i>nad5</i>	ATT	T	ATT	TAA	ATT	T	ATT	T	ATT	TAA	ATT	TAA	ATG	TAA
<i>nad4</i>	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAG	ATG	TAA	ATG	TAA	ATG	TAA
<i>nad4l</i>	ATG	TAG	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA
<i>nad6</i>	ATA	TAA	ATA	TAA	ATA	TAA	ATA	TAA	ATA	TAA	ATA	TAA	ATG	TAA
<i>cytb</i>	ATA	TAA	ATG	TAA	ATA	TAA	ATA	TAA	ATA	TAA	ATA	TAA	ATG	TAA
<i>nad1</i>	ATG	TAA	ATG	TAA	ATG	TAA	ATG	TAA	ATA	TAA	ATA	TAA	ATG	TAA

The PCGs of *A. assamensis* were also compared with seven different organisms belonging to various levels of taxonomic classification in Insecta class at nucleotide level and corresponding amino acid level in order to determine the conserved, variable and phylogenetically informative sites (**Table 3.4**). Various species of *Antheraea* showed 84-90% conserved sites at nucleotide level while at amino acid level it was comparatively higher (86-98%). This might be due to synonymous substitution where non-beneficial mutation is purified by same amino acid encoding codons. However, *atp8* gene exhibited lower values (76.8%) at amino acid level which may suggest that mutation in the gene was needed to help the organism in its evolution. Further, various comparative analyses were carried out for the organisms of different genus, family, superfamily and order. The conserved sites were found to decrease with higher hierarchical organisms while variable and phylogenetically informative sites increased in number. Among all PCGs, *cox1* and *cox2* genes showed maximal conservation both at nucleotide and amino acid levels across the organisms of various classification levels. Similarly, *M. sexta* and *P. maraho* were also reported to exhibit higher conservation for *cox1* and *cox2* genes (Cameron & Whiting, 2008; Wu *et al.*, 2010).

Table 3.4 Gene by gene divergences in the protein coding genes at different taxonomic levels, both at nucleotide level and amino acid levels (in percentage). C here denotes conserved sites, PI denotes phylogenetically informative sites, V denotes variable sites, N denotes nucleotide level, A denotes amino acid level. CVI percentage was calculated as No. of sites/total gene length*100.

Sites (in %)	Hierarchy level	<i>nad2</i>	<i>cox1</i>	<i>cox2</i>	<i>atp8</i>	<i>atp6</i>	<i>cox3</i>	<i>nad3</i>	<i>nad5</i>	<i>nad4</i>	<i>nad4l</i>	<i>nad6</i>	<i>cytb</i>	<i>nad1</i>
C-N	Species	84.1	88.9	89.2	86.3	87.8	87.1	90.5	90.1	89.9	90.5	83.1	85.3	86.3
	Genus	78.8	84	85	81.5	82.4	80.6	80.4	85.9	85.2	85	73.3	78.6	81.6
	Family	70.7	78.5	80	64.3	73.5	71.5	70.4	76.9	76.1	75.9	61.3	70.4	71
	Superfamily	66	76.1	78	56	71.3	68.6	67.9	72.6	70.8	73.8	56.1	66.4	68.1
	Order	51.2	68.5	69.2	41.7	62.5	60.7	58.4	59.7	59.7	58.2	41.9	58.1	59.5
V-N	Species	12.7	10.4	10.4	13.7	11.9	12.5	9.22	9.05	9.81	8.5	16.3	13.2	11.5
	Genus	17.9	15.2	14.6	18.5	17.4	19.1	19.3	13.4	14.5	13.9	26.1	20.1	16.9
	Family	27.2	20.7	19.6	35.7	26.2	28.2	29.3	22.6	23.6	24.1	38.7	28.9	28.1
	Superfamily	31.9	23.1	21.6	44	28.4	31.1	31.8	27.1	28.9	26.2	43.9	32.8	30.9
	Order	46.8	30.9	30.8	58.3	37.2	38.9	41.3	40	40	41.8	58.1	41.1	39.5
PI-N	Species	0	0	0	0	0	0	0	0	0	0	0	0	0
	Genus	3.34	2.72	3.21	2.38	3.82	2.65	1.68	2.56	2.75	2.04	3.89	3.08	3.77
	Family	18.4	12.2	12.3	24.4	18.8	16.7	18.2	14.9	14.9	12.9	24.3	16.1	18.3
	Superfamily	20	14.2	13.3	28.6	20	18.6	19.6	16.5	16.1	15.3	27.4	18.2	19.4
	Order	22	15.9	15.8	33.3	22.5	21.7	23.5	18.5	19.3	19	32	21.5	21.5
C-A	Species	86	98.2	96.9	76.8	94.2	93.9	90.8	91.8	91.3	87.8	83.3	92.5	91.8
	Genus	81.1	97.9	94.3	69.6	92	90.9	80.7	88.5	87.5	81.6	74.4	88.4	86.2

Chapter 3: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkmoth endemic to Assam and North-eastern part of India

V-A	Family	66.5	91.1	89	42.9	78.3	80.3	68.9	75.2	72.5	67.3	55.6	78.4	67
	Superfamily	61	90.1	86.8	39.3	76.1	77.3	64.7	69.4	67	64.3	52.2	74.8	63.2
	Order	40.1	79.8	71.9	25	61.5	67.8	53.8	53.7	52.7	46.9	31.7	62	51.6
	Species	10.6	1	2.63	21.4	5.31	5.3	8.4	7.4	8.26	10.2	15.6	5.66	5.7
	Genus	15.5	1.4	5.26	28.6	7.52	8.33	18.5	10.8	12.1	16.3	24.4	10	11.9
PI-A	Family	31.2	8.2	10.5	55.4	21.2	18.9	30.3	24.1	27	31.6	43.9	20.6	31.8
	Superfamily	36.7	9.1	12.7	58.9	23.5	22	34.5	30.1	32.6	34.7	47.2	24.2	35.5
	Order	57.9	19.6	28.1	73.2	38.1	31.4	45.4	45.8	46.9	52	67.8	37	47.2
	Species	0	0	0	0	0	0	0	0	0	0	0	0	0
	Genus	2.9	0.2	1.32	1.79	2.65	1.14	0.84	2.6	2.01	1.02	1.7	1.8	1.3
	Family	24.4	6	6.14	37.5	18.6	13.6	22.7	19.8	19	20.4	33.3	11.6	20.4
	Superfamily	25.5	6.2	6.58	37.5	19	15.5	23.5	20.5	19.6	22.4	35	12.6	21.7
	Order	27.5	6.4	8.3	44.6	20.4	16.3	26.1	21.5	22.5	25.5	38.9	13.4	23.9

The study of nucleotide variation pattern (transition-transversion ratio) among the PCGs of *A. assamensis* with respect to various members (*A. pernyi*, *A. yamamai*, *S. ricini*, *M. sexta*, *B. mori* and *B. mandarina*) of Bombycoidea superfamily revealed that the transition-transversion ratio decreased with distant organisms in comparison with closer ones across all the PCGs (**Table 3.5**) (DeSalle *et al.*, 1987). For example, ratio for *atp8* gene varied from 1.7 to 0.1 when *A. assamensis* was compared with its close relative *A. pernyi* (same genus) to distant organism *M. sexta* (different family). However in case of *nad4l* gene, a lower ratio was observed for *A. pernyi* (0.24) than distant ones. Further comparative analysis of all the sequenced species of *Antheraea* genus was carried out in order to find out variation in *nad4l* gene within the genus. The *nad4l* gene of *A. assamensis* showed lower ratio across the species whereas higher ratio was observed in comparison with other species. Interestingly it exhibited higher ratio with *S. ricini* which belongs to different genus of same family. This apparently depicts that *nad4l* gene is a decisive marker for evolution to form *A. assamensis* within *Antheraea* genus. Further we may hypothesize that *A. assamensis* might have acquired gene from *S. ricini* during its evolution as both the organism are found in Northeast India and share the same evolutionary space (Sarmah, 2011).

Table 3.5 Nucleotide variation pattern in terms of transitions (ts) and transversions (tv) ratio among PCGs of *A. assamensis* and other selected members.

Gene	A. <i>pernyi</i>	A. <i>yamamai</i>	S. <i>ricini</i>	B. <i>mori</i>	B. <i>mandarina</i>	M. <i>sexta</i>
<i>nad2</i>	1.522	1.428	1.192	0.368	0.328	0.609
<i>cox1</i>	1.143	0.997	0.799	0.504	0.435	0.745
<i>cox2</i>	0.808	1.112	0.971	0.731	0.489	0.602
<i>atp8</i>	1.681	0.767	0.801	0.221	0.221	0.104
<i>atp6</i>	0.980	1.041	0.799	0.394	0.354	0.471
<i>cox3</i>	1.905	1.353	0.866	0.583	0.538	0.479
<i>nad3</i>	1.420	1.026	0.649	0.279	0.256	0.411
<i>nad5</i>	0.960	0.898	0.839	0.403	0.434	0.346
<i>nad4</i>	0.826	0.701	0.480	0.379	0.370	0.261
<i>nad4l</i>	0.242	0.460	0.944	0.540	0.291	0.305
<i>nad6</i>	1.384	1.127	0.833	0.261	0.301	0.406
<i>cytb</i>	1.121	1.202	0.916	0.673	0.593	0.634
<i>nad1</i>	0.821	1.427	0.867	0.354	0.358	0.401

where ts= transitional substitutions per site between two sequences; tv= transversional substitutions per site between two sequences

Different evolutionary rates among the genes reflect different functional constraints affecting the mitogenome that varies among species. Therefore, in order to determine the evolutionary rates in *A. assamensis*, the PCGs were compared with that of seven different organisms belonging to various hierarchy levels in Insecta class. The average rate of synonymous substitutions (Ks) and average rate of non-synonymous substitutions (Ka) along with their ratios (Ka/Ks) were calculated for all 13 PCGs. The ratio Ka/Ks less than, greater than and equal to 1 indicates that genes are under negative (purifying) selection, positive (adaptive) selection and neutral evolution, respectively (Meiklejohn *et al.*, 2007). Among 13 PCGs, *atp8* gene encoding ATPase subunit 8 exhibited 1.5 ratio with reference to that of *D. melanogaster* indicating positive/ relaxed selection acting on this gene (**Figure 3.5**).

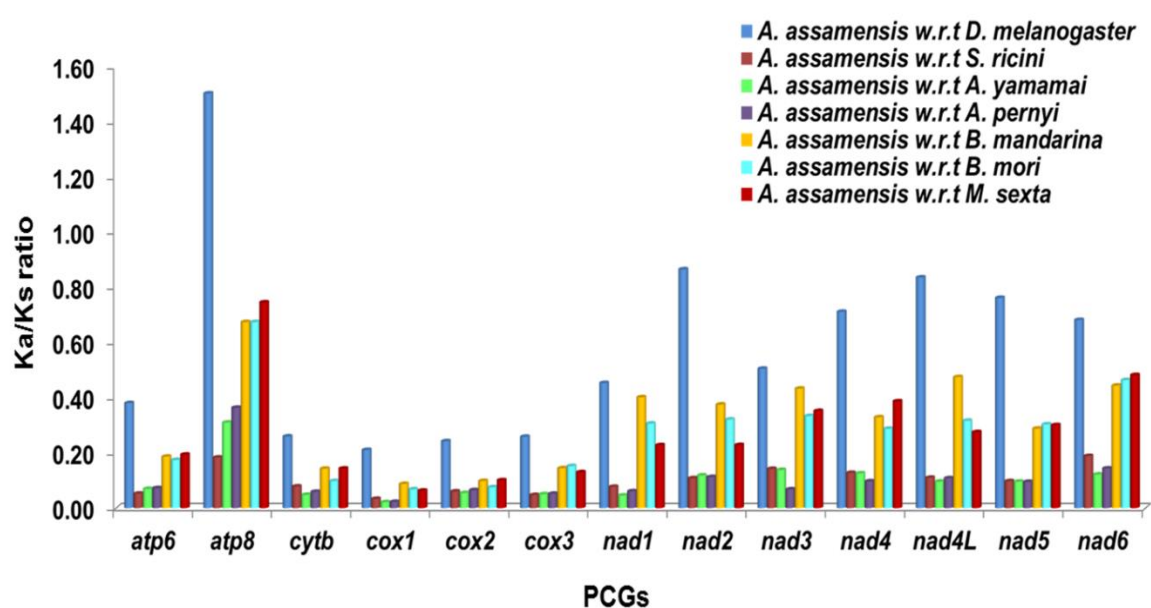


Figure 3.5 Evolutionary rates (Ka/Ks) of individual PCG of *A. assamensis* with selected species. w.r.t. here denotes with reference to.

This indicates that mutation in *atp8* of *A. assamensis* was due to a necessity to re-organize its structure. These changes may be due to the requirement of energy for the production and secretion of silk fibers which are not observed in *D. melanogaster*. It depicts that functional change in an organism needs commitment mutation in responsible genes in order to support survival of the fittest. Notably, the ratio for remaining PCGs of *A. assamensis* with all organisms was found to be less than one which suggests that mutation was against the requirement and hence mutation was replaced by synonymous nucleotides. This indicates that all the protein coding genes evolved under strong purifying (negative) selection as is evident from the number of conserved and variable sites. The Ka/Ks ratio in *atp8* with reference to *B. mandarina*, *B. mori* and *M. sexta* was found to be significantly higher. These organisms belong to different family with respect to *A. assamensis* and hence significant variation in *atp8* gene can be expected due to certain

changes in the mitogenome kinetics. As *S. ricini*, *A. yamamai*, *A. pernyi* belong to the same family of *A. assamensis*, significant Ka/Ks ratio was not observed across the PCGs. Overall comparison of PCGs exhibited that mutation in *atp8* gene was essential in insects to yield several strains with different kinetics as other PCGs showed lower ratios. The lowest evolutionary rates were observed for *cox1* gene indicating that it is least susceptible to variation in protein sequence and hence shows potential as a barcode for evolutionary studies in silkworms. In addition, other genes like *cox2*, *cox3*, *cytb* and *nad* genes with slightly low evolutionary rates next to *cox1* can also serve as barcode markers. Furthermore, the effect of GC content on Ka/Ks ratio was studied and we found a significant negative correlation between them (**Figure 3.6**). This indicates that change in GC content may result in the variation in nucleotide substitution pattern or evolutionary pattern among the PCGs.

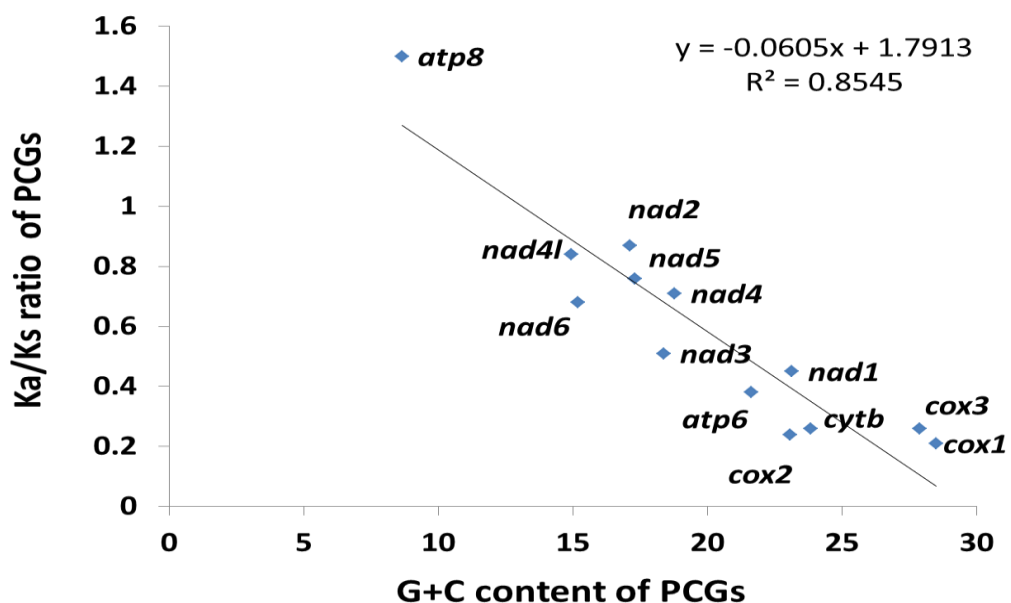


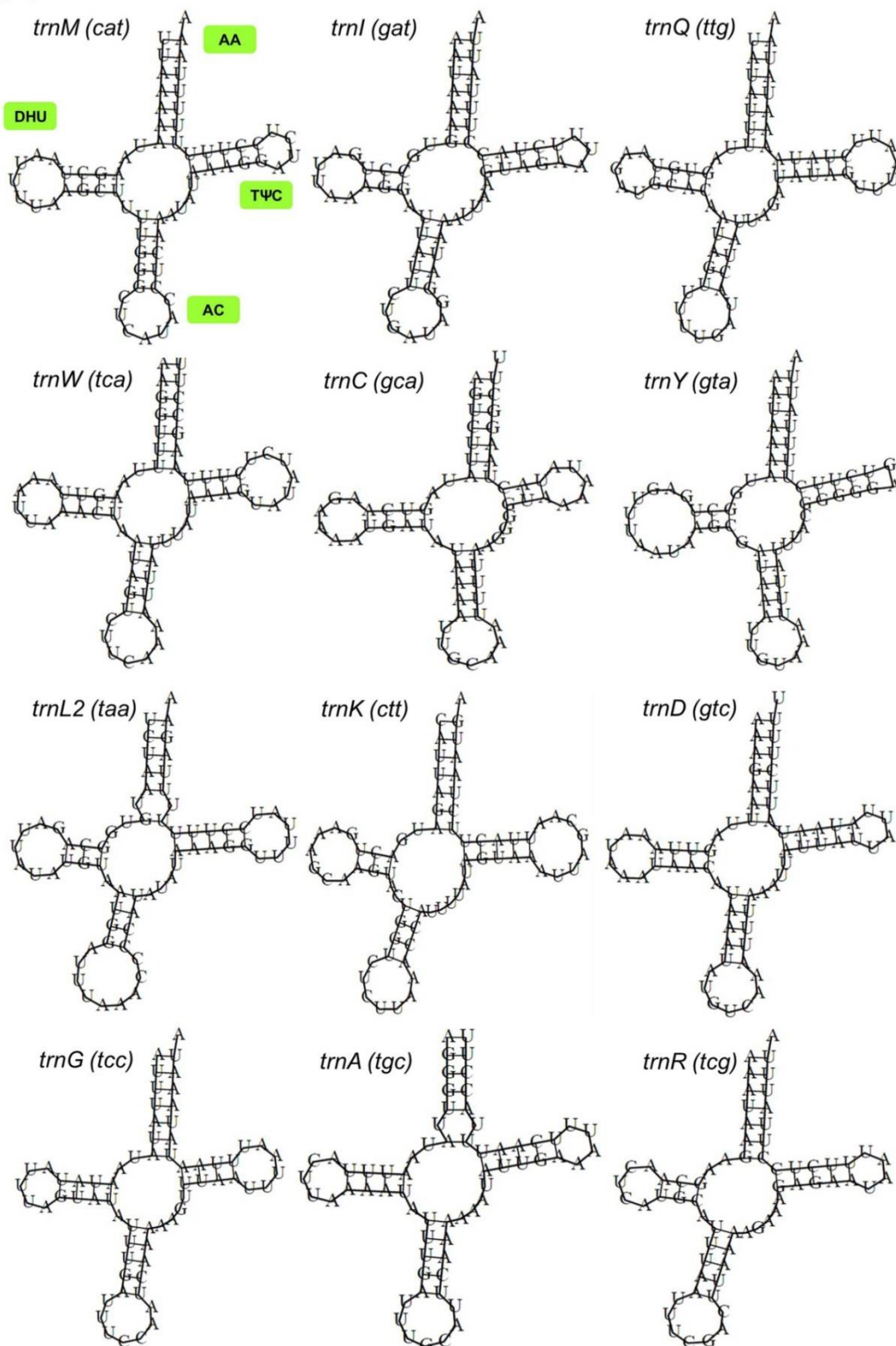
Figure 3.6 G+C content versus Ka/Ks ratios of 13 concatenated PCGs in *A. assamensis* mitogenome.

3.3.2.3 Comparative analysis among transfer RNAs (tRNAs)

Twenty two tRNA genes with a total length of 1465 bp were found to be present in the mitogenome of *A. assamensis*. 14 tRNA genes were encoded by J- strand while 8 tRNA genes were encoded by N- strand (**Figure 3.3B**). The individual length of each gene varied from 64 to 71 bp as reported for other lepidopterans (**Table 3.2**).

The secondary structures of tRNAs of *A. assamensis* exhibited a typical clover-leaf structure similar to other lepidopteran species (**Figure 3.7**). This depicts that arrangement of all the tRNA genes is mostly conserved among insects. However, some variation in the structures of *A. assamensis* was observed such as aminoacyl acceptor stem of *tRNA^{Met}*, TΨC loop of *tRNA^{Ile}*, *tRNA^{Cys}*, *tRNA^{Tyr}*, *tRNA^{Arg}*, *tRNA^{Asn}*, *tRNA^{Glu}*, *tRNA^{Phe}*, *tRNA^{Trp}*, *tRNA^{Ser2}*, dihydrouridine (DHU) arm of *tRNA^{Ser1}*, etc. The lack of stable DHU arm in *tRNA^{Ser1}* and TΨC loop in *tRNA^{Tyr}* has also been reported in lepidopterans such as *S. ricini*, *A. yamamai*, *A. pernyi*, *B. mori*, *B. mandarina*, *M. sexta*, *E. pyretorum*, *A. fylloides*, *B. panterinaria*, *E. autonoe*, etc. (Yukuhiro *et al.*, 2002; Liu *et al.*, 2008; Kim *et al.*, 2009; Jiang *et al.*, 2009; Kim *et al.*, 2012). These structural variations might be attributed to the variation in number of base pairs responsible for the formation of aminoacyl acceptor (AA) stem, DHU arm, TΨC arm and anticodon (AC) stem-loop (**Figure 3.8**).

(A)



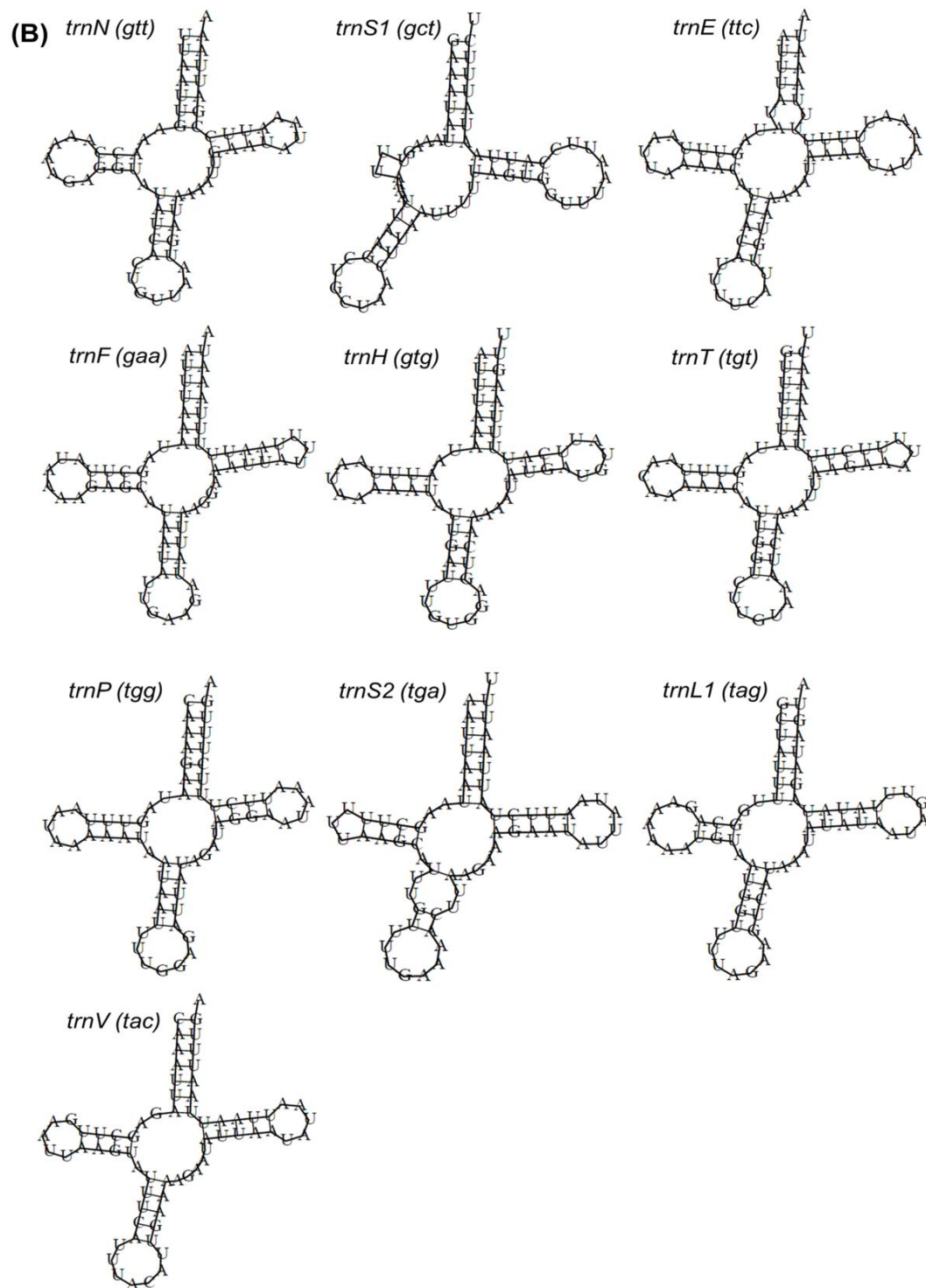
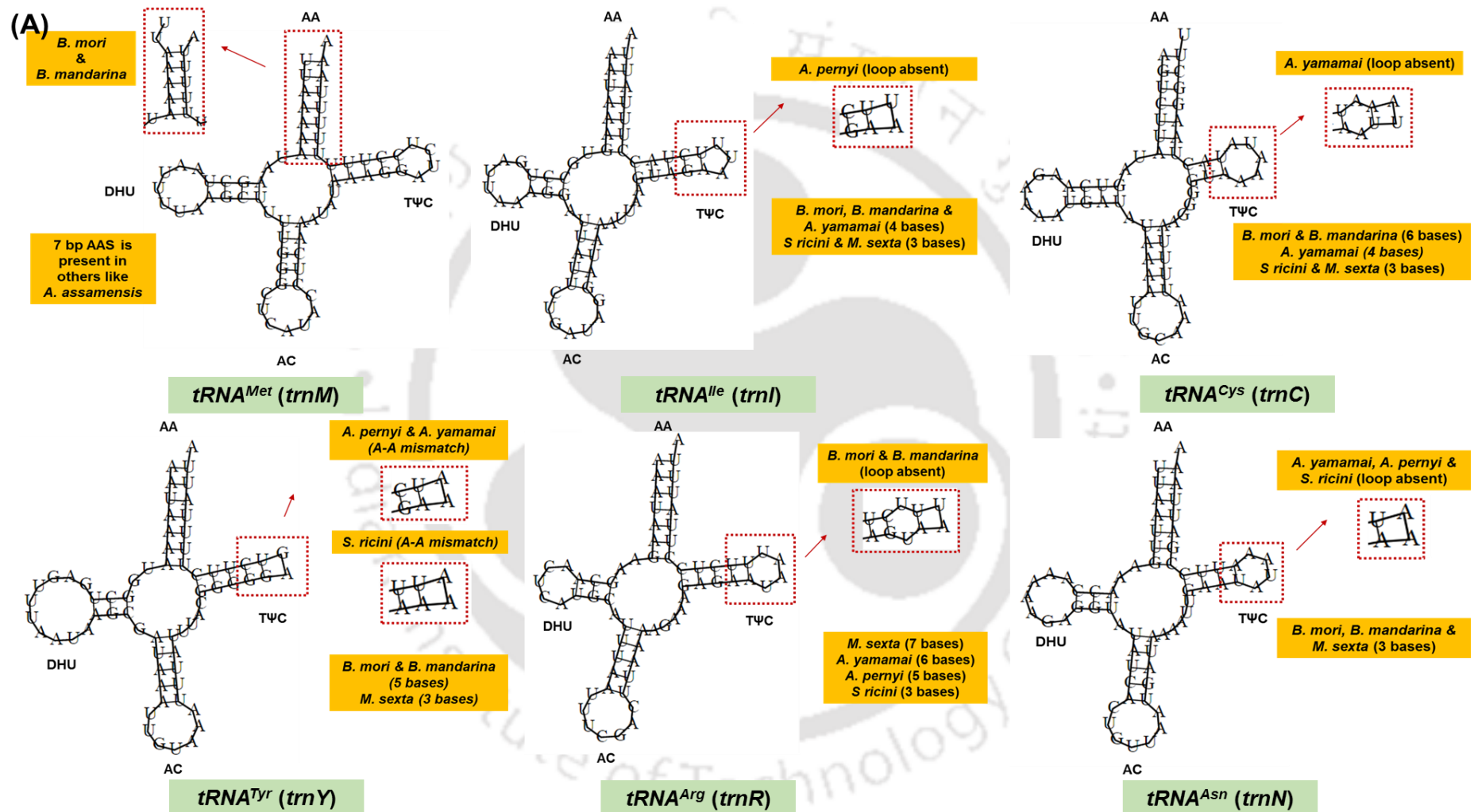


Figure 3.7 Secondary structures of 22 tRNAs of *A. assamensis*. (A) *trnM* to *trnR* (B) *trnN* to *trnV*. *trnY* lacked TΨC loop and *trnS1* lacked DHU loop. tRNA here are represented as *trn*. AA denotes amino acyl arm and AC denotes anticodon arm.

Among various structural parts of tRNAs, AA stem of *A. assamensis* was found to be more consistent in size across all the tRNAs and in other lepidopterans as observed from comparative analysis. AC stem, AC loop and DHU stem also showed consistency in the number of base pairs across lepidopterans, however, varied with the type of tRNAs. In addition, anticodons of tRNAs were found to be identical to the selected Bombycoids (**Table 3.6**) and majority of other lepidopteran insects (Hong *et al.*, 2008; Pan *et al.*, 2008; Kim *et al.*, 2009; Yang *et al.*, 2013). On the contrary, TΨC stem, TΨC loop and DHU loop exhibited randomness in the number of base pairs across the organisms and tRNAs. For instance, TΨC loop was absent in *tRNA^{Tyr}* of *A. assamensis* while *tRNA^{Trp}* of *A. assamensis* and *tRNA^{Tyr}* of *B. mori* had 7 bp and 5 bp loops, respectively. The consistency in the number of base pairs of AA stem, AC stem, AC loop and DHU stem was probably to avoid dysfunctioning of tRNAs actively participating in protein synthesis and thus critical for the survival in the evolutionary milieu.

We identified a total of 35 mismatched base pairs in tRNAs of *A. assamensis* scattered over the AA stem, TΨC stem, AC and DHU regions of 19 tRNAs. 28 mismatches were observed in G-U combinations while 6 and 1 mismatches were found to be in U-U and G-A combinations, respectively (**Table 3.6**). These mismatches lead to changes in tRNA structures; for instance two U-U mismatches in *tRNA^{Ser2}* resulted in the formation of an extra loop in AC stem.



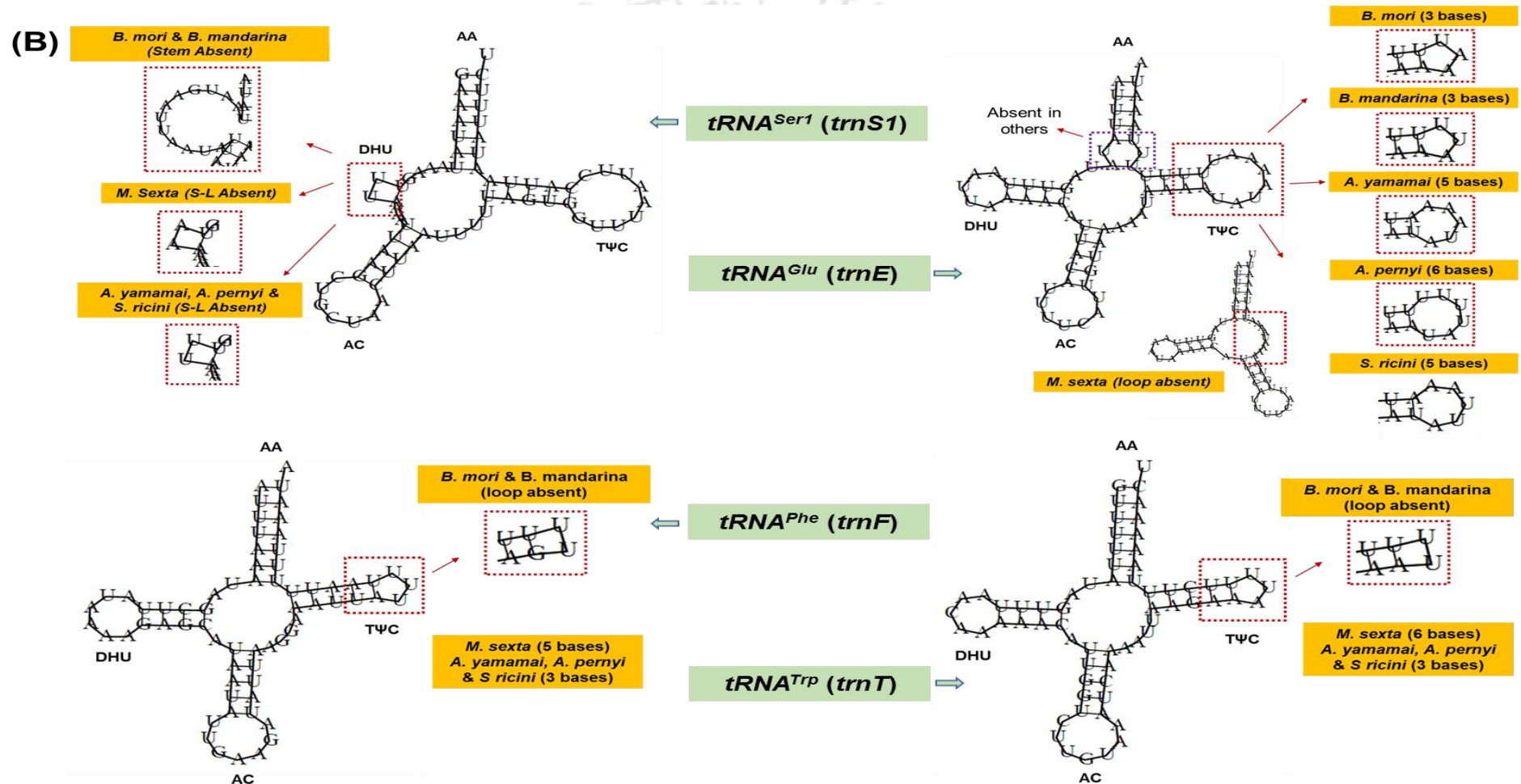


Figure 3.8 (A&B) Variation in the secondary structures of tRNAs in *A. assamensis* and selected Bombycoids. AA denotes Aminoacyl arm and AC denotes Anticodon arm. S-L denotes Stem-Loop of tRNA gene.

From the comparative analysis, we found that the number of total mismatches was higher in *A. assamensis* (35 No.s) than the selected Bombycoids like *S. ricini* (32 No.s), *B. mori* (29 No.s), *M. sexta* (25 No.s), etc. Numerically, this mismatch was higher than many Saturniids like *E. pyretorum* (24 No.s), *S. cynthia* (28 No.s), etc. Similarly, mismatches such as G-U, U-U, G-A, A-A, A-C and C-U have been previously reported for many lepidopterans like *E. autonoe* (26 No.s), *A. fylloides* (27 No.s), etc. (Kim *et al.*, 2010; Zhao *et al.*, 2013). Among various mismatches in *A. assamensis*, G-U mismatches were found be the highest (28 No.s) than the selected Bombycoids and other lepidopterans e.g. *A. atlas* (25 No.s) and *E. autonoe* (20 No.s). The details of mismatches found in various stem-loop structures of different tRNAs of Bombycoid members are compared in **Table 3.6**. The non-canonical pairing might occur due to the insertion/deletion of nucleotides or nucleotide substitution of bases within the tRNA genes as a form of ancient insertional editing. The presence of high number of mismatches indicates higher nucleotide substitution in the tRNA sequences of *A. assamensis* as compared to the selected Bombycoids which may result in genome structure and sequence evolution. These mismatches are likely to be corrected by RNA editing mechanism as observed in other arthropods (Lavrov *et al.*, 2000).

Table 3.6 Anticodons (AC) of tRNAs and number of mismatches present in Amino-acyl (AA) stem, TΨC-stem, DHU-stem and Anticodon (AC) stem of mitochondrial tRNAs of *A. assamensis* with respect to the selected Bombycoids.

tRNA	Stem	<i>A. assamensis</i>	<i>S. ricini</i>	<i>A. pernyi</i>	<i>A. yamamai</i>	<i>B. mandarina</i>	<i>B. mori</i>	<i>M. sexta</i>
<i>tRNA^{Met}</i>	AC	CAT	CAT	CAT	CAT	CAT	CAT	CAT
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	DHU Stem	-	-	-	-	-	-	-
<i>tRNA^{Ile}</i>	AC	GAT	GAT	GAT	GAT	GAT	GAT	GAT
	AA Stem	-	-	-	GU-1	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	GU-1	GU-1	GU-1	GU-1	-	-	GU-1
	DHU Stem	-	-	-	-	-	-	-
<i>tRNA^{Gln}</i>	AC	TTG	TTG	TTG	TTG	TTG	TTG	TTG
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	GU-1	GU-1	GU-1	GU-1	-	GU-1	GU-1
<i>tRNA^{Trp}</i>	AC	TCA	TCA	TCA	TCA	TCA	TCA	TCA
	AA Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	GU-1	GU-1	GU-1	GU-1	-	-	-
	DHU Stem	-	-	-	-	-	-	-
<i>tRNA^{Cys}</i>	AC	GCA	GCA	GCA	GCA	GCA	GCA	GCA
	AA Stem	GU-1	GU-1	GU-1	GU-2	GU-1	GU-1	GU-1
	TΨC Stem	-	-	-	AA-1	-	-	-
	AC Stem	-	-	-	-	-	-	-

<i>tRNA^{Tyr}</i>	DHU Stem	GU-1	GU-1	GU-1	-	-	-	-
	AC	GTA	GTA	GTA	GTA	GTA	GTA	GTA
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	GU-3, GA-1	AA-1	AA-1	AA-1	-	-	-
<i>tRNA^{Leu2}</i>	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
	AC	TAA	TAA	TAA	TAA	TAA	TAA	TAA
	AA Stem	GU-1, UU-1	UU-1	UU-1	UU-1	UU-1	UU-1	UU-1
<i>tRNA^{Lys}</i>	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	GU-1	GU-1	GU-1	GU-1	-	GU-1	GU-1
	AC	CTT	CTT	CTT	CTT	CTT	CTT	CTT
<i>tRNA^{Asp}</i>	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
<i>tRNA^{Gly}</i>	AC	GTC	GTC	GTC	GTC	GTC	GTC	GTC
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	GU-1
<i>tRNA^{Ala}</i>	DHU Stem	-	-	-	-	-	-	-
	AC	TCC	TCC	TCC	TCC	TCC	TCC	TCC
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
<i>tRNA^{Ala}</i>	AC Stem	-	-	-	-	-	-	-
	DHU Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	AC	TGC	TGC	TGC	TGC	TGC	TGC	TGC

Chapter 3: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkworm endemic to Assam and North-eastern part of India

	AA Stem	GU-1, UU-1	GU-1, UU-1	GU-1, UU-1	GU-1, UU-1	GU-1	GU-1	GU-1, UU-1
<i>tRNA^{Arg}</i>	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
	AC	TCG	TCG	TCG	TCG	TCG	TCG	TCG
	AA Stem	-	-	-	-	-	UU-1	-
<i>tRNA^{Asn}</i>	TΨC Stem	-	-	-	-	-	UU-1	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
	AC	GTT	GTT	GTT	GTT	GTT	GTT	GTT
	AA Stem	GU-1	-	-	-	-	-	-
<i>tRNA^{Ser1}</i>	TΨC Stem	-	AA-1	AA-1	AA-1	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
	AC	GCT	GCT	GCT	GCT	GCT	GCT	GCT
	AA Stem	-	-	-	-	-	-	-
<i>tRNA^{Glu}</i>	TΨC Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	UU-1	UU-1	UU-1	UU-1	-	-	AA-1
	AC	TTC	TTC	TTC	TTC	TTC	TTC	TTC
	AA Stem	UU-1	-	-	GU-1	-	-	-
<i>tRNA^{Phe}</i>	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	-	-	-	-	-	-	-
	AC	GAA	GAA	GAA	GAA	GAA	GAA	GAA
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	GU-1	GU-1, UU-1	GU-1, UU-1	-

<i>tRNA^{His}</i>	AC Stem	-	-	-	-	-	-	-
	DHU Stem	GU-1	GU-1	GU-1	GU-1	GU-2	GU-2	GU-2
	AC	GTG	GTG	GTG	GTG	GTG	GTG	GTG
	AA Stem	GU-1	-	-	-	-	-	-
<i>tRNA^{Thr}</i>	TΨC Stem	-	GU-1	GU-1	-	GU-1	GU-1	-
	AC Stem	GU-1	GU-1	GU-1	-	GU-1	GU-1	UU-1
	DHU Stem	-	-	-	-	GU-1	GU-1	-
	AC	TGT	TGT	TGT	TGT	TGT	TGT	TGT
<i>tRNA^{Pro}</i>	AA Stem	-	GU-1, UU-1	-	-	-	-	-
	TΨC Stem	-	-	-	-	UU-1	UU-1	-
	AC Stem	GU-1	-	GU-1	GU-1	-	GU-1	GU-1
	DHU Stem	-	-	-	-	-	-	-
<i>tRNA^{Ser2}</i>	AC	TGG	TGG	TGG	TGG	TGG	TGG	TGG
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	GU-1	GU-1	GU-1	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
<i>tRNA^{Leu1}</i>	DHU Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	AC	TGA	TGA	TGA	TGA	TGA	TGA	TGA
	AA Stem	-	-	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
<i>tRNA^{Leu1}</i>	AC Stem	UU-2	UU-2	UU-2	UU-2	UU-2	GU-1, UU-2	UU-1
	DHU Stem	-	-	-	-	-	-	-
	AC	TAG	TAG	TAG	TAG	TAG	TAG	TAG
	AA Stem	GU-2	GU-2	GU-2	GU-1	GU-1	GU-1	UU-1
<i>tRNA^{Leu1}</i>	TΨC Stem	-	GU-1	-	-	-	-	-

Chapter 3: Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkworm endemic to Assam and North-eastern part of India

<i>tRNA^{Val}</i>	AC Stem	GU-2	GU-1	GU-2	GU-2	GU-2	GU-2	GU-1
	DHU Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	AC	TAC	TAC	TAC	TAC	TAC	TAC	TAC
	AA Stem	-	GU-1	-	-	-	-	-
	TΨC Stem	-	-	-	-	-	-	-
	AC Stem	-	-	-	-	-	-	-
	DHU Stem	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1	GU-1
	GU Mismatches	28	24	23	22	18	22	17
	UU Mismatches	6	6	5	5	5	7	5
	GA Mismatches	1	2	2	3	0	0	0
	AA Mismatches	0	0	0	0	0	0	1
	Total Mismatches	35	32	30	30	23	29	23

3.3.2.4 Comparative analysis among ribosomal RNAs (rRNAs)

A. assamensis comprised of two genes for rRNA coded by (–) strand of the mitogenome. *rrnL* (1rRNA) is located between *tRNA^{Leu1}* and *tRNA^{Val}* genes, whereas *rrnS* (srRNA) is accommodated between *tRNA^{Val}* and control region as reported for other lepidopterans. The lengths of *rrnL* and *rrnS* were 1344 bp and 779 bp, respectively (**Table 3.2**) similar to many reported lepidopteran insects (Zhao *et al.*, 2013; Yuan & Zhang, 2013). The sequence homology of two rRNA genes of *A. assamensis* with the selected Bombycoids revealed that sequence identity of *rrnS* gene (84 to 95 %) was higher than that of *rrnL* (80 to 89 %). The divergence study in rRNA genes (*rrnL* and *rrnS*) also exhibited low variability sites which suggest that these are highly conserved sites and may have potential application in molecular systematics. This may be due to the natural effort to preserve the structure and function of rRNA as protein synthesis depends on it.

3.3.2.5 Comparative analysis among control region (A+T rich region)

The control region is the longest non-coding region of *A. assamensis* mitogenome and with a length of 328 bp, is similar to other Bombycoids. This region is located between *rrnS* and *MIQ* cluster. The comparative sequence homology of control region displayed higher similarity with its close relatives such as *A. yamamai* (89.2%) and *A. pernyi* (82.7%) while showed lower similarity with distant organisms such as *B. mori* (~68%). It was found that two sequence blocks of more than 50 bp in length displayed 96-100% similarity with other *Antheraea* species while *S. ricini* (78-96%), *M. sexta* (68-88%) and *B. mori* (73-80%) showed low similarity. Additionally, some conserved structures and repeats were observed in the control region of *A. assamensis* (**Figure 3.9**).

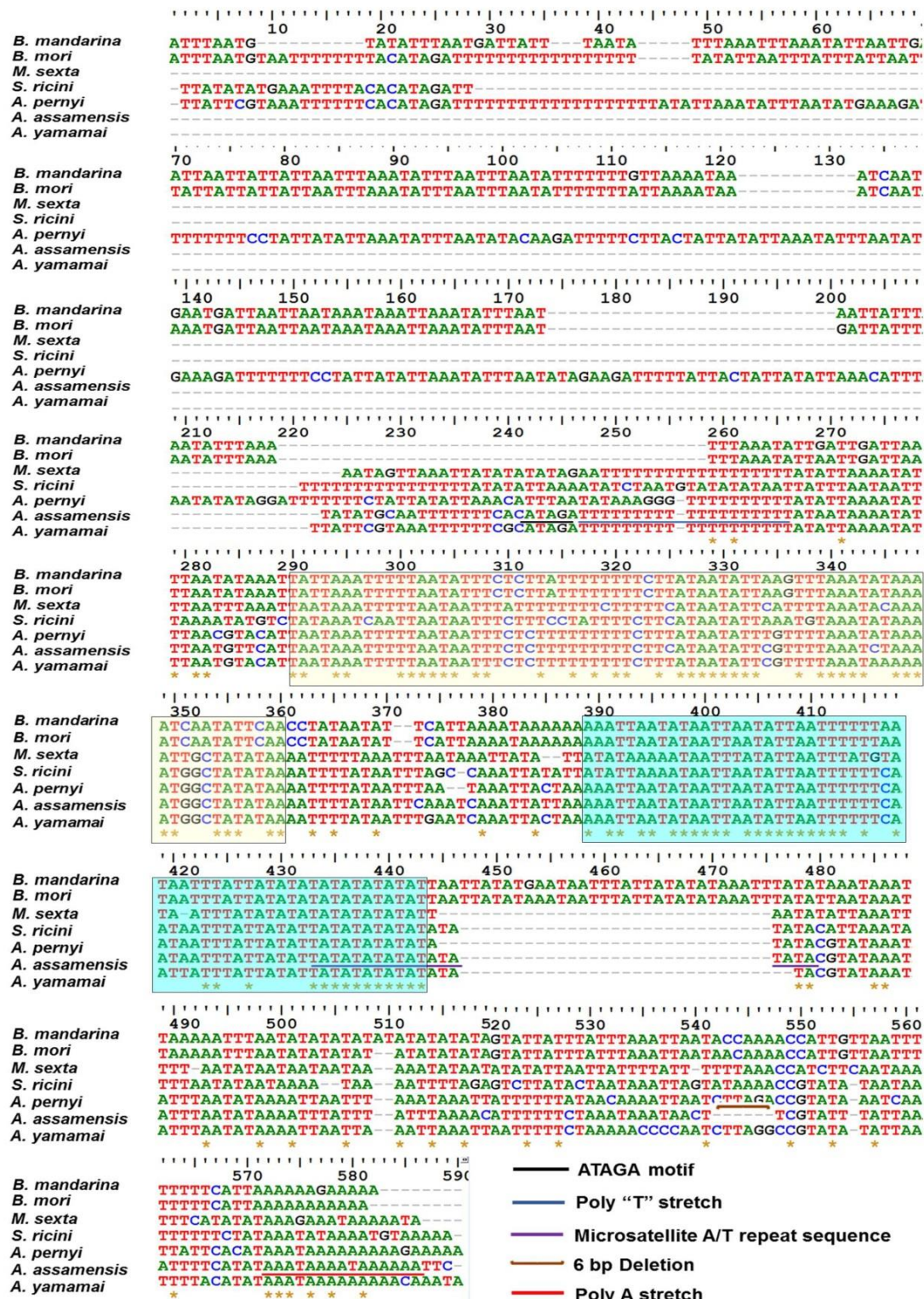


Figure 3.9 Multiple sequence alignment of control region of *A. assamensis* mitogenome showing conserved structural elements. Yellow and green highlighted blocks represent 1st and 2nd conserved blocks, respectively.

These were 'ATAGA' motif, poly T stretch (T)₁₉, microsatellite (TA)₉ repeat sequence, poly-A region and a 6 bp deletion where CTTAGAA/G was found to be conserved among *A. pernyi* and *A. yamamai*. Poly thymidine stretch was found to be variable among these Bombycoids e.g. (T)₁₉ in *A. yamamai*, *A. pernyi*; (T)₁₈ in *S. ricini*, *M. sexta*, *B. mori*; and (T)₁₁ in *B. mandarina*, etc. 'ATAGA' motif and (T)₁₉ are known to be important for gene regulation and serve as recognition site for replication initiation of minor or light strand. The poly A tail has been proposed to be required for RNA maturation and serve as a sequence for controlling transcription or replication initiation in insects (Zhang & Hewitt, 1997). The microsatellites (TA)₉ also known as simple sequence repeats (SSR) are proposed to be used as molecular markers due to their abundance and highly polymorphic nature (Kim *et al.*, 2012).

A microsatellite (AT)_n element is a well conserved site and can be used for evolutionary and conservation genetic studies (Liu *et al.*, 2008; Kim *et al.*, 2009; Stolle *et al.*, 2013). The control region of *A. assamensis* lacked tandem repeat elements similar to other completely sequenced *Antheraea* species except *A. pernyi* which harbored six 38 bp tandem repeats. The significance of these tandem repeats is unclear and needs further taxon sampling and mitogenome characterization. Presence of highly conserved yet distinct consensus sequences in the control region thus shows its high potential as a phylogenetic marker at lower taxonomic levels.

3.3.2.6 Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions

The overlapping sequences (OS) and intergenic spacers (IGS) are commonly found in the mitogenome of lepidopterans. These regions may vary in length and location from species to species during their evolution. In case of *A. assamensis*, six OS and seventeen IGS were found to be present in the mitogenome. The size of OS and IGS were in the range of 1-8 bp (cumulative length of 23 bp) and 1-50 bp (cumulative length of 171 bp) respectively. The presence of OS, IGS and their corresponding lengths were similar to other Bombycoid insects as shown in **Table 3.7**. The OS situated between *tRNA^{Trp}* and *tRNA^{Cys}* (8 bp) was the longest similar to the other Bombycoid species and also in many other lepidopteran species like *A. selene*, *A. atlas*, *R. menciaana* and *G. menyuanensis* (Liu *et al.*, 2012; Yuan & Zhang, 2013; Chen *et al.*, 2014; Kong & Yang, 2015). A characteristic conserved overlapping junction (ATGATAA) was observed in between *atp8* and *atp6* genes as reported for most of the lepidopterans (Lee *et al.*, 2006; Kim *et al.*, 2010; Zhao *et al.*, 2013).

Table 3.7 Intergenic spacers (IGS) and overlapping sequence (OS) between different genes in *A. assamensis* with respect to the selected Bombycoid species.

Spacer	<i>A. assamensis</i>	<i>S. ricini</i>	<i>A. pernyi</i>	<i>A. yama mai</i>	<i>B. mandarina</i>	<i>B. mori</i>	<i>M. sexta</i>
<i>tRNA^{Met}</i> - <i>tRNA^{Ile}</i>	1	1	9	9	-2	-2	10
<i>tRNA^{Ile}</i> - <i>tRNA^{Gln}</i>	-3	-3	-3	-3	-3	-3	-3
<i>tRNA^{Gln}</i> - <i>nad2</i>	50	54	56	53	48	47	54
<i>nad2</i> - <i>tRNA^{Trp}</i>	7	9	8	8	5	5	0
<i>tRNA^{Trp}</i> - <i>tRNA^{Cys}</i>	-8	-8	-8	-8	-8	-8	-8
<i>tRNA^{Cys}</i> - <i>tRNA^{Tyr}</i>	4	0	19	10	6	6	10
<i>tRNA^{Tyr}</i> - <i>cox1</i>	6	0	2	2	17	15	4

cox1-	0	0	0	-5	0	0	0
tRNA^{Leu2}							
tRNA^{Leu2} -	0	0	0	0	0	0	0
cox2							
cox2 -	0	0	0	0	0	0	0
tRNA^{Lys}							
tRNA^{Lys} -	17	33	23	16	-1	-1	21
tRNA^{Asp}							
tRNA^{Asp} -	0	0	0	0	0	0	0
atp8							
atp8 -	-7	-7	-7	-7	-7	-7	-7
atp6							
atp6 -	-1	-1	0	-1	12	14	18
cox3							
cox3 -	2	2	2	2	2	2	6
tRNA^{Gly}							
tRNA^{Gly} -	-3	-3	0	-3	3	3	0
nad3							
nad3 -	-1	-2	0	-2	53	54	20
tRNA^{Ala}							
tRNA^{Ala} -	0	24	-1	-1	50	31	1
tRNA^{Arg}							
tRNA^{Arg} -	0	26	0	0	0	2	50
tRNA^{Asn}							
tRNA^{Asn} -	2	13	1	2	0	1	3
tRNA^{Ser1}							
tRNA^{Ser1} -	0	3	1	1	10	17	0
tRNA^{Glu}							
tRNA^{Glu} -	11	-2	10	6	-1	-1	-2
tRNA^{Phe}							
tRNA^{Phe} -	0	0	0	0	4	4	0
nad5							
nad5 -	0	-3	0	-3	16	21	12
tRNA^{His}							
tRNA^{His} -	4	1	9	12	59	56	0
nad4							
nad4 -	1	3	5	3	-1	-1	84
nad4l							
nad4l -	7	12	7	5	4	4	13
tRNA^{Thr}							
tRNA^{Thr} -	0	0	0	0	-1	0	-1
tRNA^{Pro}							
tRNA^{Pro} -	2	2	2	2	1	2	1
nad6							
nad6 -	2	7	2	-1	55	49	17
cytb							
cytb -	31	5	25	24	-1	3	21
tRNA^{Ser2}							
tRNA^{Ser2} -	23	24	19	24	25	24	31
nad1							

<i>nad1</i> -	1	2	2	1	-6	-6	0
<i>tRNA^{Leu1}</i>							
<i>tRNA^{Leu1}</i> -	0	0	0	0	33	0	0
<i>rrnL</i>							
<i>rrnL</i> -	0	0	0	0	0	-1	0
<i>tRNA^{Val}</i>							
<i>tRNA^{Val}</i> -	0	0	0	0	0	0	0
<i>rrnS</i>							
<i>rrnS</i> -	0	0	0	0	0	0	0
control							
region							
control							
region-	0	0	0	0	0	0	0
<i>tRNA^{Met}</i>							
IGS	17	17	18	17	18	20	18
regions							
Total IGS	171	221	202	180	403	360	376
basepairs							
OS	6	8	4	10	10	9	5
regions							
Total OS	23	29	19	34	31	30	21
basepairs							

The total IGS in *A. assamensis* was lower than selected Bombycoids (**Table 3.7**). However, it was higher than the other selected lepidopterans like *A. selene* (137 bp over 13 regions), *A. atlas* (129 bp over 12 regions), *P. bremeri* (138 bp over 14 regions), *R. menciaana* (149 bp over 17 regions) but lower than *E. pyretorum* (186 bp over 17 regions), *C. boisduvali* (194 bp over 16 regions), *E. hippophaecolus* (289 bp over 18 regions), *P. xylostella* (287 bp over 16 regions), etc.

The spacer between the genes *tRNA^{Gln}* and *nad2* (*tRNA^{Gln}* - *nad2* spacer) was found to be the largest (50 bp) in *A. assamensis* but slightly shorter than that observed in *A. yamamai* (53 bp), *S. cynthia* (54 bp), etc. Similarly, it has been detected as the largest spacer in many other lepidopterans like *A. atlas*, *A. selene*, *S. boisduvali*, *E. pyretorum*, *P. bremeri* and *A. honmai* (Lee *et al.*, 2006; Hong *et al.*, 2008; Jiang *et al.*, 2009; Kim *et al.*, 2009; Liu *et al.*, 2012; Chen *et al.*, 2014). Further, spacer *cytb-tRNA^{Ser2}* (31 bp), *tRNA^{Ser2}-nad1* (23 bp), *tRNA^{Lys}- tRNA^{Asp}*

(17 bp), etc. were also found to exist in *A. assamensis* mitogenome (**Figure 3.3B**).

The spacer *cytb-tRNA^{Ser2}* may serve as a binding site for the mtTERM, which is a transcription termination peptide. A conserved motif 'ATACTAA' was found in the spacer *tRNA^{Ser2}-nad1* on comparison of *A. assamensis* with selected Bombycoid species (**Figure 3.10**). This spacer is implicated in mitochondrial transcription termination where the 'ATACTAA' motif is essential as a recognition site for mtTERM (Cameron & Whiting, 2008; Kim *et al.*, 2009; Kim *et al.*, 2012).



Figure 3.10 Multiple sequence alignment of spacer *tRNA^{Ser2}-nad1* of selected Bombycoids showing a conserved 7 bp motif "ATACTAA".

The spacers in *A. assamensis* mitogenome also exhibited sequence homology with their adjacent genes as observed in many of the lepidopteran insects (Kim *et al.*, 2009; Kim *et al.*, 2012). For instance, spacer *tRNA^{Gln}-nad2* showed 68% similarity with its adjacent *nad2* gene. Similarly, around 77%, 71%, 68%, 70%, 76% and 72% similarity was reported earlier in *S. ricini*, *A. pernyi*, *A. yamamai*, *B. mandarina*, *B. mori* and *M. sexta* respectively. Such homology has also been observed in many other lepidopterans like *A. selene* (94.23%), *A. atlas* (87.5%), *C. boisduvali* (75%), *P. bremeri* (70%), etc. This might be because of the partial duplication of *nad2* gene in order to provide another origin of replication (Hong *et al.*, 2008; Cameron & Whiting, 2008; Kim *et al.*, 2009). This also suggests that *A. assamensis* may have undergone rapid sequence divergence attributed to the

non-coding nature of IGS as seen in other organisms (Kim et al., 2009). Apart from adjacent genes, IGS also showed sequence homology with their distant genes in *A. assamensis*. For example, spacer $tRNA^{Lys}$ - $tRNA^{Asp}$ showed homology with *rrnS* (82%), *nad5* (76%) and *atp8* (82%) genes. Similarly, same spacer in *S. ricini* has also been reported to exhibit homology with *rrnS* (78%) and *nad5* (72%) genes (Kim et al., 2012). Other spacers like *cytb*- $tRNA^{Ser2}$ and $tRNA^{Ser2}$ -*nad1* also displayed similar kind of homology as shown in **Figure 3.11**. This might be attributed to gene duplication and degeneration.

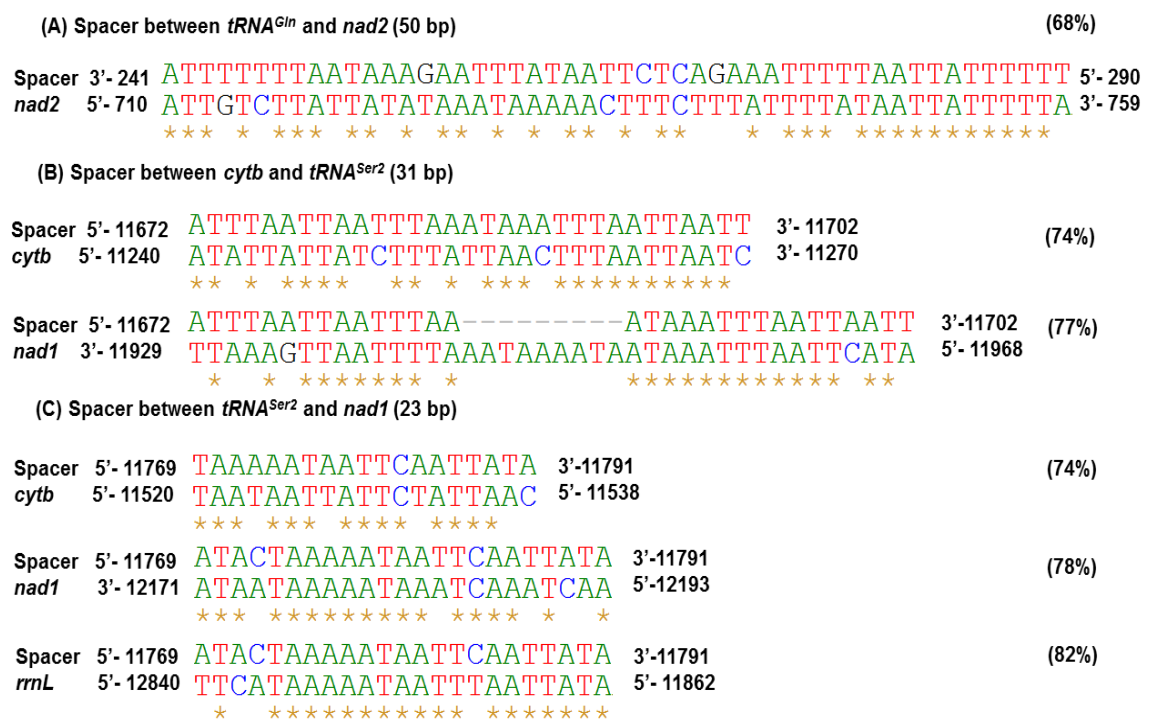


Figure 3.11 Sequence alignments of relatively large intergenic spacers (IGS) of *A. assamensis*. (A) Alignment of the IGS located between $tRNA^{Gln}$ and *nad2* with its neighboring partial gene *nad2*. (B) Alignment of the IGS located between *cytb* and $tRNA^{Ser2}$ with partial neighboring genes *cytb* and *nad1* genes. (C) Alignment of the IGS located between $tRNA^{Ser2}$ and *nad1* with partial *cytb*, *nad1* and *rrnS* genes. Asterisks represent the consensus sequences between the IGS and partial genes. Percentage values in () denote the sequence homology between IGS and partial genes.

3.3.2.7 Comparative analysis with respect to nucleotide composition, skewness and codon usage

The nucleotide composition in mitogenome of *A. assamensis* was analyzed in terms of specific nucleotide content, AT and GC skewness. Whole mitogenome nucleotide composition revealed 80.2% AT content (40.8% T + 39.4% A) and remaining 19.8% GC content (12.1% C + 7.7% G). Similar to *A. assamensis*, many lepidopterans insects also showed enhanced AT content (Kim *et al.*, 2011; Gong *et al.*, 2014). This AT content was found to vary with different regions of the mitogenome (**Figure 3.12**). The AT content was maximal (~91.2%) in the control region followed by rRNAs genes (84.3%), tRNAs genes (80.8%) and PCGs (78%). The individual PCGs also exhibited variation in the corresponding AT content (**Table 3.8**). For instance, *atp8*, *nad4l* and *nad6* genes had the highest AT content while *cox1* and *cox3* genes had lowest values. This variation pattern was also detected in the other lepidopterans like *E. pyretorum*, *P. atrilineata*, *G. menyuanensis*, etc. (Jiang *et al.*, 2009; Yang *et al.*, 2009; Yuan & Zhang, 2013). A close analysis of the third codon position in all 13 PCGs of *A. assamensis* exhibited AT biasness which was 93% than the first (73%) and second (70%) codon positions (**Table 3.9**) similar to the other Bombycoids (Liu *et al.*, 2008; Jiang *et al.*, 2009; Kim *et al.*, 2012). This may be attributed to more relaxed constraints on A+T content at 3rd position than at 1st and 2nd position due to the phenomena of degeneracy in genetic code.

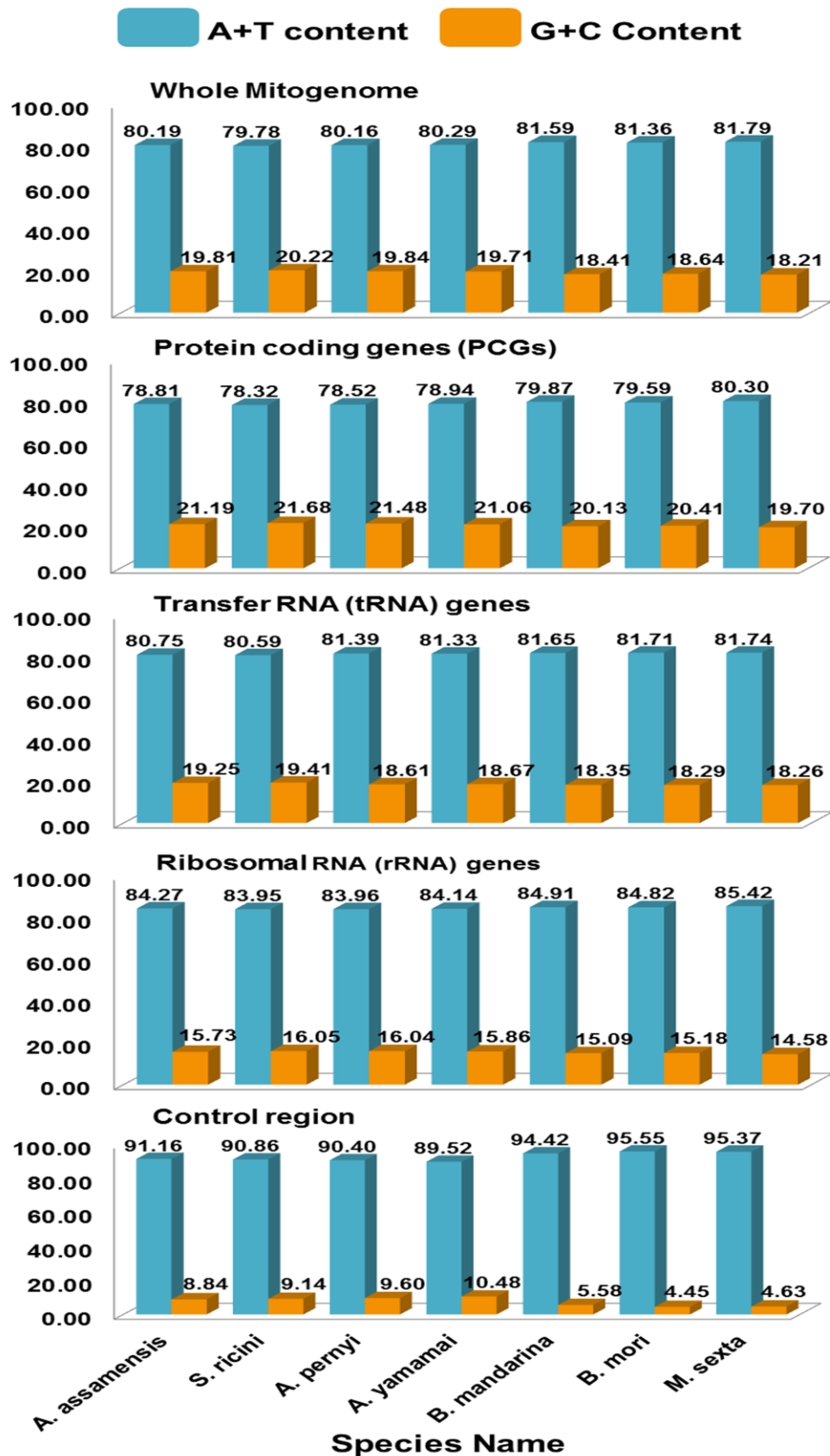


Figure 3.12 Nucleotide compositions of different regions of mitogenome of *A. assamensis* with respect to other Bombycoids.

Table 3.8 A+T content of individual PCGs in Bombycoids.

Gene	<i>A. assamensis</i>	<i>S. ricini</i>	<i>A. pernyi</i>	<i>A. yamamai</i>	<i>B. mandarina</i>	<i>B. mori</i>	<i>M. sexta</i>
<i>nad2</i>	82.89	82.89	82.10	82.89	87.35	85.98	85.60
<i>cox1</i>	71.50	71.15	70.98	71.57	73.01	72.76	72.68
<i>cox2</i>	76.95	75.73	76.36	76.36	77.09	76.36	77.97
<i>atp8</i>	91.36	91.36	91.52	91.52	88.68	87.42	89.51
<i>atp6</i>	78.37	79.56	78.58	79.26	80.74	79.70	80.89
<i>cox3</i>	72.14	71.88	71.63	73.79	75.06	74.05	75.67
<i>nad3</i>	81.64	80.79	81.77	83.62	83.62	83.05	84.05
<i>nad5</i>	82.70	81.44	81.67	82.33	81.70	82.11	83.04
<i>nad4</i>	81.24	80.72	81.17	81.17	80.54	81.02	81.91
<i>nad4l</i>	85.07	84.38	85.07	85.42	85.42	83.33	84.91
<i>nad6</i>	84.83	85.23	85.58	86.14	87.12	87.88	87.50
<i>cytb</i>	76.18	74.59	75.92	75.52	76.19	76.15	77.52
<i>nad1</i>	76.88	77.24	77.67	76.67	79.19	79.09	79.10

Table 3.9 Base composition at each codon position of the concatenated 13 protein-coding genes in Bombycoidea superfamily members.

Species Name	1st codon position					2nd codon position					3rd codon position				
	T-1	C-1	A-1	G-1	A+T	T-2	C-2	A-2	G-2	A+T	T-3	C-3	A-3	G-3	A+T
<i>A. assamensis</i>	37.1	10.4	35.9	16.6	73.0	48.5	16.6	21.8	13.1	70.4	51.8	4.5	41.1	2.6	92.9
<i>S. ricini</i>	37.0	10.6	35.5	16.5	72.5	48.0	16.6	21.8	13.3	69.8	51.0	4.8	40.5	3.4	91.5
<i>A. pernyi</i>	38.0	10.5	35.4	16.6	73.4	48.0	16.6	21.7	13.3	69.7	51.0	4.9	40.9	2.8	91.9
<i>A. yamamai</i>	37.0	10.7	35.7	16.4	72.7	49.0	16.5	21.7	13.1	70.7	52.0	4.0	41.4	2.6	93.4
<i>B. mandarina</i>	37.0	9.6	37.3	15.6	74.3	49.0	15.8	22.4	13.2	71.4	50.0	3.5	43.6	2.8	93.6
<i>B. mori</i>	38.0	9.5	37.2	15.7	75.2	49.0	15.9	22.2	13.2	71.2	49.0	4.3	43.7	2.7	92.7
<i>M. sexta</i>	38	9.5	37.1	15.7	74.8	49	15.9	22.2	13.1	71.0	51	2.6	43.6	2.5	94.9

In addition to the above regions, non-coding regions were also biased towards A and T nucleotides. Among the IGS, three longest ones- *tRNA^{Gln}* and *nad2*, *cytb-tRNA^{Ser2}* and *tRNA^{Ser2}-nad1* in *A. assamensis* had an A+T content of 92%, 100% and 91% respectively. The AT % of the longest spacer i.e. *tRNA^{Gln}* and *nad2* was similar to *A. pernyi* (92.8%) but lower than *B. mandarina* (95.8%), *B. mori* (95.7) and higher than *S. ricini* (85.2), *M. sexta* (88.9%), *A. yamamai* (90.6%), *A. atlas* (87.50%), etc. (Kim *et al.*, 2009; Kim *et al.*, 2013; Chen *et al.*, 2014). Similarly, the A+T content of *tRNA^{Ser2}-nad1* lies within the range of other Bombycoids like *B. mandarina* (88%) to *M. sexta* (96.8%).

Nucleotide skewness was determined to measure the relative number of Gs to Cs and As to Ts. The AT skewness in *A. assamensis* mitogenome was found to be negative (-0.02) which indicates the occurrence of more Ts than As and lies within the range of Saturniid moths (Liu *et al.*, 2008; Kim *et al.*, 2009). This value differed from the selected Bombycidae species having positive AT skew values e.g. 0.06 in *B. mori*, *B. mandarina*, etc. (Yukuhiro *et al.*, 2002; Pan *et al.*, 2008). Similarly, mitogenome GC skewness also exhibited negative value (-0.22) as observed in many Bombycoids such as *M. sexta* (-0.18) and *S. ricini* (-0.23). The GC skewness varied with major and minor strands of PCGs, however no significant change was observed in AT skewness. Minor strand was G-skewed i.e. rich in G nucleotides while more 'C's were encoded by major strand (**Table 3.10**). The strand asymmetry is common in insects and has been reported in all the lepidopterans which occur due to asymmetrical mutation pressure on mitogenome. These skewness patterns were also observed in rRNAs, tRNAs, control region and are similar to the other sequenced lepidopteran insects.

Table 3.10 AT skewness and GC skewness in the whole mitogenome sequence, protein coding genes (PCGs), tRNAs, rRNAs and control region of *A. assamensis* and the selected Bombycoid species.

Species Name	Whole mitogenome		Majority strand		Minority strand	
	AT Skew	GC Skew	AT Skew	GC Skew	AT Skew	GC Skew
<i>A. assamensis</i>	-0.02	-0.22	-0.13	-0.13	-0.12	0.15
<i>S. ricini</i>	-0.01	-0.23	-0.12	-0.13	-0.13	0.08
<i>A. pernyi</i>	-0.02	-0.22	-0.13	-0.13	-0.12	0.06
<i>A. yamamai</i>	-0.02	-0.22	-0.13	-0.13	-0.12	0.08
<i>B. mandarina</i>	0.06	-0.20	-0.04	-0.09	-0.15	0.08
<i>B. mori</i>	0.06	-0.22	-0.03	-0.12	-0.15	0.07
<i>M. sexta</i>	-0.01	-0.18	-0.10	-0.07	-0.13	0.08
Species Name	PCGs		Majority strand PCGs		Minority strand PCGs	
	AT Skew	GC Skew	AT Skew	GC Skew	AT Skew	GC Skew
<i>A. assamensis</i>	-0.16	0.01	-0.14	-0.14	-0.19	0.32
<i>S. ricini</i>	-0.16	0.02	-0.14	-0.15	-0.20	0.33
<i>A. pernyi</i>	-0.17	0.01	-0.15	-0.15	-0.19	0.31
<i>A. yamamai</i>	-0.16	0.02	-0.15	-0.15	-0.18	0.32
<i>B. mandarina</i>	-0.14	0.05	-0.05	-0.10	-0.27	0.31
<i>B. mori</i>	-0.13	0.03	-0.05	-0.13	-0.27	0.33
<i>M. sexta</i>	-0.14	0.05	-0.11	-0.09	-0.19	0.31
Species Name	tRNA		Majority strand tRNAs		Minority strand tRNAs	
	AT Skew	GC Skew	AT Skew	GC Skew	AT Skew	GC Skew
<i>A. assamensis</i>	0.02	0.18	0.01	0.02	0.04	0.46
<i>S. ricini</i>	0.01	-0.15	0.03	0.02	-0.02	-0.44
<i>A. pernyi</i>	0.00	-0.14	0.02	0.03	-0.03	-0.43
<i>A. yamamai</i>	0.01	-0.13	0.04	0.02	-0.04	-0.42
<i>B. mandarina</i>	0.03	-0.15	0.05	0.01	0.00	-0.44
<i>B. mori</i>	0.03	-0.15	0.05	0.02	0.00	-0.46
<i>M. sexta</i>	0.00	-0.11	0.01	0.03	-0.01	-0.38
Species Name	rRNA		<i>rrnL</i>		<i>rrnS</i>	
	AT Skew	GC Skew	AT Skew	GC Skew	AT Skew	GC Skew
<i>A. assamensis</i>	-0.03	-0.36	-0.02	-0.37	-0.05	-0.34
<i>S. ricini</i>	-0.02	-0.39	-0.01	-0.38	-0.04	-0.40
<i>A. pernyi</i>	-0.03	-0.39	-0.01	-0.40	-0.05	-0.37
<i>A. yamamai</i>	-0.03	-0.36	-0.02	-0.38	-0.05	-0.34
<i>B. mandarina</i>	0.03	-0.37	0.05	-0.39	0.01	-0.32
<i>B. mori</i>	0.03	-0.40	0.05	-0.40	0.00	-0.38
<i>M. sexta</i>	-0.03	-0.34	-0.03	-0.33	-0.04	-0.35
Species Name	Control Region					
	AT Skew			GC Skew		
<i>A. assamensis</i>	-0.10			-0.45		

<i>S. ricini</i>	-0.02	-0.21
<i>A. pernyi</i>	-0.09	-0.13
<i>A. yamamai</i>	-0.07	-0.31
<i>B. mandarina</i>	-0.02	-0.04
<i>B. mori</i>	-0.06	-0.27
<i>M. sexta</i>	-0.06	-0.33

Nucleotide biasness was also reflected in the codon usage of PCGs that determines the frequency of synonymous codon usage (SCU) in various genomes at inter and intra levels. Based on the codon usage and relative SCU results, codons with A or T nucleotides at 3rd codon position were highly used in comparison with G or C nucleotides as discussed above. Among all codons, five most frequently used amino acids were Leu, Ile, Phe, Met and Asn (**Figure 3.13**). The most prevalent codons for the corresponding amino acids were UUA, AUU, UUU, AUA and AAU with the usage proportion values ranging from 45.53% to 48.95% (**Figure 3.14**). Cys was found to be the least used amino acid as detected in other selected Bombycoids. Biasness in SCU has been observed in many other lepidopterans to avoid errors due to mis-incorporation of amino acids which depends upon various factors like natural selection (gene length and function), mutation bias (base mutation and GC content), etc. (Kim *et al.*, 2012; Yang *et al.*, 2013). This A+T biasness may result in the change in amino acid composition similar to the other lepidopterans. Therefore, the codon usage analysis will have great significance in studying gene expression and evolutionary studies in silkworms.

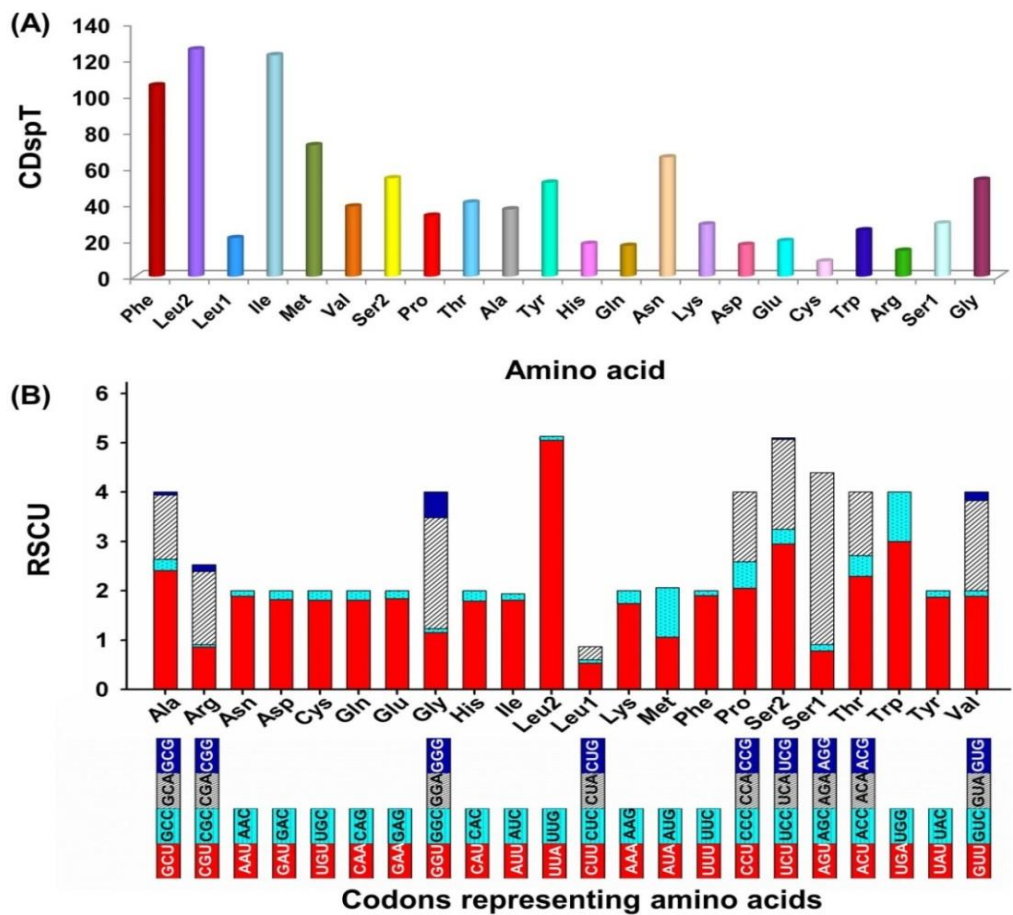


Figure 3.13 Codon usage in *A. assamensis* mitogenome. (A) Codon distribution where CDspT = codons per thousand codons (B) Relative synonymous codon usage (RSCU). Termination codons were excluded in the study.

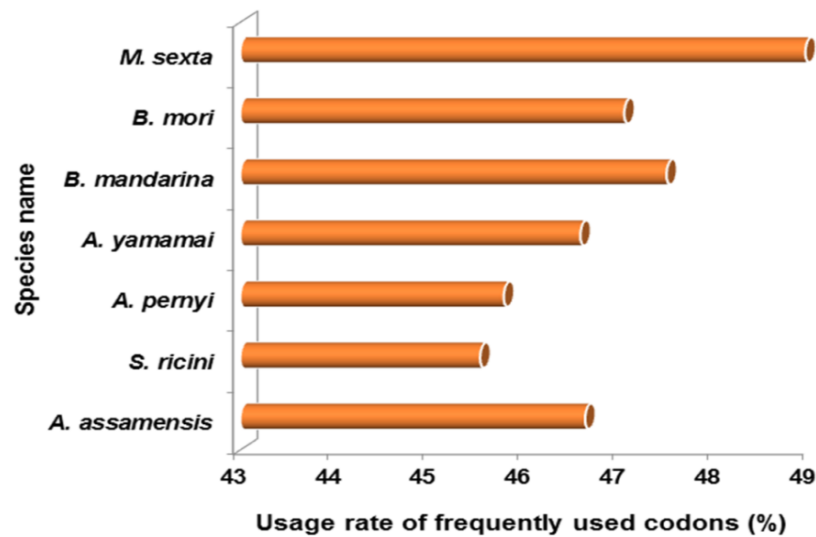


Figure 3.14 The usage rate of five frequently used codons in Bombycoidea members.

3.3.3 Comparative phylogenetic analysis

Phylogenetic trees constructed using both the datasets based on ML and BI methods showed that *A. assamensis* belongs to Saturniidae family which is clustered in the same branch with other families (Bombycidae+Sphingidae) of Bombycoidea superfamily within the order Lepidoptera (**Figure 3.15-3.18**).

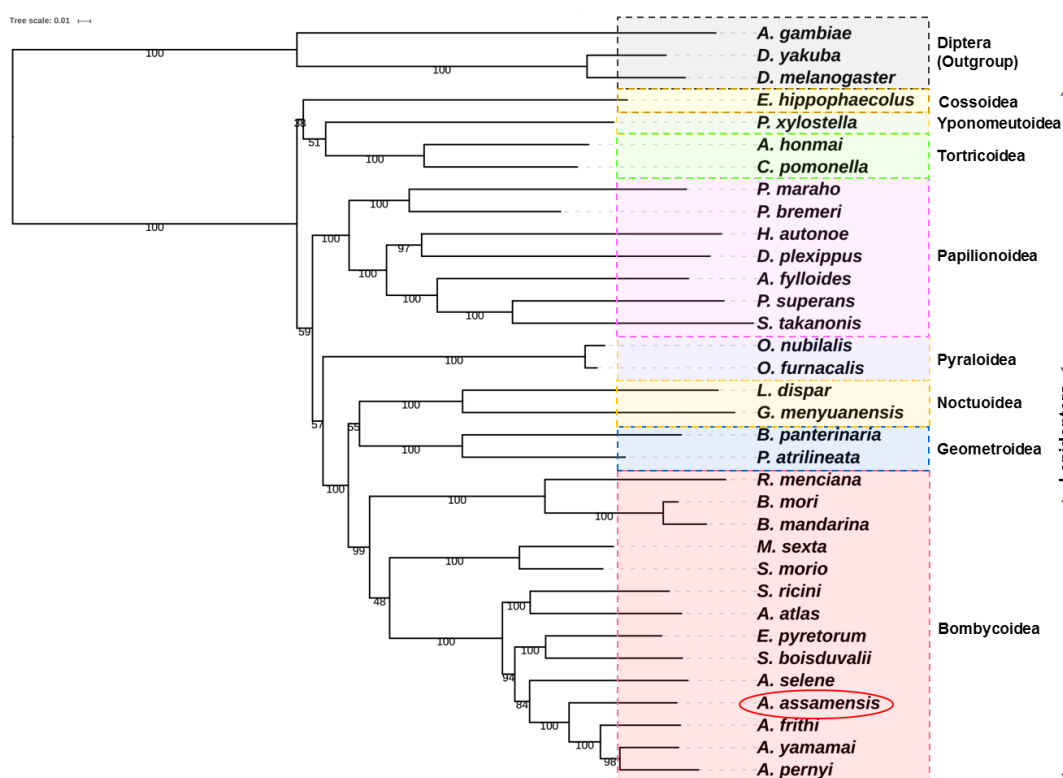


Figure 3.15 Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms based on maximum likelihood (ML) in RaxmlGUI v1.3 (1000 bootstrap replicates). The tree is drawn to scale, with bootstrap values indicated above the branches.

Within the Saturniidae family, *A. assamensis* was clustered around with other *Antheraea* species and was also close to *A. selene*. Further, our study supported the triplet superfamily (Bombycoidea+Geometroidea+Noctuoidea) relationship under Macroheterocera which along with Pyraloidea and Papilionoidea formed Obtectomera, a group within Apoditrysia (Timmermans et al., 2014). These results were consistent with the previously reported studies (Kim et al., 2009; Yang et al., 2013). More mitogenome sequence information on Bombycoidea superfamily is

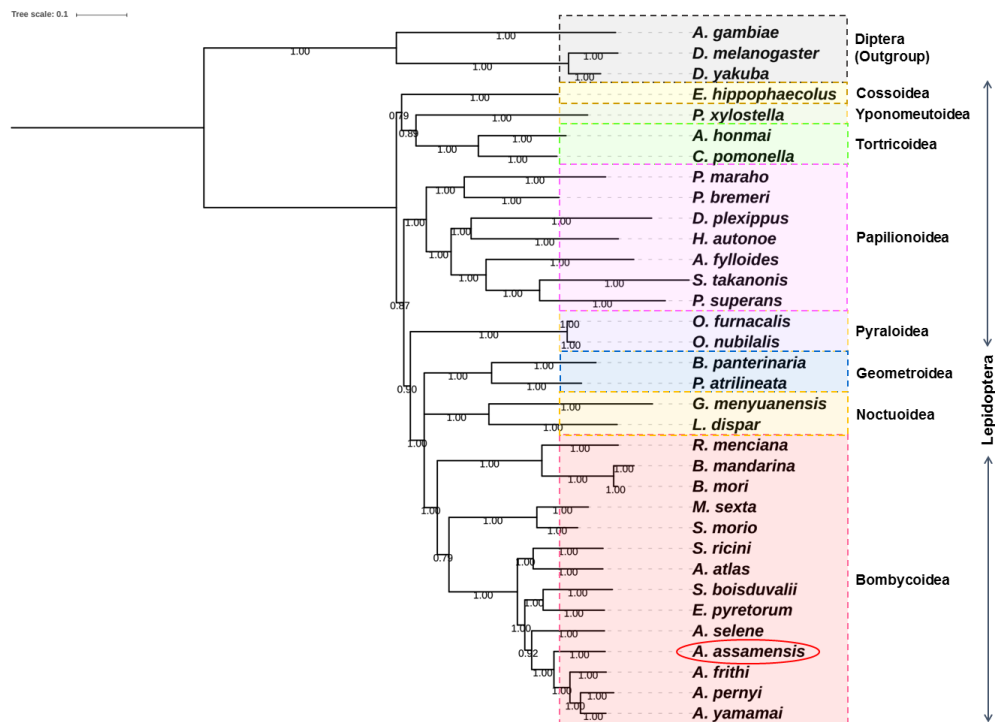


Figure 3.17 Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms based on bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

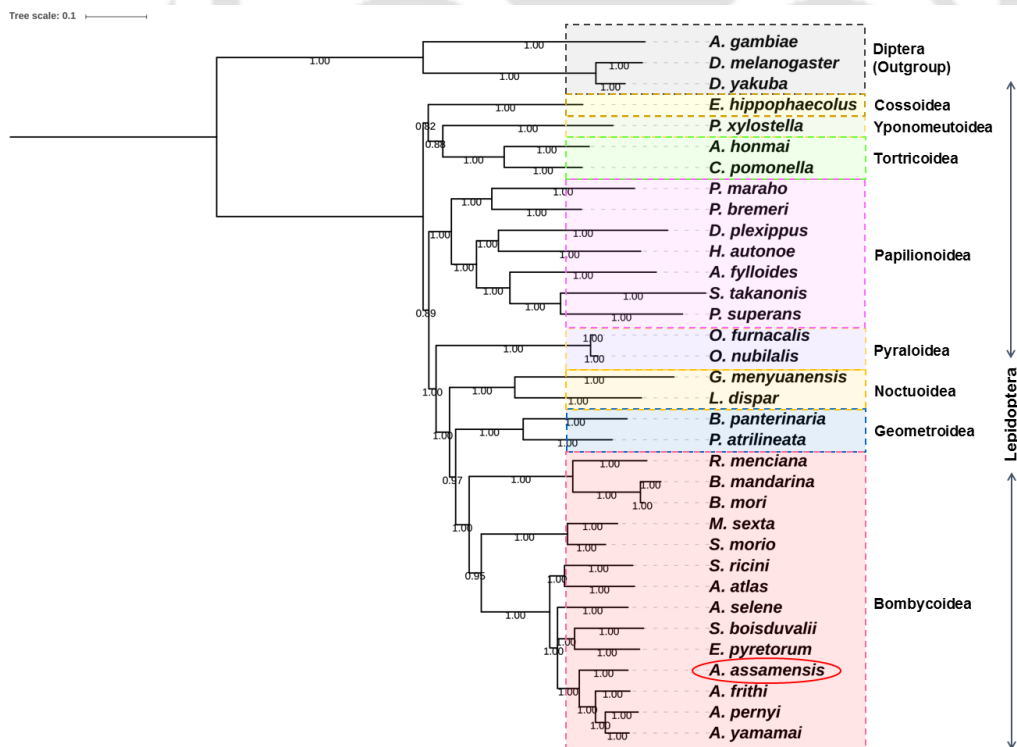


Figure 3.18 Phylogenetic tree inferred from 13PCGs+2rRNAs of 34 organisms based on bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

3.4 Conclusions

The present study deciphered the complete mitogenome of *A. assamensis* which shares similarity with majority of the lepidopterans, particularly Saturniids, in several characteristics such as genome organization and content, PCG size and structure, AT/GC skewness, tRNA structure and anti-codons, OS and IGS regions. Our analysis indicates that *A. assamensis* PCGs evolved under strong purifying selection and tRNAs genes showed high base substitution/mismatches. Transition-transversion ratio suggests that *nad4l* gene may be the significant contributor for evolution to form *A. assamensis* within *Antheraea* genus. A six-bp indel (deletion) and two highly conserved sequence blocks in control region suggest its potential application as a marker for delineating other closely related *Antheraea* sp. and subsequent usage in phylogenetic studies at lower taxonomic levels. The largest *tRNA^{Gln}-nad2* spacer found in the *A. assamensis* mitogenome may serve as another origin of replication. Our study assigns the taxonomic status of *A. assamensis* using an optimal model based phylogenetic tree construction. This is the 12th representative organism of Saturniidae family with a completely sequenced mitogenome. As *A. assamensis* is the sole producer of unique golden Muga silk which is the backbone of Assam's sericulture industry, our study on its mitogenomic landscape is an important addition to the existing genome informatics resources on silkworms.



Chapter 4

**Mitochondrial genome
sequencing of *Cricula*
trifenestrata, a wild silkworm
from Assam**

CHAPTER 4

Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkworm from Assam

4.1 Introduction

Cricula trifenestrata Helfer (Cricula silkworm), locally named as 'Amphutukoni muga', is one of the important wild sericigenous insect reported in Assam and other North-eastern regions of India. It is a multi- or bi-voltine silkworm that is also distributed in other countries like Srilanka, Bangladesh, Myanmar, Philippines and Indonesia (Tikader *et al.*, 2014; Saikia *et al.*, 2017). It is a polyphagous insect which attacks on many plants like mango, cashew, litchi, cinnamon, large cardomom and groundnut tress (Ahmad, 1994; Sujata & Anand, 2003; Pal & Medda, 2006). *Cricula* is also a serious pest of *A. assamensis* (muga) host plants (*P. bombycina*, *Litsea monopetala*, *Litsea citrata*, etc.) that migrates from one location to another depending on food availability (Negi & Sengupta, 1993; Ahmed *et al.*, 2012). It is therefore also known as a competitor of muga silkworm and is one of the main causes of damage to muga silk industry (Choudhury, 1981). Apart from being a pest, *Cricula* produces golden yellow colored fine silk which has potential as a natural silk fiber that can be exploited as biomaterial or silk textiles. Few studies have been performed to explore its potential in utilizing its silk for commercial importance and in the development of strategies to control the damage (Suriana *et al.*, 2011; Ahmed *et al.*, 2012; Nindhia *et al.*, 2014). However, no research has been done to identify or characterize the organism and to study evolutionary relationships with other lepidopteran insects.

As discussed in previous chapter, mitochondrial genomics or mitogenome characterization studies provide a useful resource for phylogenetic and comparative genomics studies among various species. Advanced sequencing technologies like next generation sequencing (NGS) has also led to the generation of plethora of information on the mitogenome of various insects of Lepidoptera order (Timmermans *et al.*, 2014). However, the utility of mitogenomics area in the family of wild silkmoths (Saturniidae) is still limited. Till now, only 12 species (including *Antheraea assamensis*) have been sequenced with complete mitogenome sequence and there exist many including *Cricula* which are still not much explored or unexplored. Therefore, exploration of these wild silkmoths including *Cricula* is thus essential for gaining deeper insights into the evolutionary relationships among wild and other lepidopteran insects.

In the present study, we utilized NGS technology to decode the complete mitogenome of *C. trifenestrata* and compared it with other sequenced lepidopteran insects. The comparison was based on various characteristics such as mitogenome arrangement, sequence homology, evolutionary rates of protein coding genes (PCGs), transfer RNA (tRNA) structures, ribosomal RNAs (rRNAs), location and number of intergenic spacers (IGS) and overlapping sequences (OS), presence of conserved regions, base composition, codon usage, etc. Finally, the phylogenetic trees were inferred using PCGs and PCGs+rRNAs datasets to elucidate the relationships among lepidopteran insects. This is the first report illustrating the complete mitogenome of wild silkmoth *C. trifenestrata* and will be helpful in future evolutionary studies among Saturniid and other lepidopteran insects.

4.2 Materials and methods

4.2.1 Sample processing, sequencing and assembly

The larvae of *Cricula trifenestrata* were collected from the Som (*Persea bombycina*) plantations of Central Muga Eri Research and Training Institute, Lahdoigarh, Jorhat, Assam, India (Sample ID: CMERI-Ct-001). The larvae were dissected and chopped under sterile conditions after washing with 70% ethanol to prevent contamination. The dissected tissue samples were processed for mitochondrial DNA sequencing and the remaining larvae were stored in 95% absolute ethanol at -80°C freezer for future use. Total DNA was isolated using standard CTAB (Cetyl trimethylammonium bromide) based buffer and silica column by the service provider (Genotypic Technology Pvt. Ltd. Bengaluru, India). The quantity and purity of extracted DNA was determined using NanoDrop Spectrophotometer (Thermofischer, USA) and Qubit Fluorometer (Thermofischer, USA). This was followed by mitochondrial DNA enrichment from total DNA, library preparation, sequencing and assembly as described in **Section 3.2.1 (Chapter 3)**. The complete overview of mitochondrial genome sequencing of *C. trifenestrata* and its analysis is similar to that of *A. assamensis* as represented in **Figure 3.1**.

4.2.2 Genome annotation, visualization and comparative analysis

The assembled scaffold of *C. trifenestrata* was annotated using MITOS web server (Bernt *et al.*, 2013) to determine the gene boundaries of PCGs, tRNAs and rRNAs as described in **Section 3.2.2 (Chapter 3)**. The number and location of overlapping or spacer regions between the genes were visualized and calculated manually. The control region or A+T rich region was validated by sequence comparison with the related organisms and searched for tandem repeats and microsatellites using Tandem Repeat Finder (Benson, 1999; Thompson *et al.*, 2002) and Microsatellite Finder (http://insilico.ehu.eus/mini_tools/microsatellites/)

tools, respectively. Circular Genome Viewer (CG View) tool was used to construct whole mito-map of the organism (Grant & Stothard, 2008). Finally, the SRA data and the annotation file of whole mitogenome sequence of *Cricula* prepared using NCBI Sequin tool (<http://www.ncbi.nlm.nih.gov/Sequin/>) were submitted to NCBI GenBank.

Further, a comparative analysis of mitogenome of *C. trifenestrata* (organism under study) was performed with other lepidopteran insects as listed in **Table 4.1** to study evolutionary relationships among them. The sequences of whole mitogenome, PCGs, tRNAs, rRNAs and control region of the selected organisms were retrieved from NCBI GenBank database. Various characteristics used for comparative analysis include mitogenome, PCGs, tRNAs, rRNAs, control region, IGS, OS, nucleotide composition, skewness and codon usage analysis. The detailed analysis methodology is discussed in **3.2.2.1 to 3.2.2.7 (Chapter 3)**.

4.2.3 Comparative phylogenetic analysis

For phylogenetic analysis, altogether 95 organisms representing 13 super-families and 32 families of the order Lepidoptera with complete mitogenome were selected. Seven organisms belonging to the order Trichoptera, Diptera and Hymenoptera were taken as outgroups as listed in **Table 4.1**. Phylogenetic relationships between *C. trifenestrata* and other organisms were elucidated using two datasets- (i) concatenated nucleotide sequences of 13 PCGs and (ii) concatenated nucleotide sequences of 13 PCGs and 2 rRNAs.

Table 4.1 List of the organisms with completely sequenced mitogenome belonging to the order Lepidoptera and the outgroup orders Trichoptera, Diptera and Hymenoptera. (Highlighted ones represent the selected organisms used for comparative mitogenome analysis with *C. trifenestrata*).

S.No.	Superfamily	Family	Species Name	Size	Accession No.	References
Lepidoptera						
1	Bombycoidea	Saturniidae	<i>C. trifenestrata</i>	15,425	KY644697	Present study
2	Bombycoidea	Saturniidae	<i>A. assamensis</i>	15,272	KU379695.1	Chapter 3
3	Bombycoidea	Saturniidae	<i>Actias selene</i>	15,236	NC_018133.1	Liu <i>et al.</i> , 2012
4	Bombycoidea	Saturniidae	<i>Eriogyna pyretorum</i>	15,327	NC_012727.1	Jiang <i>et al.</i> , 2009
5	Bombycoidea	Saturniidae	<i>Antheraea pernyi</i>	15,566	NC_004622.2	Liu <i>et al.</i> , 2008
6	Bombycoidea	Saturniidae	<i>Saturnia boisduvalii</i>	15,360	NC_010613.1	Hong <i>et al.</i> , 2008
7	Bombycoidea	Saturniidae	<i>Antheraea yamamai</i>	15,338	NC_012739.1	Kim <i>et al.</i> , 2009
8	Bombycoidea	Saturniidae	<i>Attacus atlas</i>	15,282	NC_021770.1	Chen <i>et al.</i> , 2014
9	Bombycoidea	Saturniidae	<i>Samia ricini</i>	15,384	NC_017869.1	Kim <i>et al.</i> , 2012
10	Bombycoidea	Saturniidae	<i>Antheraea frithi</i>	15,338	NC_027071.1	GenBank
11	Bombycoidea	Saturniidae	<i>Actias artemis</i>	15,243	KF927042.1	Park <i>et al.</i> , 2016
12	Bombycoidea	Endromidae	<i>Andraca theae</i>	15,737	NC_032694.1	Gu <i>et al.</i> , 2016
13	Bombycoidea	Bombycidae	<i>Bombyx huttoni</i>	15,638	NC_026518.1	Peng <i>et al.</i> , 2016
14	Bombycoidea	Bombycidae	<i>Bombyx mandarina</i>	15,928	NC_003395.1	Yukuhiro <i>et al.</i> , 2002
15	Bombycoidea	Bombycidae	<i>Bombyx mori</i>	15,643	NC_002355.1	GenBank
16	Bombycoidea	Bombycidae	<i>Rondotia menciana</i>	15,301	NC_021962.1	Kong & Yang, 2015
17	Bombycoidea	Lasiocampidae	<i>Dendrolimus punctatus</i>	15,411	NC_027156.1	Qin <i>et al.</i> , 2015
18	Bombycoidea	Lasiocampidae	<i>Dendrolimus spectabilis</i>	15,411	NC_025763.1	Tang <i>et al.</i> , 2014
19	Bombycoidea	Lasiocampidae	<i>Dendrolimus tabulaeformis</i>	15,411	NC_027157.1	Qin <i>et al.</i> , 2015
20	Bombycoidea	Lasiocampidae	<i>Euthrix laeta</i>	15,368	NC_031507.1	Wu <i>et al.</i> , 2016
21	Bombycoidea	Sphingidae	<i>Manduca sexta</i>	15,516	NC_010266.1	Cameron & Whiting, 2008
22	Bombycoidea	Sphingidae	<i>Sphinx morio</i>	15,299	NC_020780.1	Kim <i>et al.</i> , 2013
23	Tortricoidea	Tortricidae	<i>Adoxophyes honmai</i>	15,680	NC_008141.1	Lee <i>et al.</i> , 2006

24	Tortricoidea	Tortricidae	<i>Cydia pomonella</i>	15,253	NC_020003.2	Shi <i>et al.</i> , 2013
25	Tortricoidea	Tortricidae	<i>Spilonota lechriaspis</i>	15,368	NC_014294.1	Zhao <i>et al.</i> , 2011
26	Noctuoidea	Erebidae	<i>Amata formosae</i>	15,463	NC_021416.1	Lu <i>et al.</i> , 2013
27	Noctuoidea	Erebidae	<i>Hyphantria cunea</i>	15,481	NC_014058.1	Liao <i>et al.</i> , 2010
28	Noctuoidea	Noctuidae	<i>Agrotis ipsilon</i>	15,377	NC_022185.1	Wu <i>et al.</i> , 2015
29	Noctuoidea	Noctuidae	<i>Eutelia adaltricoides</i>	15,360	NC_026840.1	Yang <i>et al.</i> , 2015
30	Noctuoidea	Noctuidae	<i>Helicoverpa armigera</i>	15,347	NC_014668.1	Yin <i>et al.</i> , 2010
31	Noctuoidea	Noctuidae	<i>Mythimna separata</i>	15,332	NC_023118.1	Li <i>et al.</i> , 2015
32	Noctuoidea	Noctuidae	<i>Sesamia inferens</i>	15,413	NC_015835.1	Chai & Du, 2012
33	Noctuoidea	Noctuidae	<i>Spodoptera frugiperda</i>	15,365	NC_027836.1	Liu <i>et al.</i> , 2016
34	Noctuoidea	Lymantriidae	<i>Gynaephora menyuanensis</i>	15,770	NC_020342.1	Yuan & Zhang, 2013
35	Noctuoidea	Lymantriidae	<i>Lymantria dispar</i>	15,569	NC_012893.1	Yajun <i>et al.</i> , 2010
36	Noctuoidea	Notodontidae	<i>Ochrogaster lunifer</i>	15,593	NC_011128.1	Salvato <i>et al.</i> , 2008
37	Noctuoidea	Notodontidae	<i>Phalera flavesces</i>	15,659	NC_016067.1	Hao, 2012
38	Noctuoidea	Nolidae	<i>Risoba prominens</i>	15,343	NC_026841.1	Yang <i>et al.</i> , 2015
39	Hepialoidea	Hepialidae	<i>Ahamus yunnanensis</i>	15,816	NC_018095.1	Cao <i>et al.</i> , 2012
40	Hepialoidea	Hepialidae	<i>Thitarodes gonggaensis</i>	15,940	NC_026903.1	Shi <i>et al.</i> , 2016
41	Hepialoidea	Hepialidae	<i>Thitarodes renzhiensis</i>	16,173	NC_018094.1	Cao <i>et al.</i> , 2012
42	Copromorphoidea	Carposinidae	<i>Carposina sasakii</i>	15,611	NC_023212.1	Wu <i>et al.</i> , 2016
43	Cossoidea	Cossidae	<i>Eogystia hippophaecolus</i>	15,431	NC_023936.1	Gong <i>et al.</i> , 2014
44	Geometroidea	Geometridae	<i>Operophtera brumata</i>	15,748	NC_027723.1	Derks <i>et al.</i> , 2015
45	Geometroidea	Geometridae	<i>Phthonandria atrilineata</i>	15,499	NC_010522.1	Yang <i>et al.</i> , 2009
46	Geometroidea	Geometridae	<i>Jankowskia athleta</i>	15,534	NC_027948.1	Xu <i>et al.</i> , 2016
47	Geometroidea	Geometridae	<i>Biston panterinaria</i>	15,517	NC_020004.1	Yang <i>et al.</i> , 2013
48	Geometroidea	Geometridae	<i>Apocheima cinerarium</i>	15,722	NC_024824.1	Liu <i>et al.</i> , 2014
49	Yponomeutoidea	Lyonetiidae	<i>Leucoptera malifoliella</i>	15,646	NC_018547.1	Wu <i>et al.</i> , 2012
50	Yponomeutoidea	Yponomeutidae	<i>Prays oleae</i>	16,499	NC_025948.1	van Asch <i>et al.</i> , 2016
51	Yponomeutoidea	Plutellidae	<i>Plutella xylostella</i>	16,014	NC_025322.1	Dai <i>et al.</i> , 2016
52	Hesperioidea	Hesperiidae	<i>Potanthus flavus</i>	15,267	NC_024650.1	Kim <i>et al.</i> , 2014

Chapter 4: Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkmoth from Assam, India

53	Hesperioidea	Hesperiidae	<i>Lerema accius</i>	15,338	NC_029826.1	Cong & Grishin, 2016
54	Hesperioidea	Hesperiidae	<i>Lobocla bifasciatus</i>	15,366	NC_024649.1	Kim <i>et al.</i> , 2014
55	Hesperioidea	Hesperiidae	<i>Choaspes benjaminii</i>	15,300	NC_024647.1	Kim <i>et al.</i> , 2014
56	Hesperioidea	Hesperiidae	<i>Erynnis montanus</i>	15,530	NC_021427.1	Wang <i>et al.</i> , 2014
57	Hesperioidea	Hesperiidae	<i>Daimio tethys</i>	15,350	NC_024648.1	Kim <i>et al.</i> , 2014
58	Gelechioidea	Stathmopodidae	<i>Atrijuglans hetaohei</i>	15,379	NC_028168.1	Wang <i>et al.</i> , 2016
59	Gelechioidea	Gelechiidae	<i>Dichomeris ustalella</i>	15,410	NC_029810.1	Park <i>et al.</i> , 2016
60	Gelechioidea	Gelechiidae	<i>Mesophleps albilinella</i>	15,274	NC_029811.1	Park <i>et al.</i> , 2016
						Ramírez-Ríos <i>et al.</i> , 2016
61	Gelechioidea	Gelechiidae	<i>Tecia solanivora</i>	15,251	NC_029386.1	
62	Gelechioidea	Elachistidae	<i>Promalactis suzukiella</i>	15,507	NC_026697.1	Park <i>et al.</i> , 2016
63	Gelechioidea	Oecophoridae	<i>Stathmopoda auriferella</i>	15,456	NC_031831.1	Jeong <i>et al.</i> , 2016
64	Zygaenoidea	Limacodidae	<i>Monema flavescens</i>	15,396	NC_032683.1	Liu <i>et al.</i> , 2016
65	Zygaenoidea	Zygaenidae	<i>Rhodopsona rubiginosa</i>	15,248	NC_025761.1	Tang <i>et al.</i> , 2014
66	Papilionoidea	Riodinidae	<i>Abisara fylloides</i>	15,301	NC_021746.1	Zhao <i>et al.</i> , 2013
67	Papilionoidea	Riodinidae	<i>Apodemia mormo</i>	15,262	NC_024571.1	Kim & Kim, 2016
68	Papilionoidea	Nymphalidae	<i>Fabriciana nerippe</i>	15,140	NC_016419.1	Kim <i>et al.</i> , 2011
69	Papilionoidea	Nymphalidae	<i>Heliconius hecale</i>	15,338	NC_024744.1	Shen & Wang, 2016
						Servín-Garcidueñas & Martínez-Romero, 2014
70	Papilionoidea	Nymphalidae	<i>Danaus plexippus</i>	15,314	NC_021452.1	
71	Papilionoidea	Nymphalidae	<i>Hipparchia autonoe</i>	15,489	NC_014587.1	Kim <i>et al.</i> , 2010
72	Papilionoidea	Nymphalidae	<i>Sasakia funebris</i>	15,233	NC_022134.1	Wang <i>et al.</i> , 2013
73	Papilionoidea	Papilionidae	<i>Papilio maraho</i>	16,094	NC_014055.1	Wu <i>et al.</i> , 2010
74	Papilionoidea	Papilionidae	<i>Parnassius bremeri</i>	15,389	NC_014053.1	Kim <i>et al.</i> , 2009
75	Papilionoidea	Lycaenidae	<i>Protantigius superans</i>	15,248	NC_016016.1	Kim <i>et al.</i> , 2011
76	Papilionoidea	Lycaenidae	<i>Spindasis takanonis</i>	15,349	NC_016018.1	Kim <i>et al.</i> , 2011
77	Papilionoidea	Pieridae	<i>Catopsilia pomona</i>	15,142	NC_022687.1	Hao <i>et al.</i> , 2014
78	Papilionoidea	Pieridae	<i>Leptidea morsei</i>	15,122	NC_022686.1	Hao <i>et al.</i> , 2014
79	Papilionoidea	Pieridae	<i>Pieris melete</i>	15,140	NC_010568.1	Hong <i>et al.</i> , 2009
80	Pyraloidea	Crambidae	<i>Ostrinia furnacalis</i>	14,536	NC_003368.1	Coates <i>et al.</i> , 2005

81	Pyraloidea	Crambidae	<i>Paracymoriza distinctalis</i>	15,354	NC_023471.1	Ye & You, 2016
82	Pyraloidea	Crambidae	<i>Scirpophaga incertulas</i>	15,220	NC_031329.1	GenBank
83	Pyraloidea	Crambidae	<i>Maruca vitrata</i>	15,385	NC_024099.1	GenBank
84	Pyraloidea	Crambidae	<i>Chilo auricilius</i>	15,367	NC_024644.1	Cao & Du, 2014
85	Pyraloidea	Pyralidae	<i>Plodia interpunctella</i>	15,264	NC_027961.1	Tang <i>et al.</i> , 2015
86	Pyraloidea	Pyralidae	<i>Hypsopygia regina</i>	15,212	NC_030508.1	GenBank
87	Pyraloidea	Pyralidae	<i>Lista haraldusalis</i>	15,213	NC_024535.1	Ye <i>et al.</i> , 2015
88	Pyraloidea	Pyralidae	<i>Corcyra cephalonica</i>	15,273	NC_016866.1	Wu <i>et al.</i> , 2012
Outgroup Organisms						
89	Hemiptera:	Apidae	<i>Apis mellifera</i>	16,343	NC_001566.1	Crozier & Crozier, 1993
90	Diptera	Culicidae	<i>Anopheles gambiae</i>	15,363	NC_002084.1	Beard <i>et al.</i> , 1993
91	Diptera	Drosophilidae	<i>Drosophila incompta</i>	15,640	NC_025936.1	De Ré <i>et al.</i> , 2014
92	Diptera	Drosophilidae	<i>Drosophila melanogaster</i>	19,524	NC_024511.2	GenBank
93	Diptera	Tachinidae	<i>Blepharipa</i> sp.	15,080		Unpublished
94	Trichoptera	Limnephilidae	<i>Limnephilus decipiens</i>	16,667	NC_026219.1	GenBank
95	Trichoptera	Phryganeidae	<i>Eubasilissa regina</i>	15,021	NC_023374.1	Wang <i>et al.</i> , 2013

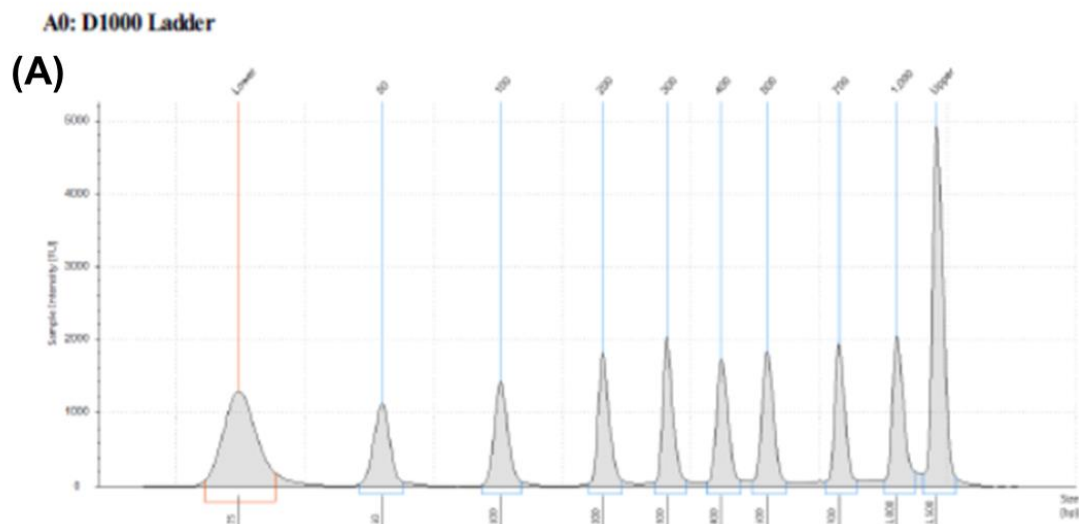
The alignment of the nucleotide sequences of PCGs was done using MAFFT algorithm and GBlocks option in the TranslatorX server (Abascal *et al.*, 2010). rRNA sequences were aligned using Q-INS-I method in MAFFT tool (Kato & Standley, 2013) followed by removal of ambiguous positions and divergent regions using GBlocks server (http://molevol.cmima.csic.es/castresana/Gblocks_server.html). The rRNA sequence blocks obtained were then concatenated with nucleotide sequences of PCGs. Optimal model of nucleotide substitution was selected for the given dataset using jModeltest 2.1.7 which was GTR+G+I (Guindon & Gascuel, 2003; Darriba *et al.*, 2012). Further, phylogenetic trees were constructed based on Maximum Likelihood (ML) and Bayesian Inference (BI) methods using RaxmlGUI v1.3 and MrBayes v3.2.6, respectively (**Chapter-3, Section 3.2.3**) (Silvestro & Michalak, 2012; Ronquist *et al.*, 2012). The visualization of tree topologies was carried out using iTOL v3 (Interactive tree of life) tool (Letunic & Bork, 2006).

4.3 Results and discussion

4.3.1 Sample processing, sequencing and assembly

Total DNA isolated from the tissue sample of *C. trifenestrata* was found to be optimal with a concentration of 1176.8 ng/μl and 315 ng/μl by NanoDrop spectrophotometer and Qubit fluorometer, respectively. The mitochondrial DNA was then enriched from total DNA prior to library preparation. The library profiles showed around 300-1000 bp total size of the fragments with an effective insert size of 180-880 bp and adapters of size ~120bp (**Figure 4.1**). The concentration of the prepared libraries (15.9 ng/μl) was also found to be suitable for sequencing on Illumina platform (Illumina NextSeq 500 sequencer) using 2 × 150 bp paired

end chemistry. A total of 49,57,940 number of raw reads were obtained out of which around 47,30,802 number of high quality reads were processed for scaffold generation through *de novo* assembly using SPAdes-3.5.0, SSPACE and CAP3. A scaffold of 15,425 bp length was generated that represents the whole mitogenome sequence of *C. trifenestrata*.



Sample Table

Well	Conc. [ng/ul]	Sample Description	Alert	Observations
A0	20.3	D1000 Ladder		Ladder

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [nmol/l]	% Integrated Area	Peak Comment	Observations
25	5.22	-	321	-		Lower Marker
50	2.25	-	69.3	11.11		
100	2.37	-	36.5	11.71		
200	2.47	-	19.0	12.20		
300	2.55	-	13.1	12.56		
400	2.57	-	9.87	12.66		
500	2.71	-	8.33	13.36		
700	2.46	-	5.41	12.15		
1,000	2.89	-	4.44	14.25		
1,500	6.50	6.50	6.67	-		Upper Marker

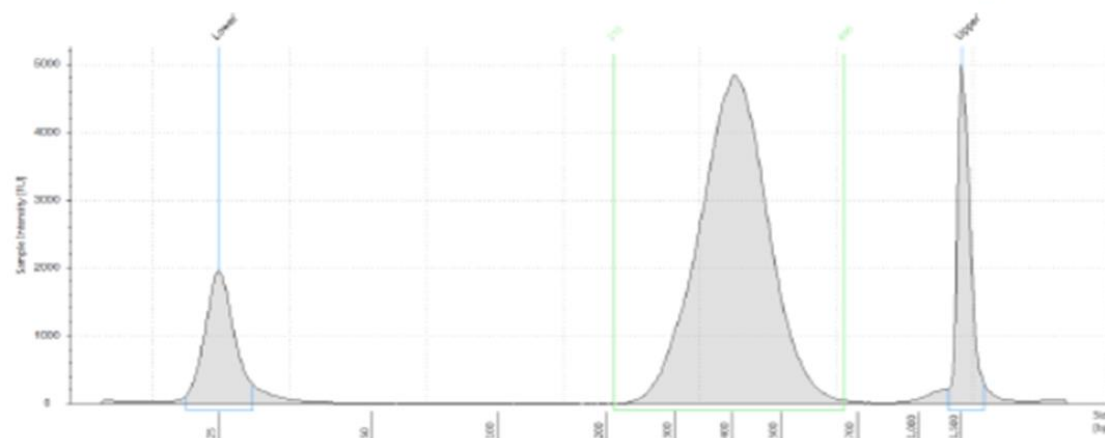
Sample Table

(B)

Well	Conc. [ng/ul]	Sample Description	Alert	Observations
E2	30.9	SO_4478_Set3_Sample_3_CT1_ePCR1_IL_WGS		

Peak Table

Size [bp]	Calibrated Conc. [ng/ul]	Assigned Conc. [ng/ul]	Peak Molarity [nmol/l]	% Integrated Area	Peak Comment	Observations
25	5.84	-	359	-		Lower Marker
404	30.9	-	117	100.00		
1,500	6.50	6.50	6.67	-		Upper Marker



Region Table

From [bp]	To [bp]	Average Size [bp]	Conc. [ng/ul]	Region Molarity [nmol/l]	% of Total	Region Comment	Color
210	666	409	34.1	132	95.37		Green

Figure 4.1 Tapestation profiles of the ladder (A) and prepared libraries (B).

4.3.2 Genome annotation, visualization and comparative analysis

The mitogenome of *C. trifenestrata* is a closed, circular molecular of 15,425 bp length comprising of 37 genes including 13 PCGs, 22 tRNAs and 2 rRNAs with an additional non-coding control region (A+T rich region). About 24 genes are seated on major coding strand or J-strand (+) while the remaining 13 genes on minor coding strand or N-strand (-). The features of *Cricula* mitogenome like gene location, arrangement, direction, etc. along with the mito-map are depicted in the **Figure 4.2**. Both sequence (full annotated mitogenome) and SRA data of *C. trifenestrata* have been deposited to NCBI GenBank under the accession numbers KY644697 and SRR3990715, respectively.

4.3.2.1 Comparative mitogenome analysis

The size of *C. trifenestrata* mitogenome was smaller than that of *M. sexta* (15, 516 bp), *A. pernyi* (15, 566 bp), *B. mori* (15, 643 bp), *A. theae* (15, 737 bp) and *B. mandarina* (15, 928 bp) while bigger than that of the other selected Bombycoid insects (**Figure 4.3**). However, it lies within the range of other sequenced lepidopteran insects (15,122 bp in *Leptidea morsei* to 16,499 bp in *Prays oleae*) (Hao *et al.*, 2014; van Asch *et al.*, 2016). The gene content (i.e. 37 genes and a CR) of mitogenome of *C. trifenestrata* was also typical of other lepidopteran insects (Lee *et al.*, 2006; Hong *et al.*, 2008; Chen *et al.*, 2014; Liu *et al.*, 2016). We found that mitogenome arrangement i.e. gene location and presence of conserved gene clusters was conserved in *C. trifenestrata* (**Figure 4.2A**).

It showed typical $tRNA^{Met}-tRNA^{Ile}-tRNA^{Gln}$ (or MIQ) gene arrangement similar to that found in *A. assamensis* mitogenome and in other known lepidopterans of Ditrysia lineage till date. However, this arrangement differed from ancestral type ($tRNA^{Ile}-tRNA^{Gln}-tRNA^{Met}$ or IQM) present in the members of family Hepialidae of Aditrysian lineage of Lepidoptera. This suggests that ancestral IQM cluster had undergone rearrangement after splitting of Hepialoidea from other lepidopteran lineages (Cao *et al.*, 2012; Shi *et al.*, 2016). Apart from MIQ $\leftrightarrow$ IQM cluster, gene rearrangements have also been observed in *tRNA* (*ARNS1EF*) gene cluster in few lepidopteran insects (Timmermans *et al.*, 2014; Wang *et al.*, 2014; Park *et al.*, 2016). *ARNS1EF* is a typical arrangement observed in *Cricula* and most of the other lepidopteran insects sequenced till now (Jiang *et al.*, 2009; Liu *et al.*, 2016).

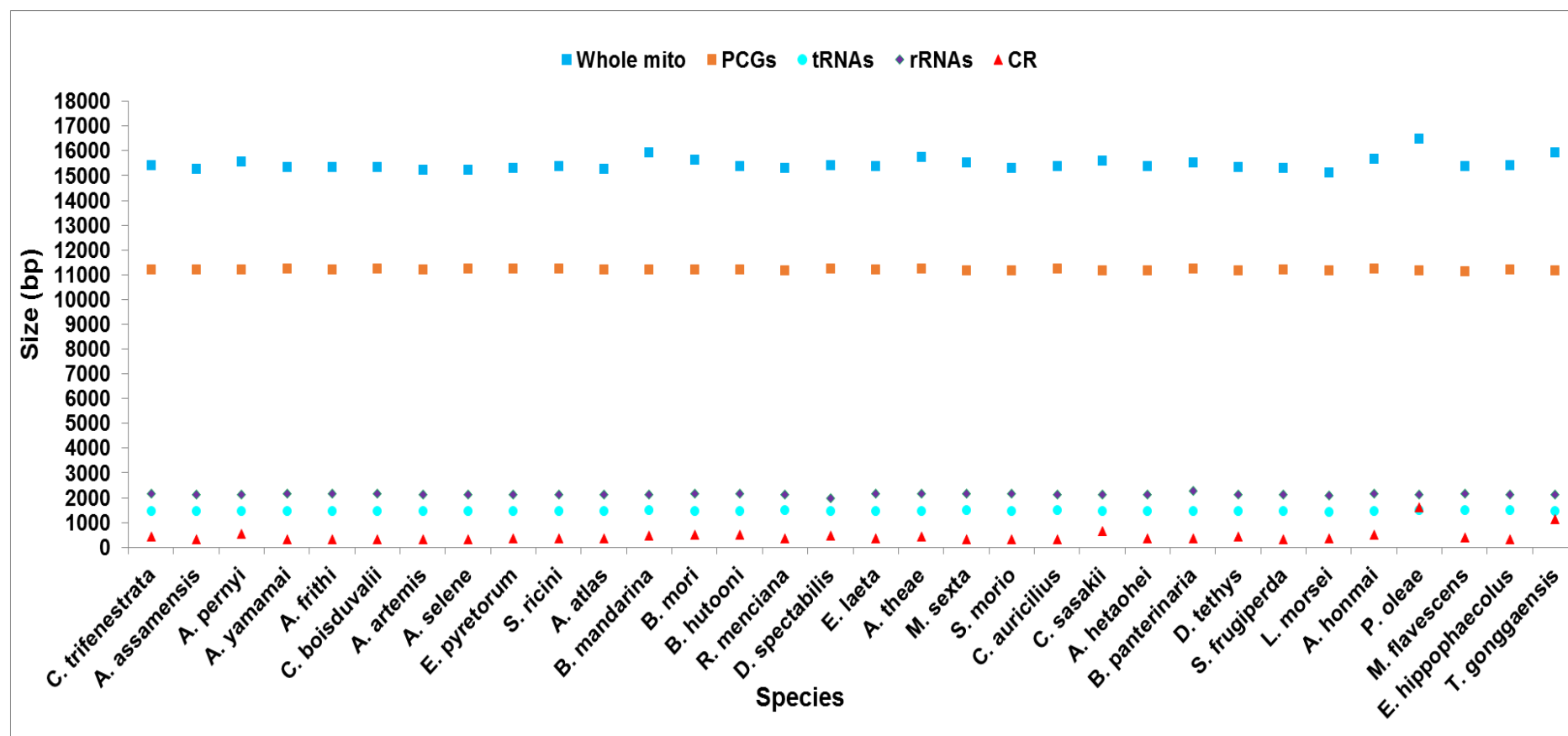


Figure 4.3 Size variation in different regions of mitogenome in the selected species (Whole mito: Whole mitogenome, PCGs: Protein coding genes, tRNAs: Transfer RNAs, rRNAs: Ribosomal RNAs; CR: Control region).

The sequence similarity of *C. trifenestrata* mitogenome with other selected silkmoths also revealed maximum sequence similarity with Saturniid species (*Antheraea* species, *E. pyretorum*, *C. boisduvalii*, etc.) followed by Sphingiids, Bombycids, etc. and lowest with *T. gonggaensis* (77.09%). This suggests that the level of sequence conservation decreases from lower taxonomic groups to higher taxonomic groups.

4.3.2.2 Comparative analysis among protein coding genes (PCGs)

As other related silkmoths, the mitogenome of *C. trifenestrata* also comprised of 13 PCGs of total length of 11,219 bps (~72.72% of the total mitogenome) spread over J or major strand and N or minor strand (**Figure 4.3**). The total size of PCGs was higher than that of *A. assamensis* (11, 208 bp) while lying within the range of other selected lepidopteran insects (11, 151 bp in *M. flavescens* to 11, 253 bp in *D. spectabilis*) (Tang *et al.*, 2014; Liu *et al.*, 2016). Nine genes (*nad2*, *cox1*, *cox2*, *atp8*, *atp6*, *cox3*, *nad3*, *nad6* and *cytb*) were encoded by J strand while the remaining four genes (*nad5*, *nad4*, *nad4l* and *nad1*) by N strand (**Figure 4.2B**).

All PCGs initiated with typical ATN codons as observed in other silkmoths except *cox1* gene which commenced with CGA as start codon (**Table 4.2**). The occurrence of such non-canonical start codons for *cox1* gene has been described in various family members of the order Lepidoptera (Jiang *et al.*, 2009; Gong *et al.*, 2014; Park *et al.*, 2016; Peng *et al.*, 2016; Wu *et al.*, 2016; Liu *et al.*, 2016). For instance, CGA was found to be a start codon in *A. assamensis*, *E. pyretorum*, *A. frithi* and *A. artemis* (Saturniidae), *B. huttoni* and *B. mori* (Bombycidae), *C. sasakii* (Copromorphoidae), *E. hippophaecolus* (Cossoidae), *M. flavescens* (Zygaenoidea), etc. as given in **Table 4.2**. Other atypical start codons for *cox1* gene also exist in Bombycoidea members like *S. boisduvalii* (TTG), *A. pernyi*, *A.*

yamamai, *A. selene* and *B. mandarina* (TTAG), etc. and lepidopterans like *O. furncalis*, *O. nubilalis* (a six base TATTAG) (Coates *et al.*, 2005; Hong *et al.*, 2008; Liu *et al.*, 2008; Kim *et al.*, 2009).

Similarly, a typical TAA termination codon pattern occurred in all the PCGs of *C. trifenestrata* mitogenome, except in *cox2* and *nad5* which utilized truncated termination codon (T-). Such non-canonical termination codons are common evolutionary characteristics of many lepidopterans as listed in **Table 4.3**. These truncated codons produce functional stop codons (TAA) through the mechanism of polycistronic transcriptional cleavage and polyadenylation (Cameron & Whiting, 2008; Liu *et al.*, 2008; Kim *et al.*, 2009).

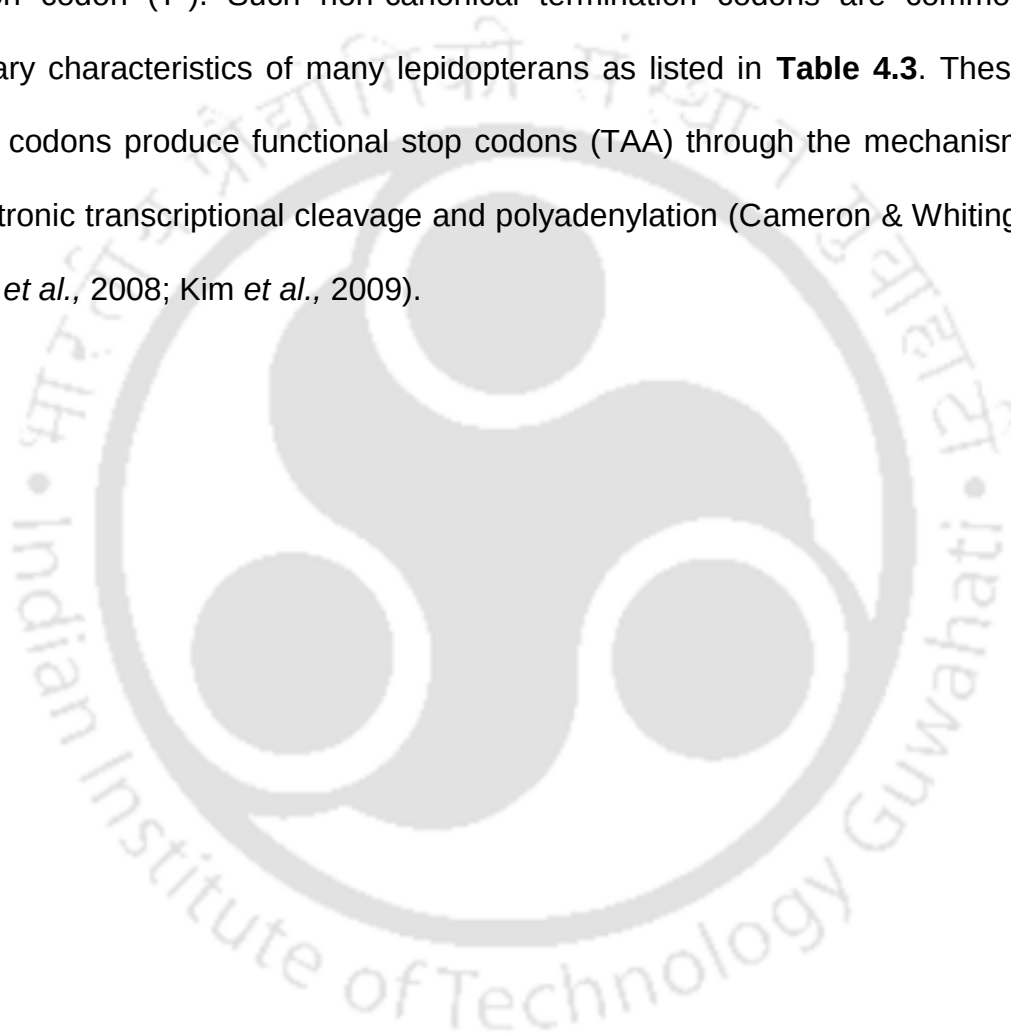


Table 4.2 Initiation codons of 13 protein coding genes in the mitogenome of *C. trifenestrata* and selected lepidopteran insects.

Species Name	<i>nad2</i>	<i>cox1</i>	<i>cox2</i>	<i>atp8</i>	<i>atp6</i>	<i>cox3</i>	<i>nad3</i>	<i>nad5</i>	<i>nad4</i>	<i>nad4l</i>	<i>nad6</i>	<i>cytb</i>	<i>nad1</i>
<i>C. trifenestrata</i>	ATA	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATC	ATG	ATG
<i>A. assamensis</i>	ATT	CGA	GTG	ATT	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATA	ATG
<i>A. pernyi</i>	ATT	TTAG	ATG	ATT	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATA	ATG
<i>S. ricini</i>	ATT	ATT	GTG	ATC	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATG	ATG
<i>A. yamamai</i>	ATT	TTAG	ATG	ATT	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATA	ATG
<i>A. selene</i>	ATT	TTAG	ATG	ATC	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATG	ATG
<i>C. boisduvalii</i>	ATT	TTG	GTG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATG	ATG
<i>E. pyretorum</i>	ATT	CGA	GTG	ATT	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATG	ATG
<i>A. frithi</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATA	ATG
<i>A. artemis</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATG	ATG
<i>A. atlas</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATA	ATG
<i>B. mandarina</i>	ATA	TTAG	ATG	ATA	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATA	ATA
<i>B. mori</i>	ATA	CGA	ATG	ATA	ATG	GAA	ATA	ATT	ATG	ATG	ATT	ATA	ATT
<i>B. huttoni</i>	ATA	CGA	ATG	ATA	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATA	ATA
<i>R. menciaana</i>	ATT	ATT	ATG	ATT	ATG	ATG	ATT	ATA	ATG	ATG	ATA	ATA	ATG
<i>D. spectabilis</i>	ATT	CGA	ATA	ATT	ATG	ATG	ATG	ATT	ATG	ATG	ATA	ATG	ATG
<i>E. laeta</i>	ATA	CGA	ATA	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATG	ATG
<i>A. theae</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATC	ATT	ATG	ATG	ATA	ATG	ATT
<i>M. sexta</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATG	ATG	ATG
<i>S. mori</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATG	ATG	GTG
<i>T. gonggaensis</i>	ATT	CGA	ATG	ATG	ATA	ATG	ATT	ATT	ATG	ATG	ATA	ATG	TTG
<i>C. auricilius</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATG	ATG
<i>C. sasakii</i>	ATT	CGA	ATC	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATT	ATG	ATG
<i>A. hetaohei</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATA	ATT	ATG	ATG	ATA	ATG	ATG
<i>B. panterinaria</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATT	ATT	ATG	ATG	ATT	ATG	ATA
<i>D. tethys</i>	ATT	CGA	ATT	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATT	ATG	ATG

<i>S. frugiperda</i>	ATT	CGA	ATG	ATG	ATG	ATG	ATC	ATT	ATG	ATG	ATT	ATG	ATG
<i>L. morsei</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATA	ATT	ATA	ATG	ATT	ATG	ATG
<i>A. honmai</i>	ATA	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATG	ATA	ATA	ATG
<i>P. oleae</i>	ATC	CGA	ATG	ATT	ATG	ATG	ATG	ATT	ATG	ATG	ATT	ATG	ATG
<i>M. flavescens</i>	ATT	CGA	ATG	ATT	ATG	ATG	ATT	ATC	ATG	ATA	ATA	ATG	ATG
<i>E. hippophaecolus</i>	ATT	CGA	ATG	ATC	ATG	ATG	ATT	ATT	ATG	ATA	ATT	ATG	ATG

Table 4.3 Termination codons in the protein coding genes of *C. trifenestrata* mitogenome and other selected lepidopteran insects.

Species Name	<i>nad2</i>	<i>cox1</i>	<i>cox2</i>	<i>atp8</i>	<i>atp6</i>	<i>cox3</i>	<i>nad3</i>	<i>nad5</i>	<i>nad4</i>	<i>nad4l</i>	<i>nad6</i>	<i>cytb</i>	<i>nad1</i>
<i>C. trifenestrata</i>	TAA	TAA	T-	TAA	TAA	TAA	TAG	T-	TAA	TAG	TAA	TAA	TAA
<i>A. assamensis</i>	TAA	T-	T-	TAA	TAA	TAA	TAA	T-	TAA	TAG	TAA	TAA	TAA
<i>A. pernyi</i>	TAA	T-tRNA	T-tRNA	TAA	TA-COIII	TAA	T-tRNA	T-tRNA	TAA	TAA	TAA	TAA	TAA
<i>S. ricini</i>	TAA	T-tRNA	T-tRNA	TAA	TAA	TAA	TAG	TAA	TAA	TAA	TAA	TAA	TAA
<i>A. yamamai</i>	TAA	TAA	T-tRNA	TAA	TAA	TAA	TAG	T-tRNA	TAG	TAA	TAA	TAA	TAA
<i>A. selene</i>	TAA	T	T	TAA	TAA	TAA	TAG	T	TAA	TAA	TAA	TAA	TAA
<i>C. boisduvalii</i>	TAA	TAA	T-tRNA	TAA	TA-COIII	TAA	T-tRNA	TAA	TAA	TA-ND4	TAA	TAA	TAG
<i>E. pyretorum</i>	TAA	T	T	TAA	TAA	TAA	TAG	TAA	TAA	TAA	TAA	TAA	TAA
<i>A. frithi</i>	TAA	T	T	TAA	TAA	TAA	TAG	T	TAG	TAA	TAA	TAA	TAA
<i>A. artemis</i>	TAA	T	T	TAA	TAA	TAA	TAG	T	TAA	TAA	TAA	TAA	TAA
<i>A. atlas</i>	TAG	T-tRNA	T-tRNA	TAA	TA-	TAA	T-	T-	TAA	TA-	TAA	T-	TAA

				COII			tRNA	tRNA		ND4	tRNA		
<i>B. mandarina</i>	TAA	T++	T++	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA
<i>B. mori</i>	TAA	T-	T-	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA
<i>B. huttoni</i>	TAA	T(aa)	T(aa)	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAG
<i>R. menciaana</i>	TAA	TAA	T	TAA	TAA	TAA	TAG	TAA	TAA	TAA	TAA	TAA	TAG
<i>D. spectabilis</i>	TAA	T-	T-	TAA	TAA	TAA	TAA	TAA	T-	TAA	TAA	TAA	TAA
<i>E. laeta</i>	TAA	T	T	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA
<i>A. theae</i>	TAA	T	T	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA
<i>M. sexta</i>	T-	T-	T-	TAA	TAA	TAA	TAA	T	TAA	TAA	TAA	TAA	TAA
<i>S. mori</i>	T-	T-	T-	TAA	TAA	TAA	TAA	T-	TAA	TAA	TAA	TAA	TAA
<i>T. gonggaensis</i>	T	T	T	TAA	TAA	TA	T	T	TAA	TAA	TAA	TAA	TAA
<i>C. auricilius</i>	TAA	T	T	TAA	TAA	TAA	TAA	TAA	TA	TAA	TAA	TAA	TAA
<i>C. sasakii</i>	T	T	T	TAA	TAA	TAA	TAA	T	TAA	TAA	TAA	T	TAA
<i>A. hetaohei</i>	TAA	TAA	T	TAA	TA	TAA	TAA	TAA	TA	TAA	TAA	TAA	TAA
<i>B. panterinaria</i>	TAA	TAA	T	TAA	TAA	TAA	TAG	TAA	T	TAA	TAA	TAA	TAA
<i>D. tethys</i>	TAA	T	T	TAA	TAA	TAA	TAA	T	TAA	TAG	TAA	TAA	TAG
<i>S. frugiperda</i>	T	T	T	TAA	TA	TAA	T	TAA	T	TAG	TAA	TA	TAA
<i>L. morsei</i>	TAA	T-tRNA	T-tRNA	TAA	TAA	TAA	TAA	T-tRNA	T-tRNA	TAA	TAA	TAA	TAA
<i>A. honmai</i>	TAA	T	T	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA
<i>P. oleae</i>	TAA	T	T	TAA	TAA	TAA	TAA	TA	TAA	TAA	T	TAA	TAA
<i>M. flavescens</i>	T	T	T	TAA	TAA	TAA	TAA	TAA	T	A	TAA	T	TAA
<i>E. hippophaecolus</i>	T	T	T	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TAA	TA

The PCGs of *Cricula* mitogenome were compared with other organisms of same family (Saturniide) as well as organisms from other hierarchical levels in terms of conserved (C), variable (V) and phylogenetically informative (PI) sites (**Table 4.4**). The comparison was performed at amino acid level as it is a functional entity of an organism. We found that the conserved sites decreased with distant organisms across all PCGs. However, all the PCGs of *Cricula* showed higher conserved sites with Sphingidae family members rather than Saturniide family. This is an interesting character of this organism, which indirectly depicts that mitogenome of *Cricula* might have originated from organisms of Sphingidae family. This might be attributed to sharing of a common habitat during evolution of *Cricula* which opens the scope for further evolutionary studies on this organism. Among all 13 PCGs, highest conservation was observed for *cox1* (68-92%) and *cox2* (59-91%) genes of *Cricula* across other organisms. These are mostly employed in molecular systematic studies in Lepidoptera and concordant with those obtained for *A. assamensis* as explained in **Chapter 3**. The *atp8* gene exhibited very less conservation and highest variability (39% to 70%) with respect to other genes in all the families. This suggest than *atp8* gene has higher possibility of mutation during evolution and generation of new strains. Similarly, higher percentage of phylogenetically informative (PI) sites was observed in *atp8* gene of *Cricula* in comparison with the other genes across all organisms. A study on *M. sexta* (Sphingidae family) has also been reported to show higher variability in *atp8* gene in comparison with other genes (Cameron & Whiting, 2008). Hence, it can be useful as informative marker for phylogenetic studies.

Further, we determined the evolutionary rates of PCGs of *C. trifenestrata* with other organisms belonging to various hierarchical levels of taxonomy (**Table 4.5**).

All the PCGs of *Cricula* were found to display Ks/Ks ratio of less than 1 with respect to the selected organisms. This indicates the evolution of all PCGs under strong purifying selection. Overall, *cox1* gene showed lowest Ka/Ks values than other genes suggesting that the mutations occurring in this gene are synonymous which maintains its functions for protein synthesis. *M. sexta* has also reported to exhibit lowest Ka/Ks ratio for *cox1* gene in comparison with other PCGs (Cameron and Whiting, 2008). On other hand, the *atp8* gene of *Cricula* showed higher Ka/Ks ratio with respect to the ancestor *T. gonggaensis*. This gene encodes ATPase subunit 8 which is required for ATP synthesis thus suggesting that high energy was required by *C. trifenestrata* as compared to the ancestor moth in order to perform activities like silk production, adaptation to wild conditions, etc. *Cricula* exhibited lowest Ka/Ks ratio with other members of Saturniidae family and it was found to be increasing with the distant organisms. The *cox1* gene and *atp8* gene can be the potential markers at various hierarchical levels studies as they have lowest and highest Ka/Ks ratios, respectively.

4.3.2.3 Comparative analysis among transfer RNAs (tRNAs)

Similar to *A. assamensis* and other related silkmoths, *C. trifenestrata* also comprised of 22 tRNA genes, 14 being encoded by J strand and 8 genes by N strand. These tRNA genes varied in length from 62-71 bps which lies within the range of other lepidopteran insects (Kim *et al.*, 2012; Yang *et al.*, 2013; Chen *et al.*, 2014; Hao *et al.*, 2014). The total length (1453 bp) was higher than *L. morsei* (1436 bp) but lower than other selected insects (**Figure 4.3**).

Table 4.4 Gene-by-gene divergence in protein coding genes of *C. trifenestrata* with respect to other selected organisms at amino acid level. Conserved sites (C), variable sites (V) and phylogenetically informative sites (PI) represented in percentage of total gene length.

Sites (%)	*Hierarchy level	<i>cox1</i>	<i>cox2</i>	<i>cox3</i>	<i>atp8</i>	<i>atp6</i>	<i>cytb</i>	<i>nad1</i>	<i>nad2</i>	<i>nad3</i>	<i>nad4</i>	<i>nad4l</i>	<i>nad5</i>	<i>nad6</i>
C-sites	Satuniidae	91.1	85.1	81.4	36.1	81.1	78.4	68.6	67.4	70.3	69.6	64.4	71.1	51.0
	Bombycidae	86.9	85.1	73.1	41.0	72.7	74.9	61.8	61.6	62.7	63.2	56.4	64.9	40.9
	Sphingidae	92.2	91.2	86.7	49.2	81.5	83.0	74.5	73.3	74.6	71.2	71.3	74.8	56.6
	Bombycoidea	83.1	77.2	67.4	29.5	67.0	68.4	52.6	51.7	60.2	53.9	49.5	54.7	35.4
	Macroheterocera	76.3	68.4	56.4	21.3	60.4	62.9	44.9	38.1	49.2	44.8	40.6	46.1	23.2
	Ditrysia	70.4	62.7	51.9	21.3	47.1	53.4	40.6	31.1	44.1	36.8	25.7	38.1	19.2
	Lepidoptera	68.7	59.6	51.1	19.7	46.7	51.4	38.5	27.9	42.4	35.1	23.8	36.6	16.7
V-sites	Satuniidae	5.9	14.5	17.8	55.7	18.1	17.8	27.4	30.5	29.7	24.7	30.7	27.3	38.9
	Bombycidae	9.9	14.5	26.1	52.5	26.4	21.6	34.2	37.5	36.4	31.1	38.6	32.4	48.5
	Sphingidae	4.6	8.3	12.9	39.3	17.6	13.3	21.2	24.7	24.6	22.4	23.8	21.5	32.3
	Bombycoidea	14.4	22.4	32.2	65.6	32.2	28.3	43.4	47.4	39.8	40.6	45.5	43.8	55.6
	Macroheterocera	21.3	31.1	43.2	73.8	38.8	33.8	53.5	61.3	50.8	49.9	54.5	52.3	67.7
	Ditrysia	27.1	36.8	47.7	75.4	52.4	43.9	57.8	68.0	55.9	59.6	71.3	60.6	72.7
	Lepidoptera	28.8	39.9	48.5	77.0	52.9	45.9	60.0	71.5	57.6	61.5	73.3	62.1	75.3
I-sites	Satuniidae	1.3	3.1	8.0	21.3	7.0	8.8	11.7	14.8	14.4	13.5	17.8	13.4	24.7
	Bombycidae	4.6	4.4	8.3	19.7	11.5	7.5	13.5	13.7	12.7	9.5	7.9	12.1	15.7
	Sphingidae	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Bombycoidea	8.5	9.6	21.6	49.2	22.0	19.3	30.8	35.8	32.2	30.4	30.7	28.5	42.4
	Macroheterocera	9.7	14.9	26.9	100.0	27.3	22.3	36.6	43.0	33.9	35.5	38.6	35.9	48.0
	Ditrysia	14.2	20.6	30.7	63.9	36.1	26.3	42.5	52.6	41.5	42.5	47.5	42.6	55.6
	Lepidoptera	15.0	23.2	31.8	63.9	37.0	26.6	44.6	54.4	44.1	43.8	48.5	43.5	58.1

* Saturniidae, Bombycidae and Spingidae: family level; Bombycoidea: Superfamily level (members of Saturniidae, Bombycidae and Spingidae); Macroheterocera: A Group/ Division within Obtectomera (members of Geometroidea, Noctuoidea and Bombycoidea); Ditrysia: Clade (All organisms except *T. gonggaensis*); Lepidoptera: Order (All selected organisms).

Table 4.5 Ka/Ks ratio of individual protein coding genes (PCGs) of *C. trifenestrata* with respect to the members of different families.

*Taxonomy level	<i>atp6</i>	<i>atp8</i>	<i>cox1</i>	<i>cox2</i>	<i>cox3</i>	<i>cytb</i>	<i>nad1</i>	<i>nad2</i>	<i>nad3</i>	<i>nad4</i>	<i>nad4l</i>	<i>nad5</i>	<i>nad6</i>
Saturniidae	0.06	0.15	0.03	0.05	0.07	0.07	0.12	0.11	0.13	0.16	0.20	0.15	0.19
Bombycidae	0.13	0.24	0.05	0.06	0.10	0.10	0.27	0.19	0.22	0.25	0.31	0.25	0.32
Spingidae	0.14	0.20	0.05	0.06	0.09	0.10	0.23	0.17	0.20	0.36	0.41	0.27	0.28
Bombycoidea	0.09	0.18	0.04	0.05	0.08	0.08	0.18	0.14	0.17	0.21	0.26	0.19	0.24
Macroheterocera	0.10	0.22	0.04	0.05	0.09	0.08	0.19	0.15	0.17	0.23	0.26	0.21	0.24
Ditrysia	0.12	0.24	0.05	0.07	0.09	0.09	0.21	0.18	0.19	0.26	0.31	0.25	0.26
Lepidoptera	0.12	0.27	0.05	0.07	0.10	0.09	0.22	0.19	0.19	0.27	0.31	0.25	0.26
<i>T. gonggaensis</i>	0.16	1.13	0.08	0.08	0.14	0.12	0.34	0.37	0.26	0.57	0.48	0.44	0.43

* Saturniidae, Bombycidae and Spingidae: family level; Bombycoidea: Superfamily level (members of Saturniidae, Bombycidae and Spingidae); Macroheterocera: A Group/ Division within Obtectomera (members of Geometroidea, Noctuoidea and Bombycoidea); Ditrysia: Clade (All organisms except *T. gonggaensis*); Lepidoptera: Order (All selected organisms).

All the genes displayed typical clover leaf secondary structures except in *tRNA^{Ile}*, *tRNA^{Arg}* and *tRNA^{Ser1}* genes which lacked TΨC loop, DHU loop and stable DHU arm, respectively (**Figure 4.4**). Lack of stable DHU arm in the structure of *tRNA^{Ser1}* is commonly observed in many lepidopterans (*S. frugiperda*, *A. hetaohei*, *B. panterinaria*, etc.) and Saturniid silkmoths including *A. assamensis* (Hong *et al.*, 2008; Yang *et al.*, 2013; Chen *et al.*, 2014; Wang *et al.*, 2016; Liu *et al.*, 2016). The tRNA genes exhibited structural variations which may be attributed to variation in the number of base pairs in tRNA genes or due to the presence of mismatches in the genes. We found that among all four arms, the number of base pairs was mostly conserved in AA stem, AC stem and AC loop while the most variable regions were TΨC stem, TΨC loop and DHU loop which resulted in variation among tRNA structures across species. For instance, DHU loop was absent in *tRNA^{Arg}* of *C. trifenestrata* which was similar to *A. theae* (with one U-U mismatch) while differed from others having variable number of bases in the loop (2 bases in *A. hetaohei*; 3 bases in *A. yamamai*, *T. gonggaensis*, *S. morio*, *E. laeta*, etc.; 4 bases in selected Bombycids, *A. selene*, *C. sasakii*, *M. flavescens*; 5 bases in *A. assamensis*, *A. pernyi*, and so on) (Liu *et al.*, 2008; Kim *et al.*, 2009; Liu *et al.*, 2012; Kim *et al.*, 2012; Gu *et al.*, 2016; Liu *et al.*, 2016; Wang *et al.*, 2016; Shi *et al.*, 2016). Similarly, lack of TΨC loop in *tRNA^{Ile}* gene was observed in *A. atlas*, *D. spectabilis*, *A. pernyi*, etc. while the bases varied from 2 to 6 bases in few organisms like *C. sasakii* (2 bases), *A. assamensis*, *S. morio*, *M. sexta* (3 bases), *A. hetaohei*, selected *Bombyx* sp., *A. yamamai*, *E. laeta* (4 bases), *S. frugiperda* (5 bases), *C. auricilius*, *M. flavescens* (6 bases), etc. (Liu *et al.*, 2008; Cameron & Whiting, 2008; Tang *et al.*, 2014; Chen *et al.*, 2014; Wu *et al.*, 2016;

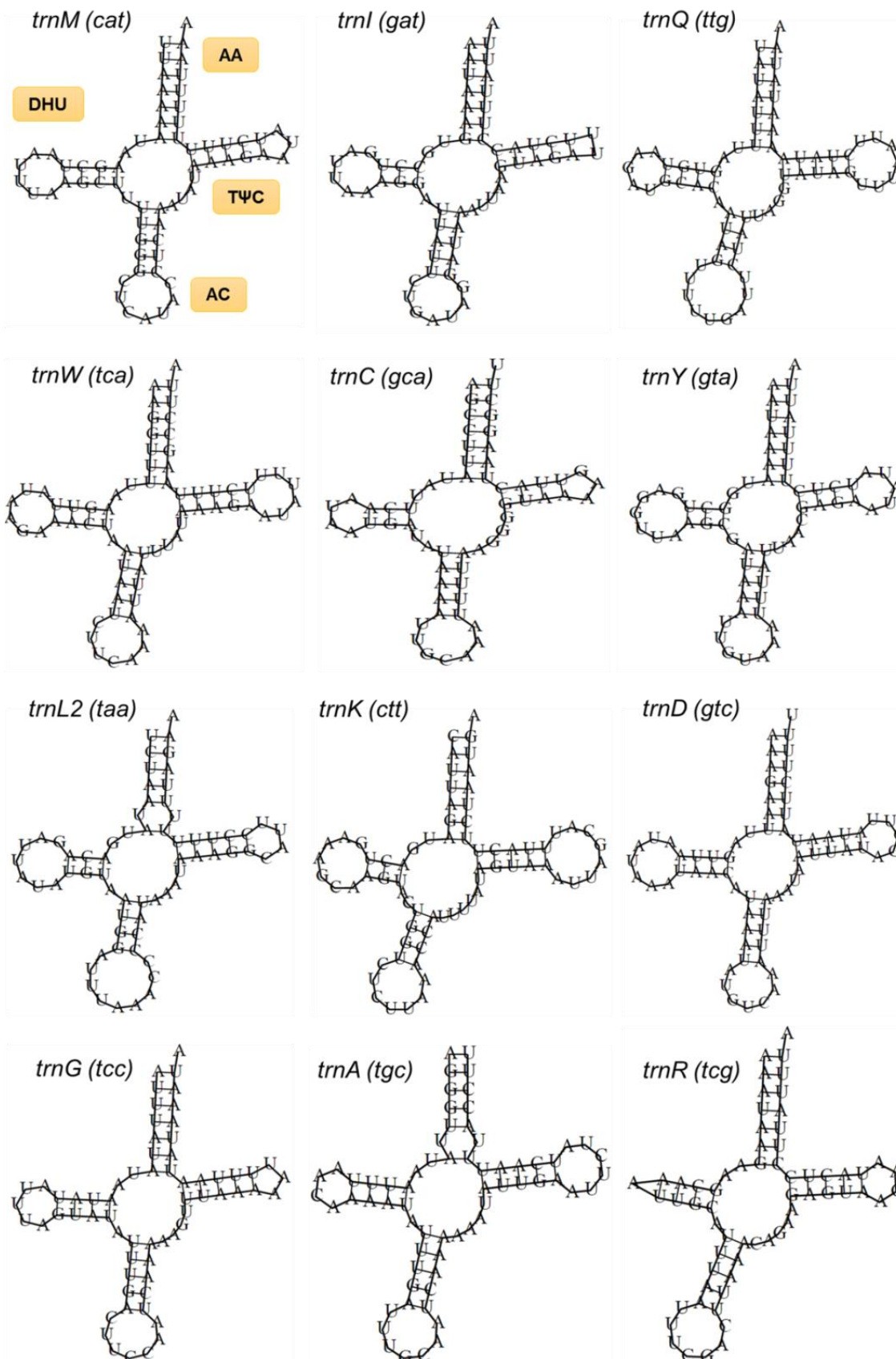
Wang *et al.*, 2016; Liu *et al.*, 2016). However, the anticodons were found to be identical to the tRNAs reported in other lepidopteran insects.

Altogether 29 mismatched pairs were identified in 16 tRNA structures of *C. trifenestrata* mitogenome (**Table 4.6**). This number was identical to *A. selene* and *E. laeta* while varied from 17 pairs (*T. gonggaensis*) to 35 pairs (*A. assamensis*) in the selected organisms (Shi *et al.*, 2016). Only two kinds of mismatches 23 “G-U” and 6 “U-U” could be identified in *C. trifenestrata* tRNAs. The number of G-U mismatches was identical to *A. pernyi*, *A. selene*, *R. menciiana* and *E. laeta* while ranged from 17 (*D. spectabilis*) to 28 pairs (*A. assamensis*) in other Bombycoids. G-U mismatches are the common non-canonical pairs (weak bond) observed in most of the lepidopterans (Liu *et al.*, 2012; Tang *et al.*, 2014; Kong & Yang, 2015; Wu *et al.*, 2016). Similar unmatched base pairs (G-A, G-U, U-U, A-A and A-C) were also reported in other lepidopterans like *T. gonggaensis* (17 No.s), *C. auricilius* (20 No.s), *S. frugiperda* (22 No.s), *Atrijuglans hetaohei* and *M. flavescens* (25 No.s), *B. panterinaria* (29 No.s), etc. (Yang *et al.*, 2013; Cao & Du, 2014; Liu *et al.*, 2016). These mismatched pairs lead to structural variations among tRNAs and also among species which ultimately results to genome sequence and structure evolution.

4.3.2.4 Comparative analysis among ribosomal RNAs (rRNAs)

The mitogenome of *C. trifenestrata* comprised of two rRNA genes, *rrnL* and *rrnS* located between *tRNA^{Leu1}*- *tRNA^{Val}* and between *tRNA^{Val}*- control region, respectively (**Figure 4.2B**). These genes were encoded by minor (-) strand similar to the selected silkmoths and other known lepidopterans (Yang *et al.*, 2013; Chen *et al.*, 2014; Tang *et al.*, 2014).

(A)



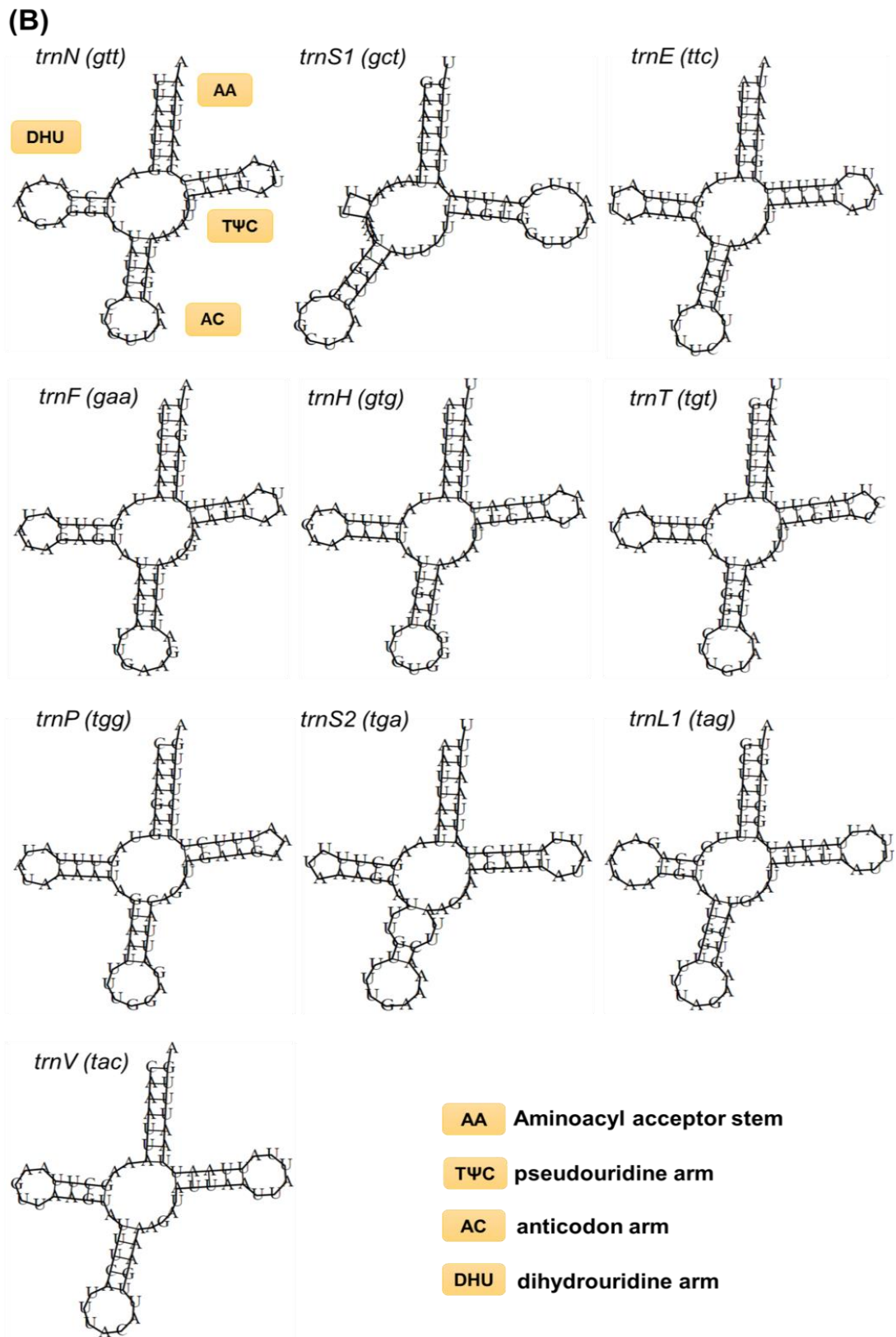


Figure 4.4 Secondary structures of 22 tRNAs of *C. trifenestrata*. (A) *trnM* to *trnR* (B) *trnN* to *trnV*. *trnY* lacked TΨC loop and *trnS1* lacked DHU loop. *tRNA* here are represented as *trn*. AA denotes amino acyl arm and AC denotes anticodon arm.

Table 4.6 Number of mismatched pairs in the tRNAs of *C. trifenestrata*.

tRNA	tRNA Stem	Mismatched pairs
<i>tRNA</i> ^{Met}	Anticodon	GU-1
<i>tRNA</i> ^{Ile}	TΨC	UU-1
	Anticodon	GU-1
<i>tRNA</i> ^{Gln}	DHU	GU-1
<i>tRNA</i> ^{Trp}	Aminoacyl acceptor	GU-1
<i>tRNA</i> ^{Leu2}	Aminoacyl acceptor	UU-1
<i>tRNA</i> ^{Gly}	DHU	GU-1
<i>tRNA</i> ^{Ala}	Aminoacyl acceptor	GU-1, UU-1
<i>tRNA</i> ^{Ser1}	TΨC	GU-1
	Anticodon	GU-1
	DHU	UU-1
<i>tRNA</i> ^{Glu}	Aminoacyl acceptor	GU-1
<i>tRNA</i> ^{Phe}	DHU	GU-2
<i>tRNA</i> ^{His}	Anticodon	GU-1
<i>tRNA</i> ^{Thr}	Anticodon	GU-1
<i>tRNA</i> ^{Pro}	Aminoacyl acceptor	GU-1
	TΨC	GU-1
	DHU	GU-1
<i>tRNA</i> ^{Ser2}	Anticodon	UU-2
<i>tRNA</i> ^{Leu1}	Aminoacyl acceptor	GU-3
	Anticodon	GU-2
	DHU	GU-1
<i>tRNA</i> ^{Val}	DHU	GU-1
GU Mismatched pairs		23
UU Mismatched pairs		6
Total Mismatched pairs		29

The size of *rrnL* and *rrnS* genes was found to be 1384 bp and 777 bp which was well within the range as reported in other lepidopteran insects (**Figure 4.3**). However, the total size of rRNAs was larger than *A. assamensis* (2123 bp) but identical to *B. mori* (2161 bp) and was lying within the range of other lepidopterans (1985 bp in *D. spectabilis* to 2261 bp in *B. panterinaria*) (Yang et al., 2013; Tang et al., 2014). Sequence similarity of *rrnL* gene demonstrated that it was highly similar to *Antheraea* species (88.9% with *A. assamensis* to 92.4% with *A.*

yamamai) and lowest with *T. gonggaensis* (80.9%). Similarly, *rrnL* genes showed highest similarity with *Antheraea* species (93.1% with *A. assamensis* to 93.6% with *A. yamamai*) and lowest with *P. oleae* (84.7%). Gene-by-gene divergence - analysis in both the genes among selected organisms also demonstrated that level of conservation (Conserved or C- sites) decreased from lower (Family) to higher (Order) taxonomical levels. Within family level, *rrnL* and *rrnS* genes revealed maximum C- sites with Sphingidae (65 & 80%) followed by Bombycidae (60 & 71%) and Saturniidae (58 & 69%), respectively. The sequence similarity level was higher in *rrnS* gene than *rrnL* gene among the selected organisms suggesting the potential of this gene as marker for species identification.

4.3.2.5 Comparative analysis among control region (A+T rich region)

The control region of *C. trifenestrata* mitogenome was found to be of 447 bp length flanked by *rrnS* and *tRNA^{Met}* genes. Comparative analysis (**Figure 4.3**) with the selected organisms revealed that this length was bigger than *A. assamensis* (328 bp) but within the size range of other selected lepidopterans (316 bp in *S. morio* to 1610 bp in *P. oleae*) (Kim *et al.*, 2013; van Asch *et al.*, 2016). The variation in length depicts that this region varies extensively among the closely related species. This may be due to the presence of variable tandem repeats which is a common characteristic feature in insects control region and differs from species to species (Liu *et al.*, 2008; Wu *et al.*, 2012; Chai & Du, 2012; van Asch *et al.*, 2016; Liu *et al.*, 2016). No tandem repeats could be detected through Tandem Repeat Finder tool but we found two 35 bp repeat units (location 77-112 & 15390-15425) in the control region of *C. trifenestrata*. The presence of tandem repeats is rare in Bombycoid insects and has been reported in few Saturniids like *A. pernyi* (six 38 bp tandem repeats), *A. proylei* (5 repeats), Bombycoids like Japanese *B.*

mandarina (three 126 bp tandem repeat unit), *B. mori* & Chinese *B. mandarina* (one repeat unit) and *R. mencia* (2.6 tandem repeats of 37 bp). These tandem repetitive elements were thought to be attributed to slipped-strand mispairing by unusual crossing during replication (Yukuhiro *et al.*, 2002; Liu *et al.*, 2008; Kong & Yang, 2015).

Sequence homology of the control region with selected organisms, however, showed some conserved features common to the lepidopteran control region. This region exhibited 84.9% similarity with *A. assamensis* and highest with *A. yamamai* (91.6%) while lowest with *L. morsei* (62.8%). Located downstream of *rrnS* gene, the motif “ATAGA” followed by poly T stretch (T)_n is known to be well-conserved in Bombycoid species and many sequenced lepidopterans as origin of minor or light strand replication (Liu *et al.*, 2008; Chen *et al.*, 2014; van Asch *et al.*, 2016; Liu *et al.*, 2016). However, this motif has been reported to be absent in the ancestor *T. gonggaensis* while in *C. trifenestrata*, “TTAGA” (**Figure 4.5**) rather than “ATAGA” was observed. We hypothesize that this motif may perform the same function like the conserved ATAGA motif as also observed in *B. huttoni* (ATAGT) motif upstream of poly T stretch. Other common features found in the Bombycoids were ATTTA elements followed by (AT/TA)_n microsatellite repeats and poly A stretch of variable length with the exception of *D. spectabilis* which lacked poly A tail. Similar to *A. assamensis*, alignment of the control region of *Cricula* with selected insects revealed two conserved blocks of more than 50 bp exhibiting greater than 90% sequence identity with Saturniids followed by Bombycids (Liu *et al.*, 2008). Thus we conclude that the control region, excluding tandem repetitive units which vary among species, remained conserved across

the Bombycoidea superfamily members suggesting their independent origins after the divergence.

>Control Region; 15099-15425, 1-120

rrnS-

TATATGTAAAATTTKTTACTTAGATTGTTTTTTTTTTTTTTTATATTAAAATATTT
AATAGACCTTAATAATTATTTAATAATTTCTCTTATTTTTTATTCATAATATTAAG
TTTAAATATAAAATGGCTATATAAAATTTGATAATTTTAAAAAAATTACTAAAAA
TTAATATAATTAATATTAATTTTTTCAATTATTATTATATATATATATATATATA
CGTATAAATATTTAATATAAAATTAAATTAATTAATTAATTTTCTAAAAACC
CCAATCTTAGGCCGTATAAACCAATTTTACATATAAATAAAAAAAATATATAT
ATACGTATAAATATTTAATATATAAAATTAATTTATTAGAAAATTAACAATTA
AAAAGATTCCTTAACCGTATAAACCAATTTTACATATAAATAAAAAAAATTAAT
ATA - *trnM*

Figure 4.5 Conserved regions in the control region of *C. trifenestrata* where **TTAGA** (orange) denotes conserved motif followed by **poly T stretch** (T)₁₉ (location 15123-15141) in green, **ATTTA** (blue) followed by **(TA)₈** in pink denotes microsatellite like element, **(AT)_n** microsatellite (location 15348-15362) and finally **poly A stretch** (A)₁₆ with intermittent T nucleotides (in red) (location 100-120). The highlighted sequence in grey denotes the **repeat element** occurring twice in the region (location 77-112 & 15390-15425).

4.3.2.6 Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions

Seventeen IGS regions with the size range of 1 to 49 bps and a total length of 166 bps were detected in the mitogenome of *C. trifenestrata*. Further, it contained seven OS ranging from 1 to 8 bp length with a total length of 21 bps (**Figure 4.2B**). In lepidopteran mitogenomes, these non-coding spacers and overlapping sequences occur frequently as a result of evolutionary pressure and contribute towards mitogenome size variation (Jiang *et al.*, 2009; Shi *et al.*, 2013; Chen *et al.*, 2014; Gong *et al.*, 2014; Liu *et al.*, 2016).

Based on comparative analysis of IGS regions of *C. trifenestrata* with the selected organisms, we found that the total size of IGS in *C. trifenestrata* was lying between *R. menciaana* (149 bp) and *A. assamensis* (171 bp) (**Table 4.7**). This size was largest in *A. theae* (509 bp) while smallest in *T. gonggaensis* (49 bp). *tRNA^{Gln}-nd2* spacer (Spacer 1) was the largest IGS among the selected silkmoths except in *M. sexta* & *B. huttoni* (2nd largest), *A. theae* (3rd largest), *B. mandarina* (5th largest) followed by *tRNA^{Lys}-tRNA^{Asp}* (Spacer 2), *tRNA^{Ser2}-nd1* (Spacer 3), *tRNA^{Cys}-tRNA^{Tyr}* and so on. Spacer 1 is common to the selected lepidopterans but absent in the ancestral moth *T. gonggaensis* (Hong *et al.*, 2008; Jiang *et al.*, 2009; Shi *et al.*, 2016).

Table 4.7 Comparative table of IGS and OS in *Cricula* mitogenome and other selected lepidopteran insects.

Species	No. of IGS	Total IGS bp	No. of OS	Total OS bp
<i>C. trifenestrata</i>	17	166	7	21
<i>A. assamensis</i>	17	171	6	23
<i>A. pernyi</i>	18	202	4	19
<i>S. ricini</i>	17	221	8	29
<i>A. yamamai</i>	17	180	10	34
<i>C. boisduvalii</i>	16	194	6	42
<i>A. selene</i>	13	137	14	56
<i>E. pyretorum</i>	17	186	7	41
<i>A. frithi</i>	18	187	7	23
<i>A. artemis</i>	13	123	11	30
<i>A. atlas</i>	12	129	5	20
<i>B. mandarina</i>	18	403	10	31
<i>B. mori</i>	20	361	5	20
<i>B. huttoni</i>	18	303	7	27
<i>R. menciaana</i>	17	149	11	34
<i>D. spectabilis</i>	17	198	7	23
<i>E. laeta</i>	18	190	9	41
<i>A. theae</i>	22	509	5	55
<i>M. sexta</i>	18	376	5	21
<i>S. morio</i>	19	210	5	21
<i>T. gonggaensis</i>	13	49	6	19
<i>B. panterinaria</i>	17	239	6	45
<i>A. hetaohei</i>	13	262	12	41
<i>C. sasakii</i>	18	198	4	19
<i>E. hippophaecolus</i>	18	289	7	32
<i>D. tethys</i>	11	179	10	25

<i>S. frugiperda</i>	13	256	7	23
<i>L. morsei</i>	11	94	13	33
<i>C. auricilius</i>	19	218	8	56
<i>A. honmai</i>	25	406	6	105
<i>P. oleae</i>	14	131	13	38
<i>M. flavescens</i>	17	167	6	29

where IGS: Intergenic spacer; OS: Overlapping sequence; bp: base pair

Based on the sequence similarity, we found only limited sequence conservation (50-62%) in this region even between species and different families of Bombycoidea superfamily. Similarly, spacer 2 showed 47-65% sequence similarity even with closely related Saturniids and is not universal as it is either absent or contain OS between $tRNA^{Lys}$ and $tRNA^{Asp}$ (e.g. in selected Bombycids, *T. gonggaensis*, *B. panterinaria*, *S. frugiperda*, etc.). Spacer 3 ($tRNA^{Ser2-nd1}$) is also conserved but variable in length among different lepidopteran insects with the exception of *D. spectabilis* (-1 bp overlap) & *E. laeta* (-2 bp overlap) (Tang et al., 2014; Wu et al., 2016). This spacer region contained a 7 bp motif “ATACTAA” which was conserved across the order (except in *T. gonggaensis*) (**Figure 4.6**). It has been proposed that this 7 bp motif helps in the site recognition by mTERM, the transcription termination peptide for which the spacer *cytB-tRNA^{Ser2}* acts as a binding site (Taanman, 1999; Salvato et al., 2008).

Similar to IGS regions, the total number of OS in *C. trifenestrata* (21 bp) was less than *A. assamensis* (23 bp) but within the range of the selected lepidopterans (19 bp in *C. sasakii* to 105 bp in *A. honmai*) (**Table 4.7**). The spacer *atp8-atp6* located between *atp8* and *atp6* genes contained a 7 bp overlapping junction “ATGATAA” conserved across Saturniids and many other lepidopteran insects (Cameron & Whiting, 2008; Liu et al., 2008; Yang et al., 2013; Gong et al., 2014; Liu et al., 2016). On the contrary, the *nad4-nad4l* overlap was absent in *C. trifenestrata* like

other closely related species *C. boisduvalii*, *A. selene* and *A. atlas* suggesting that this region is not conserved across all lepidopterans (Hong *et al.*, 2008; Kim *et al.*, 2014; Liu *et al.*, 2016; van Asch *et al.*, 2016).

Similar to that detected in *A. assamensis* and other Bombycoids, we also found sequence homology of the spacers in *C. trifenestrata* mitogenome with their adjacent genes and even with distant genes (Hong *et al.*, 2008; Kim *et al.*, 2009, Jiang *et al.*, 2009; Kim *et al.*, 2012; Liu *et al.*, 2012; Kong & Yang 2015). Spacer 1 showed 65% sequence similarity with its adjacent gene *nad2* which was less than in *A. assamensis* & *A. yamamai* (68%), *C. boisduvalii* & *E. pyretorum* (75%), *S. ricini* (77%), *A. pernyi* (71%), *B. mori* (76%), *M. sexta* (72%), *B. mandarina* (70%), etc. Similarly, the spacer 2 showed 70.97% similarity with its distant gene *nad5* while 64-66% with its adjacent genes *tRNA^{Lys}* & *tRNA^{Asp}*. Around 83-88% similarity was exhibited by spacer 3 with its distant genes *rrnL* and *nad1* which was greater than that with its adjacent genes *rrnS* & *tRNA^{Ser2}*. The occurrence of such homology with the neighbor or distant genes indicates the duplication of gene which has resulted in rapid sequence divergence among the organisms due to its non-coding nature.

4.3.2.7 Comparative analysis with respect to nucleotide composition, skewness and codon usage

The mitogenome of *C. trifenestrata* was highly AT rich with an A+T content of 77.7% (A=38.5%, T=40.4%, G=8.1%, C=13%) which was within the range of other known lepidopteran insects (Jiang *et al.*, 2009; Kim *et al.*, 2012; Cao *et al.*, 2014; Gu *et al.*, 2016).

Table 4.8 Nucleotide composition (AT and GC content) in different regions of *C. trifenestrata* and other selected lepidopteran insects.

Species	AT%					GC%				
	Whole Mito	PCGs	tRNAs	rRNAs	CR	Whole Mito	PCGs	tRNAs	rRNAs	CR
<i>A. selene</i>	78.9	77.30	80.6	83.7	87.9	21.1	22.70	19.4	16.3	12.1
<i>A. theae</i>	78.3	75.92	81.0	83.4	88.6	21.7	24.08	19.0	16.6	11.4
<i>A. assamensis</i>	80.2	78.76	80.8	84.3	91.2	19.8	21.24	19.2	15.7	8.8
<i>A. pernyi</i>	80.2	78.47	81.4	84.0	90.4	19.8	21.53	18.6	16.0	9.6
<i>A. yamamai</i>	80.3	78.89	81.4	84.1	89.5	19.7	21.11	18.6	15.9	10.5
<i>A. atlas</i>	79.3	77.68	81.2	84.2	90.5	20.7	22.32	18.8	15.8	9.5
<i>B. huttoni</i>	81.8	80.04	81.6	85.8	95.1	18.2	19.96	18.4	14.2	4.9
<i>B. mandarina</i>	81.7	79.58	81.6	85.2	95.2	18.3	20.42	18.4	14.8	4.8
<i>B. mori</i>	81.3	79.51	81.4	84.8	95.4	18.7	20.49	18.6	15.2	4.6
<i>C. trifenestrata</i>	77.7	75.54	80.5	83.6	89.9	22.3	24.46	19.5	16.4	10.1
<i>D. spectabilis</i>	79.5	77.65	81.0	83.7	92.0	20.5	22.35	19.0	16.3	8.0
<i>E. pyretorum</i>	80.8	79.35	81.9	84.5	92.2	19.2	20.65	18.1	15.5	7.8
<i>E. laeta</i>	80.2	78.65	80.6	84.7	91.4	19.8	21.35	19.4	15.3	8.6
<i>M. sexta</i>	81.8	80.24	81.7	85.4	95.4	18.2	19.76	18.3	14.6	4.6
<i>R. menciaana</i>	78.9	77.04	81.0	83.7	91.0	21.1	22.96	19.0	16.3	9.0

<i>S. ricini</i>	79.8	78.26	80.6	83.9	90.9	20.2	21.74	19.4	16.1	9.1
<i>S. boisduvalii</i>	80.6	79.10	81.9	84.5	91.5	19.4	20.90	18.1	15.5	8.5
<i>S. morio</i>	81.2	79.78	81.6	84.8	92.7	18.8	20.22	18.4	15.2	7.3
<i>T. gonggaensis</i>	81.4	79.30	83.4	85.5	89.1	18.6	20.70	16.6	14.5	10.9
<i>A. frithi</i>	80.2	78.65	81.8	84.2	89.2	19.8	21.35	18.2	15.8	10.8
<i>A. artemis</i>	78.6	76.85	80.8	83.6	91.8	21.4	23.15	19.2	16.4	8.2
<i>A. honmai</i>	80.4	78.41	80.9	84.2	94.3	19.6	21.59	19.1	15.8	5.7
<i>S. frugiperda</i>	81.1	79.55	82.4	84.3	93.9	18.9	20.45	17.6	15.7	6.1
<i>C. sasakii</i>	81.5	79.77	81.7	85.3	92.8	18.5	20.23	18.3	14.7	7.2
<i>E. hippophaecolus</i>	78.4	76.24	81.0	83.5	92.7	21.6	23.76	19.0	16.5	7.3
<i>B. panterinaria</i>	79.6	77.34	81.6	85.3	93.1	20.4	22.66	18.4	14.7	6.9
<i>P. Oleae</i>	81.8	79.05	81.3	85.0	96.2	18.2	20.95	18.7	15.0	3.8
<i>D. tethys</i>	79.3	77.40	81.4	84.0	92.0	20.7	22.60	18.6	16.0	8.0
<i>A. hetaohei</i>	81.3	79.65	81.6	85.2	95.6	18.7	20.35	18.4	14.8	4.4
<i>M. flavescens</i>	80.5	78.67	82.4	84.5	93.3	19.5	21.33	17.6	15.5	6.7
<i>L. morsei</i>	80.2	79.15	80.7	83.9	89.6	19.8	20.85	19.3	16.1	10.4
<i>C. auricilius</i>	82.0	80.56	83.1	85.5	94.1	18.0	19.44	16.9	14.5	5.9

Where CR denotes Control region

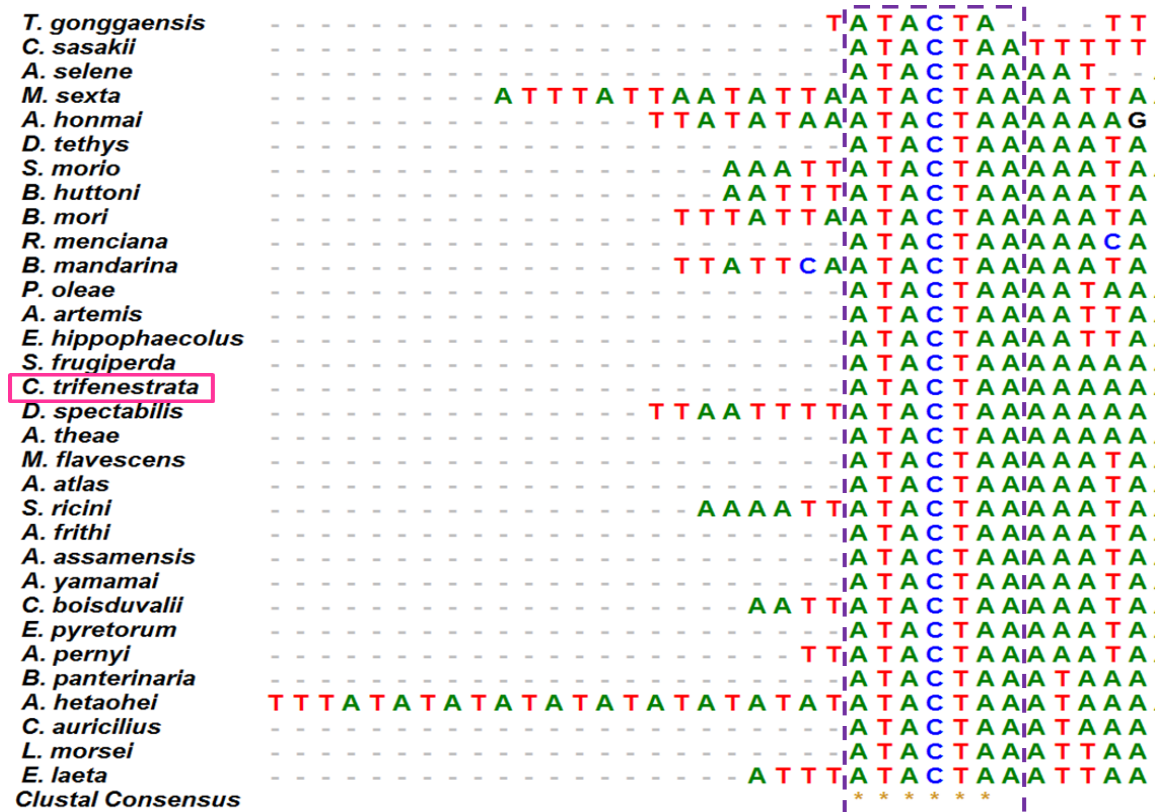


Figure 4.6 Multiple sequence alignment depicting a conserved 7 bp motif "ATACTAA" identified in *tRNA^{Ser2}-nd1* spacer of *C. trifenestrata* mitogenome similar to other selected Bombycoids. * indicates identical bases.

However, this enrichment was the lowest as compared to the selected lepidopteran insects (78.3% in *A. theae* to 82% in *C. auricilius*). On the contrary, GC content was higher (22.3%) than the selected organisms. The nucleotide content also varied between different regions of the mitogenome. We found that control region displayed highest AT biasness with an AT content of 89.9% followed by rRNAs (83.6%), tRNAs (80.5%) and PCGs (75.5%). This pattern was also followed in the selected silkmoths and many other lepidopterans (**Table 4.8**).

All PCGs exhibited high A+T content ranging from 68.7% in *cox1* gene (lowest) to 88.5% in *atp8* gene (highest) which was concordant with the selected lepidopterans except *B. mori* (*nad6* and *atp8*- highest) and *S. morio* (*nad4l*- highest). On the contrary, they exhibited low G+C content which is evident from

the evolutionary rates measured in terms of Ka/Ks ratios as discussed above. Compared to first (71.5%) and second (69.7%) codon positions, the third codon position of PCGs exhibited highest A+T content (85.4%) which has great significance in studying codon usage and gene expression among species (**Figure 4.7**). This pattern was similar to the other sequenced lepidopteran insects and indicates more relaxed constraints at third position as compared to the first and second positions due to degenerate genetic code.

In addition, the spacer 3 showed highest AT content (88.9%) than spacer 1 and 2 (83.9%) in *C. trifenestrata* which was within the range of other organisms (*B. mandarina* (88%) to *M. sexta* (96.8%)) (Yukuhiro *et al.*, 2002; Cameron & Whiting, 2008). Besides AT content, nucleotide composition of *C. trifenestrata* mitogenome was also measured in terms of relative number of Gs to Cs and As to Ts. Unlike other Saturniids, the mitogenome of *C. trifenestrata* was found to be positively AT skew (0.011) similar to *A. atlas* and ranged from -0.040 in *L. morsei* to 0.072 in *B. huttoni*. Similarly, GC skew was negative (-0.277) and lower than *A. assamensis* (-0.220) while within the range of other selected lepidopterans (-0.280 in *A. theae* to -0.162 in *C. sasakii*). This indicates that mitogenome of *Cricula* is more biased towards A and C nucleotides which also varied between major and minor strands. Such biasness was also detected in other regions like PCGs (-0.17 AT skew, -0.05 GC skew), tRNAs (0.03 AT skew, 0.16 GC skew), rRNAs (-0.02 AT skew, -0.37 GC skew) and the control region (0.04 AT skew, -0.33 GC skew).

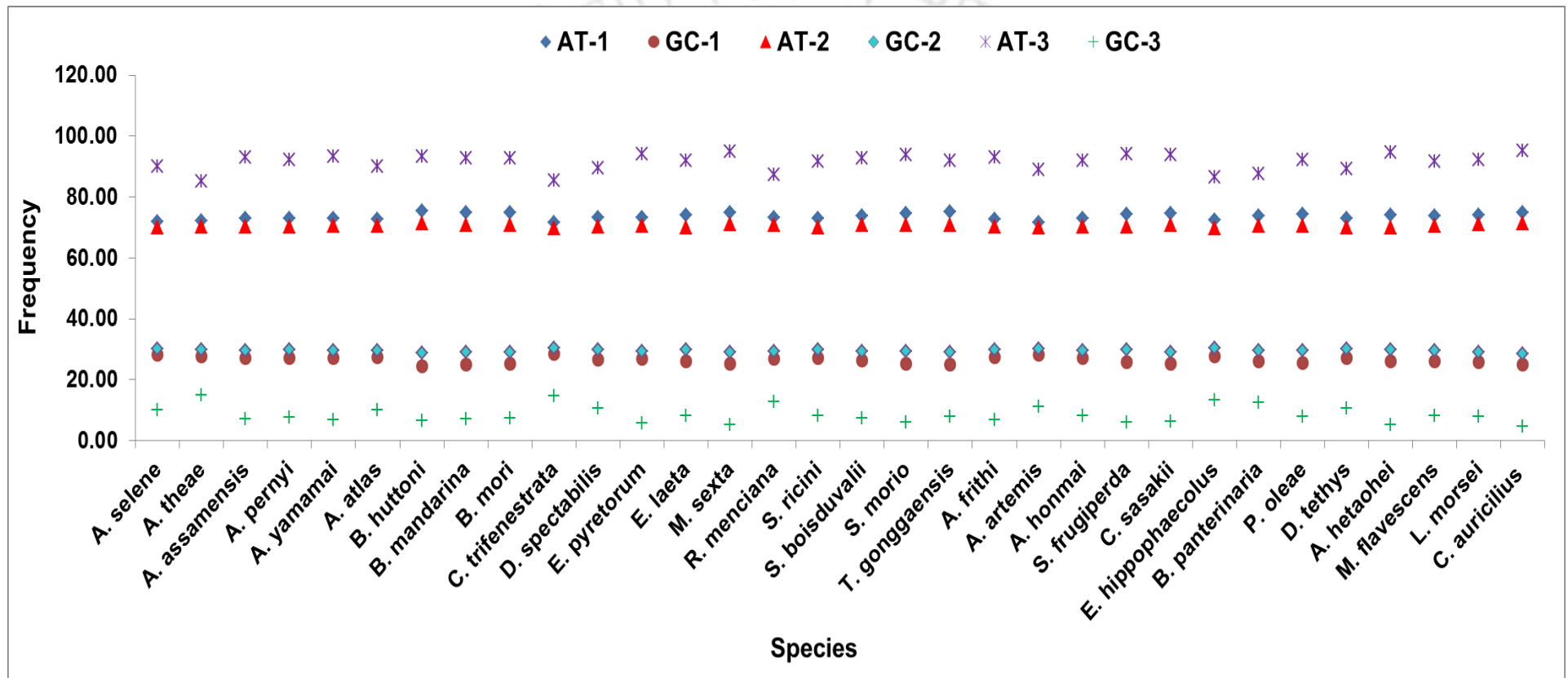


Figure 4.7 Frequency of AT and GC content at first, second and third codon positions of PCGs of *C. trifenestrata* mitogenome and other selected organisms.

Major strand of *Cricula* mitogenome exhibited negative AT skewness (-0.105) which was slightly greater than the minor strand (-0.143) while GC skewness was positive at minor strand (0.157) and negative at major strand (-0.222) similar to that obtained in *A. assamensis* mitogenome. Strand asymmetry is a common phenomenon that has also been observed in many lepidopteran insects (Kim *et al.*, 2012; Yang *et al.*, 2013; Chen *et al.*, 2014).

In order to gain insights into the evolutionary forces acting on PCGs, codon usage biasness was studied in the PCGs of *C. trifenestrata* (Kim *et al.*, 2012; Liu *et al.*, 2012; Yang *et al.*, 2013; Kong & Yang, 2015; Ramírez-Ríos *et al.*, 2016; Liu *et al.*, 2016). The relative synonymous codon usage (RSCU) of PCGs revealed high frequency of NNU or NNA codons rich in A and T nucleotides than G and C rich at the third codon position than first and second positions. This is evident from the AT content determined at the third codon position of PCGs which was higher as compared to others (**Figure 4.8A**). Among all 3728 codons (excluding termination codons) in *C. trifenestrata*, frequency of five codons UUA (L2), AUU (I), UUU (F), AUA (M) and AAU (N) was found to be the highest. These codons represent amino acids Leu, Ile, Phe, Met and Asn and were rich in A and T nucleotides thus contributing towards AT biasness (**Figure 4.8B**). The usage proportion rates (41.52%) of these prevalent codons was found to be the lowest among the selected Saturniids like *A. assamensis* (46.63%), *E. pyretorum* (46.80%), *S. boisduvalii* (46.92%), etc. as is evident from the A+T content of PCGs (75.5%). On the other hand, GC rich codons CGC, AGG, ACG, CCG, UCG, CUG, UGC, etc. were the least represented. Cys was found to be the least used amino acid similar to the other related silkmoths. From the analysis, we conclude that all codons with A or T at the third positions were overused as compared to GC codons which

were either absent or under-represented. This suggests that among various factors, compositional mutation bias or compositional constraints plays an important role in affecting codon usage. This further indicates that codon usage varies among genes as well as species and is useful for evolutionary and gene expression studies.

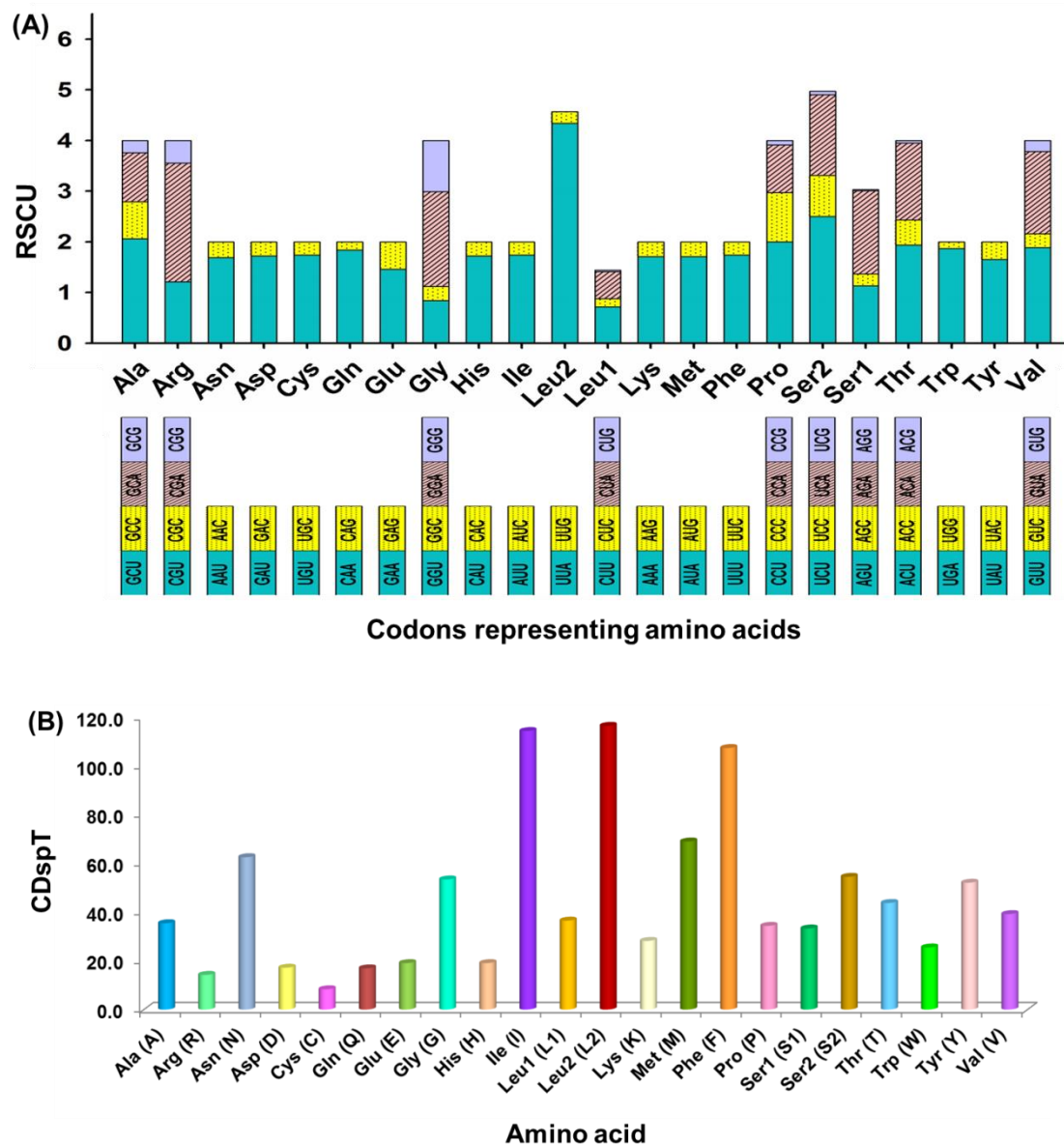


Figure 4.8 Codon usage analyses in protein coding genes (PCGs) of *C. trifenestrata* mitogenome. (A) Relative synonymous codon usage (RSCU). Codon families are represented on X-axis and RSCU values on Y-axis. (B) Codon distribution where CDspT = codons per thousand codons. Termination codons were excluded in the study.

4.3.3 Comparative phylogenetic analysis

The phylogenetic tree was constructed based on combined set of PCGs and, combined set of PCGs+rRNAs which revealed that *C. trifenestrata* belongs to Saturniidae family (**Figures 4.9 to 4.12**). The tree topology indicated that it clustered with its related species *S. bosiduvalii* and *E. pyretorum* belonging to Saturniini Tribe of Saturniidae family. The sister relationship of Saturniidae with the other two families (Bombycidae and Sphingiidae) of Bombycoidea superfamily was well supported with earlier published reports (Jiang *et al.*, 2009; Chen *et al.*, 2014). Further, sister relationship between Geometroidea and Noctuoidea was observed in our study which is also consistent with that reported in many other studies (Jiang *et al.*, 2009; Yang *et al.*, 2013). More taxon sampling and characterization of lepidopteran insects will further help in resolving the phylogenetic relationships among their members of lepidopterans at deeper levels.

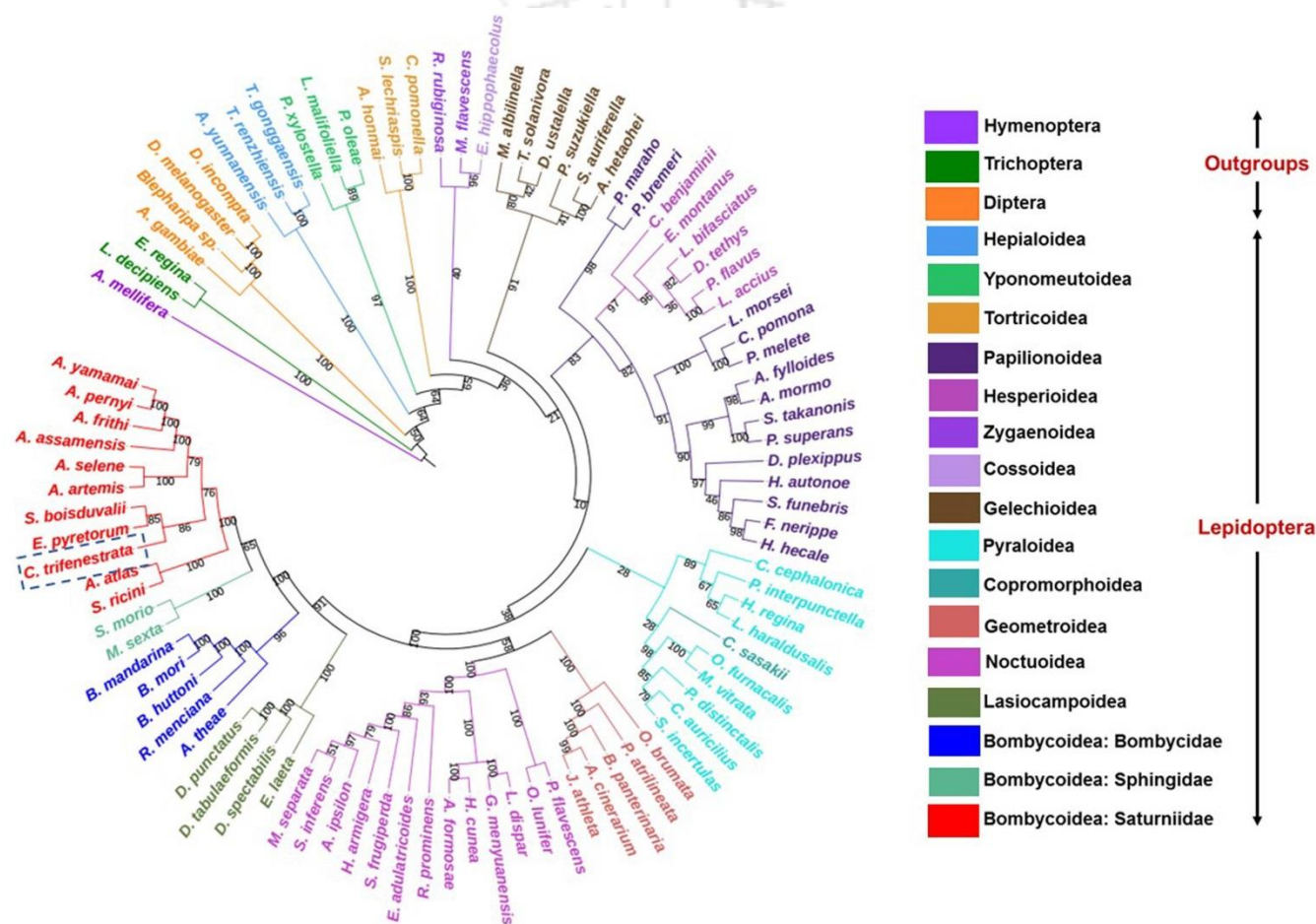


Figure 4.9 Maximum likelihood (ML) phylogenetic tree inferred using PCGs dataset in RaxmlGUI v1.3 among *Cricula* and other selected lepidopterans. The tree was obtained after 1000 bootstrap replicate analysis. Various selected families of Lepidoptera order are represented in different colors. The organism of the present study (*C. trifenestrata*) is highlighted in box.

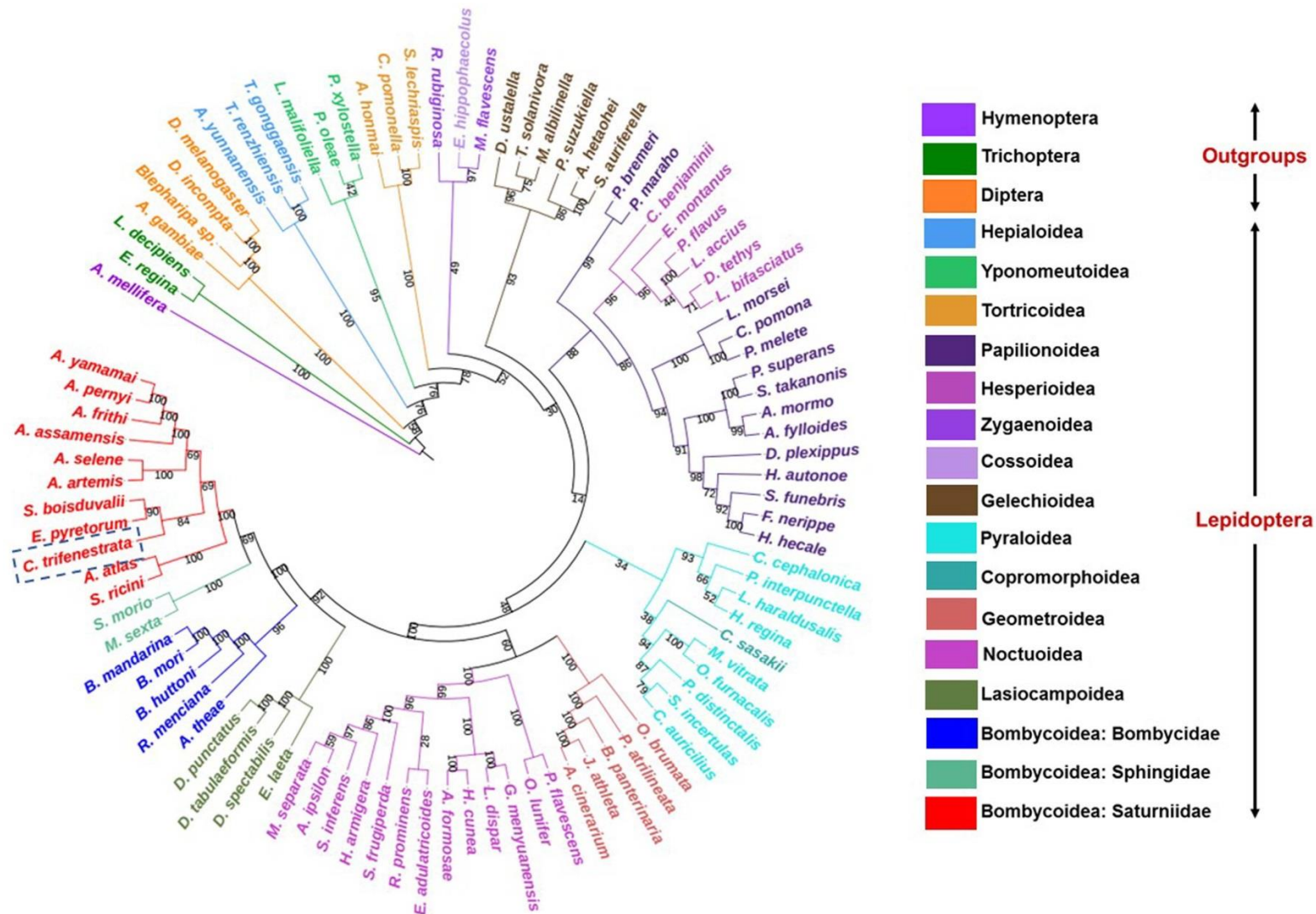


Figure 4.10 Maximum likelihood (ML) phylogenetic tree inferred using PCGs+rRNA dataset in RaxmlGUI v1.3 among *Cricula* and other selected lepidopterans. The tree was obtained after 1000 bootstrap replicate analysis. Various selected families of Lepidoptera order are represented in different colors. The organism of the present study (*C. trifenestrata*) is highlighted in box.

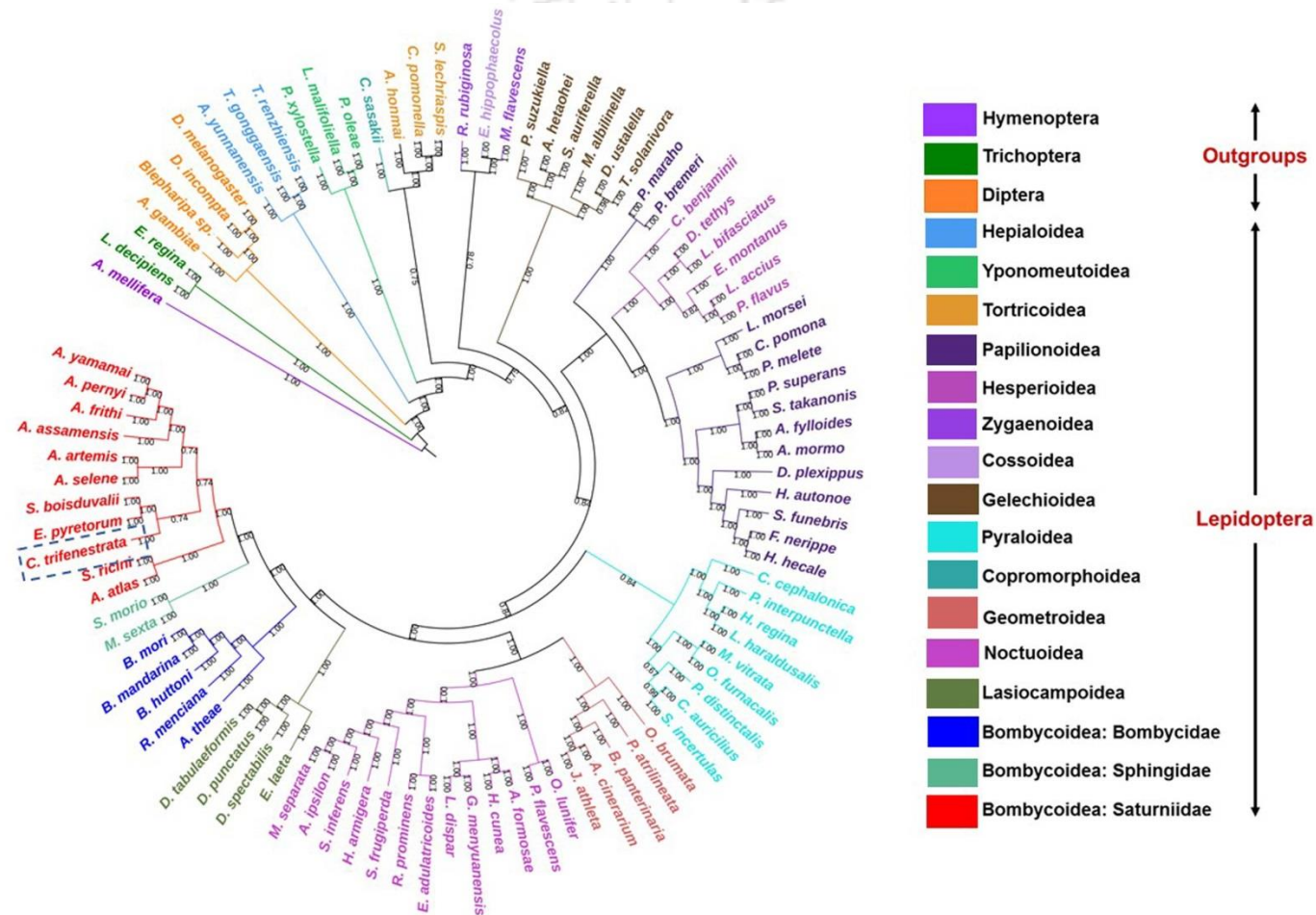


Figure 4.11 Bayesian inference (BI) phylogenetic tree inferred using PCGs dataset in MrBayes v3.2.6 among *Cricula* and other selected lepidopterans. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

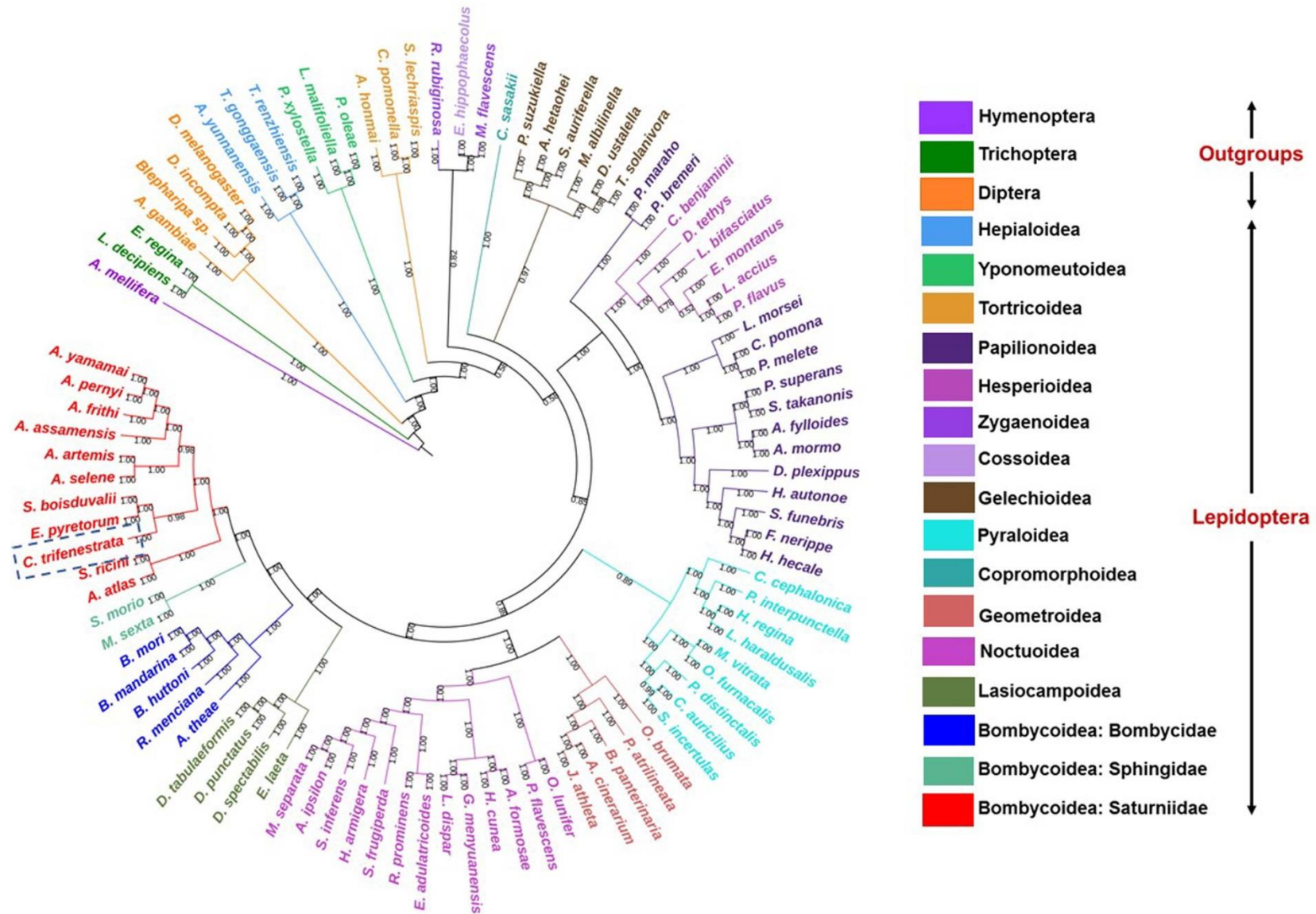


Figure 4.12 Bayesian inference (BI) phylogenetic tree inferred using PCGs+rRNAs dataset in MrBayes v3.2.6 among *Cricula* and other selected lepidopterans. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

4.4 Conclusions

The present study is a first ever report on mitogenome sequencing and characterization of *C. trifenestrata*, a wild silkworm of high economic significance. This study revealed many conserved features as that of other Saturniid insects as well as many other sequenced lepidopteran insect such as genome organization, gene content, tRNA structures, rRNA genes, conserved elements in control region, etc. Although lying within the range of other lepidopterans, the AT content was found to be lowest in *Cricula* mitogenome. Presence of TTAGA motif different from many other sequenced lepidopterans and a tandem repeat of 35 bp repeat occurring twice in the control region are key finding of the study which has potential as a marker at lower taxonomical levels. Highest variability and informative sites for *atp8* gene while lowest in *cox1* gene suggests their potential in future evolutionary studies. Moreover, compositional biasness was detected in the codon usage of protein coding genes of *Cricula* mitogenome which indicates their significance in evolutionary and gene expression studies. Phylogenetic analysis also clusters *Cricula* with its related species lying within the family Saturniidae. This study thus enriches mitogenome data of Saturniidae family which may be valuable resource for future evolutionary and marker studies.



Part B

Database Development related to Seri-bioresources

Chapter 5

Development of Seri-related Web-resources

CHAPTER 5

Development of Seri-related Web-resources

5.1 Introduction

Recent progress in the field of sequencing and analysis has led to a tremendous spike in data and development of data science tools. One of the outcomes of this scientific progress is development of numerous databases which are gaining popularity in all fields including sericulture (Flybase Consortium, 2003; Dewangan *et al.*, 2011; Federhen, 2012). As studied in earlier chapters, sericulture is the one of the important segment of agriculture due to higher demand for silk and other seri-products in the market (<http://www.inserco.org/en/statistics>). Sericultural research involves the study of silkworm diversity extending from domesticated to various wild species, their corresponding host plants, pests and pathogens which form an integral part of the seri-ecosystem (He *et al.*, 2013).

Several databases (DB) have been developed in the past decade to collate specific information on some of the components of seri-ecosystem (Mita *et al.*, 2003; Prasad *et al.*, 2005; Shimomura *et al.*, 2009; Thakur *et al.*, 2011; Zhang *et al.*, 2014). Our search attempt has led to the identification of more than 70 databases developed from 2003 till now (**Figure 5.1**) which comprise of genome, proteome, transcriptome and other data of silkworms, host plants, pest and pathogens, etc.

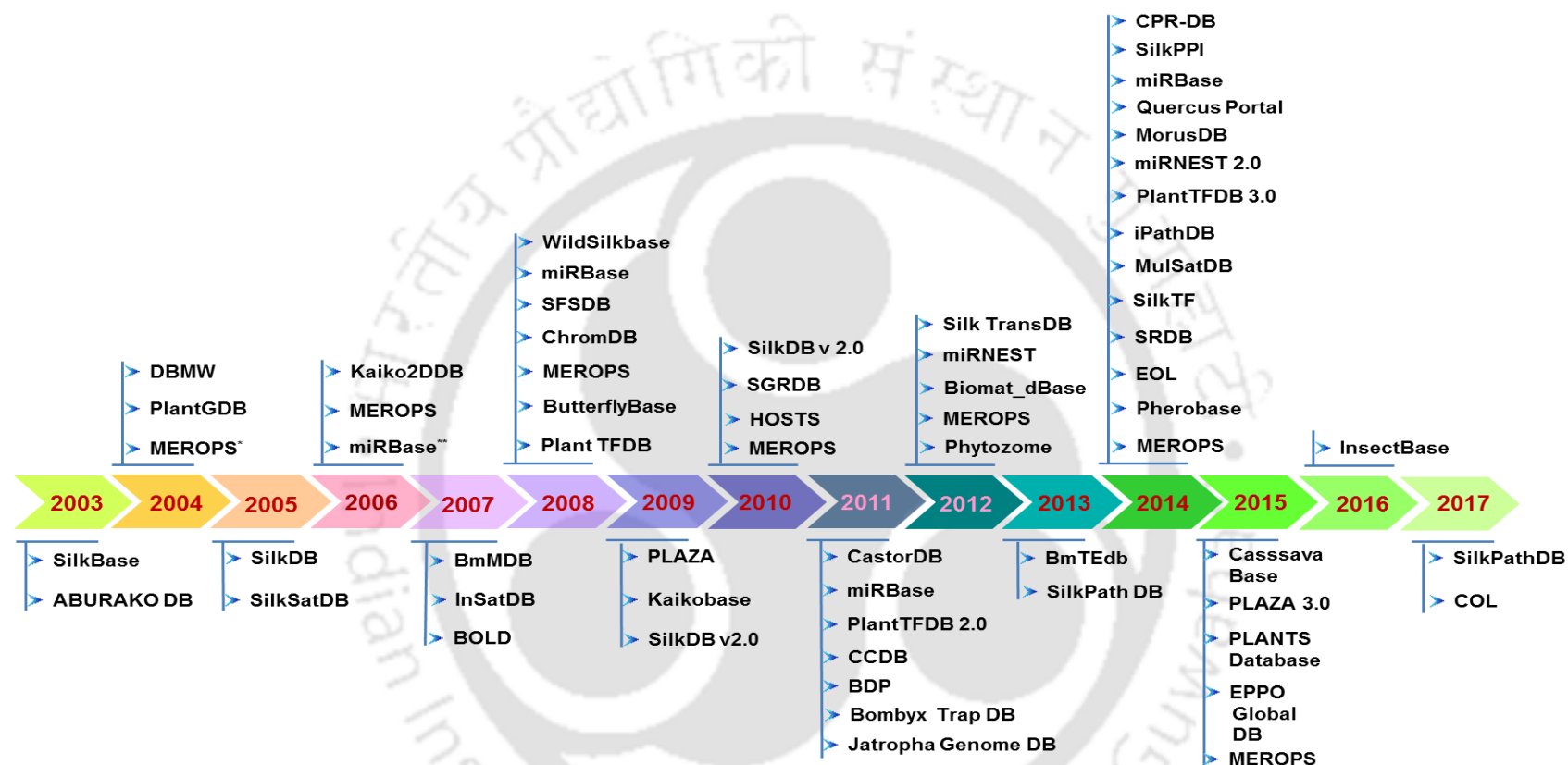


Figure 5.1 Timeline of the existing seridatabases from the year 2003 to 2017 generated using respective publication in the literature and database creation year from websites, where (*) indicates database first published in 1999 and its updated versions considered from period 2003-2017; (**) indicates the same database with updated information. (Abbreviations of the web-resources are detailed in **Figure 5.5**).

The comparative features of these databases are depicted in **Figure 5.2**. Some of these databases are being regularly updated and made accessible to researchers working in the relevant field of science such as seri-genomics, proteomics, seri-bioinformatics, etc. (Singh *et al.*, 2016). However, some of these databases are not easily available to users due to their minimal turn-up during common searches using popular search engines. Again, some of the related web-resources are not published in literature and hence, are not known to common people. So, in order to make these prevailing seri-resources easily accessible, it is essential to build a common platform comprising the data underlying these resources.

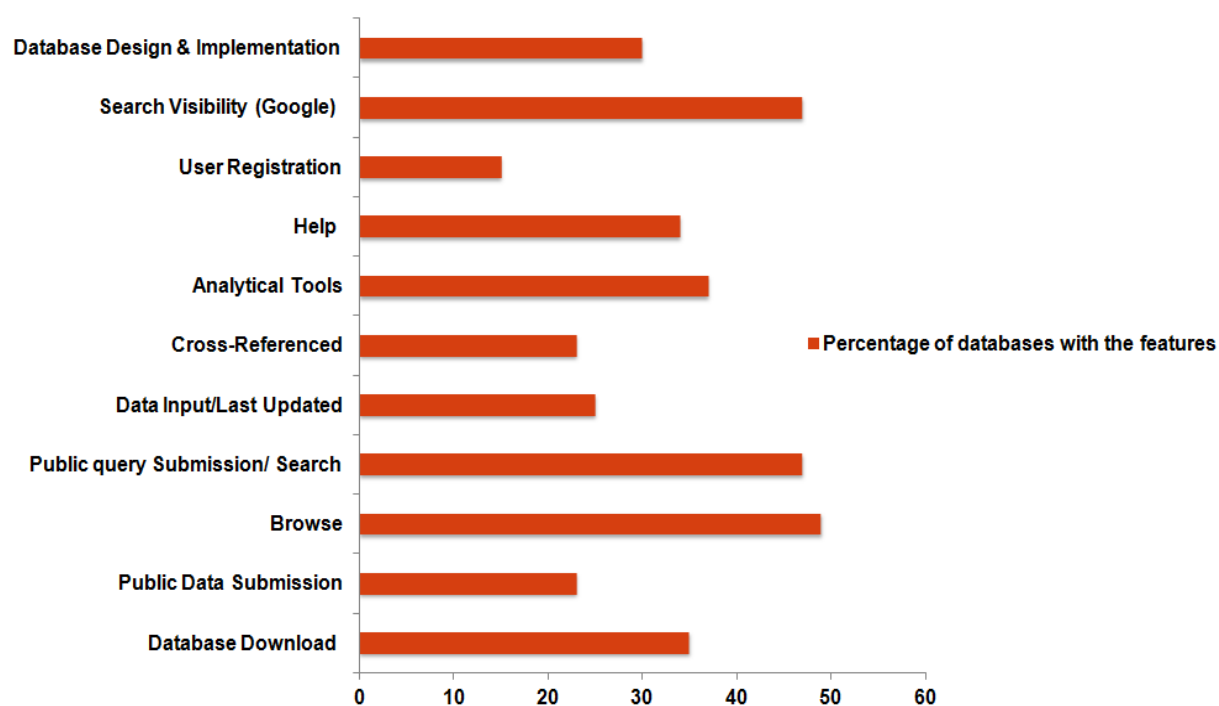


Figure 5.2 Percentage of databases with the features (Analysis was based on detailed review on various features of all available seri-related databases).

Moreover, most of the databases developed so far were more focused on domesticated silkmooths while only few web-resources are available which enlist some wild silkmooths (or Saturniid silkmooths) such as WildSilkBase, SILKS, etc. (Beccaloni *et al.*, 2003; Arunkumar *et al.*, 2008; Robinson *et al.*, 2010; Li *et al.*, 2017). Also, details of wild seri-resources including their host plants, pests and

pathogens are sparingly available in research articles (Singh & Chakravorty, 2006; Tikader *et al.*, 2013). Therefore, creation of a common platform is an immediate requirement for wild silkworm bio-resources other than domesticated ones.

Wild silkworms are distributed over the largest family (Saturniidae) of Lepidoptera order. This family exhibits great diversity in organism's size, behavior (mostly nocturnal), color (bright), silk type and production, univoltine or multivoltine depending upon the climatic conditions (Tuskes *et al.*, 2000; Kakati & Chutia, 2009). Among 1200-1800 wild sericigenous species reported based on previous literature, majority of them are not yet explored (Nassig *et al.*, 1996; Lemaire & Minet, 1998). North-East India, one of the 35 hot spots of biodiversity in the world, makes an ideal habitat for such wild sericigenous insects (Singh & Chakravorty, 2006; Lalhmingliani *et al.*, 2013). Identification of wild species and new strains along with their ecosystem is thus needed to be studied and documented. Identification and characterization through molecular tools like DNA barcoding and omics tools like genomics, organellar genomics, transcriptomics, etc. can vastly assist in their exploration and conservation of seribiodiversity.

Presently, no single data resource includes the distribution, barcode and genome data status in a common platform. Therefore, there is a need of database that can document wild seri-bioresources in a single platform. Unfortunately, there is no such database available till now which provides such comprehensive information of the wild species, their host plants, diseases, pests and predators.

In the present study, we have created two seri-related web-resources- "SeriPort" and "Seribase". "SeriPort", a web-portal, aims to systematically caterorize the available databases based on their type. The e-resources are grouped into silkworm, host plant, pest/pathogen and combined databases. These are further sub-categorized according to the core data. In addition, the genome, proteome

and transcriptome information of the existing silkworms and their resources is made accessible. On the other hand, “Seribase” consists of information on wild silkworms belonging to Saturniidae family along with their interacting partners involved in the seri-ecosystem. The information includes distribution, taxonomic classification, barcode data, whole genome, organellar genome and transcriptome sequencing status. In addition, the patents published in sericulture area are also accessible. We expect that the web-resources developed will enable the users to explore the seri-resources enlisted in the database, easily search existing seridatabases, analyze them and construct new databases thus bridging the gaps in sericulture and seri-informatics fields.

5.2 Materials and methods

SeriPort is a user-friendly and multilingual web-interface mainly comprising of a collection of web resources related to the seri-resources. Seribase enlists wild sericigenous insects along with their corresponding host plants, pest and pathogens. The workflow for the construction of SeriPort and Seribase are systematically represented in **Figure 5.3 and 5.4**.

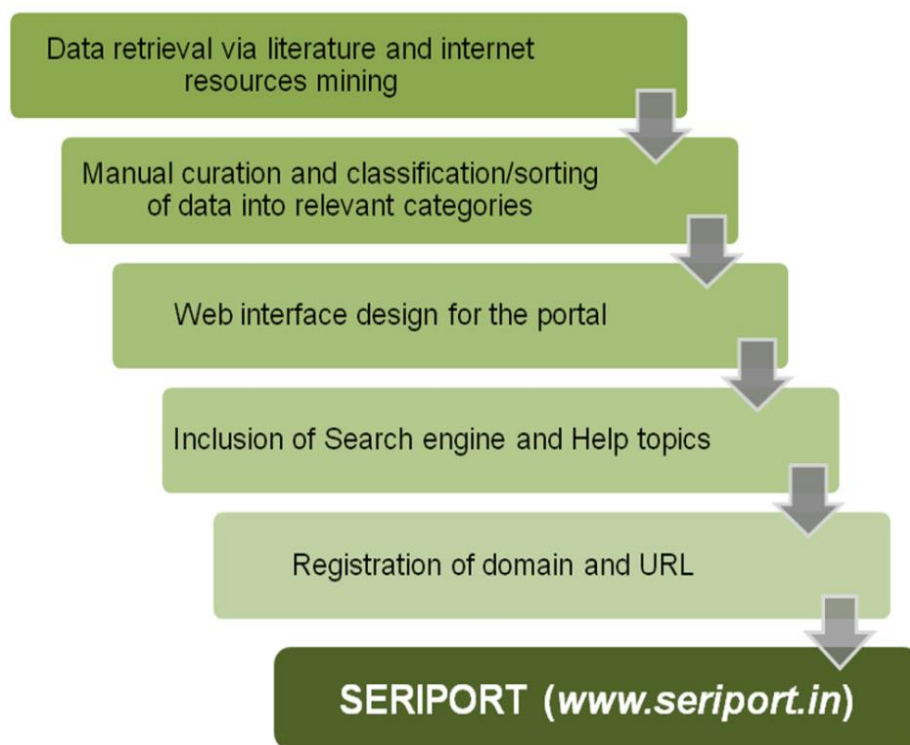


Figure 5.3 Workflow for the construction of SeriPort.

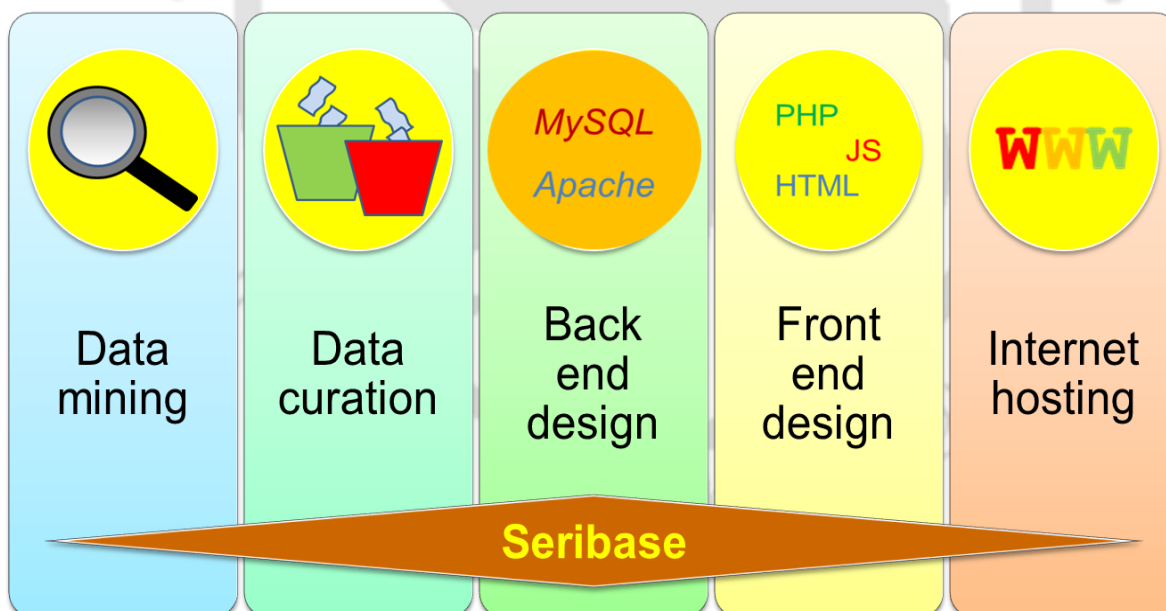


Figure 5.4 Workflow for the construction of Seribase.

The detailed information on data retrieval, content, design, data query and other necessary features of the databases are provided as below-

5.2.1 Database content- retrieval, curation and categorization

SeriPort

The data on available databases/web-resources related to sericulture area was retrieved through literature and internet mining. Data mining resulted in around 75 seri-related databases which were further classified based on the type of information. Most of the databases are full-fledged and being updated whereas few are under development. The data were verified manually and updated according to the latest revisions in the databases listed. The databases (DB) in SeriPort were organized into four main categories- 22 silkworm, 25 host plant, 2 silkworm pest/pathogen and 26 combined databases as follows-

- Silkworm DB- Set of databases directly providing data related to silkworms.
- Host plant DB- Databases providing specific data on silkworm host plants.
- Pest and Pathogen DB- Databases providing data related to pest or pathogens of silkworms.
- Combined DB- Databases comprising of data which belong to categories other than the above.

The Silkworm DB was further grouped into four sub-categories- Nucleotide DB, Protein DB, Genetic resource DB and Insect pathway DB. Nucleotide DB covers the diverse nucleotide information like ESTs, genomic sequences, microsatellites, microarrays, etc. Protein DB deals with the protein profiles at various developmental stages, protein-protein interaction information, etc. Genetic resource DB houses the characterization information of life stages of silkworm strains, varieties, etc. In addition, the insect pathway construction data (10

lepidopteran insects) based on the genomic and transcriptomic sequences can be obtained from Insect pathway DB (Shimomura *et al.*, 2009; Li *et al.*, 2012; Xu *et al.*, 2013; Zhang *et al.*, 2014). Likewise, the host plant DB was classified based on their respective host plants like Mulberry, Castor, *Jatropha*, Cassava, Papaya, Oak, etc. (Thakur *et al.*, 2011; Li *et al.*, 2014; Krishnan *et al.*, 2014). Only one silkworm pathogen database comprises of genomic information of various pathogens as well the transcriptomic profile of silkworms under infected conditions (Li *et al.*, 2017).

Moreover, the generalized data of silkworms and their resources is also included in the section 'Combined DB'. It consists of data on barcode sequences, miRNA candidates, their structure, annotated miRNA sequences, nomenclature and classification of peptidases, etc. The information regarding structure, abundance and applications of silk biomaterials; silk fabric characteristics as well as historical, field and remote-sensing image data of the ancient Silk route is also incorporated in combined DB (Ratnasingham & Hebert, 2007; Kundu, 2012). Further, databases on taxonomy, distribution, common or scientific names, pheromones of insects and their host plants have been included in the combined DB part of SeriPort. The diagrammatic representation of the various databases mentioned in SeriPort is depicted in **Figure 5.5**.

The updated genomic, proteomic and transcriptomic information of the silkworms and their resources was also retrieved from Scopus (<http://www.scopus.com/>), stored in spreadsheets and made available in SeriPort. Moreover, the portal enlists a number of sericulture related organizations and industries in a tabular format.

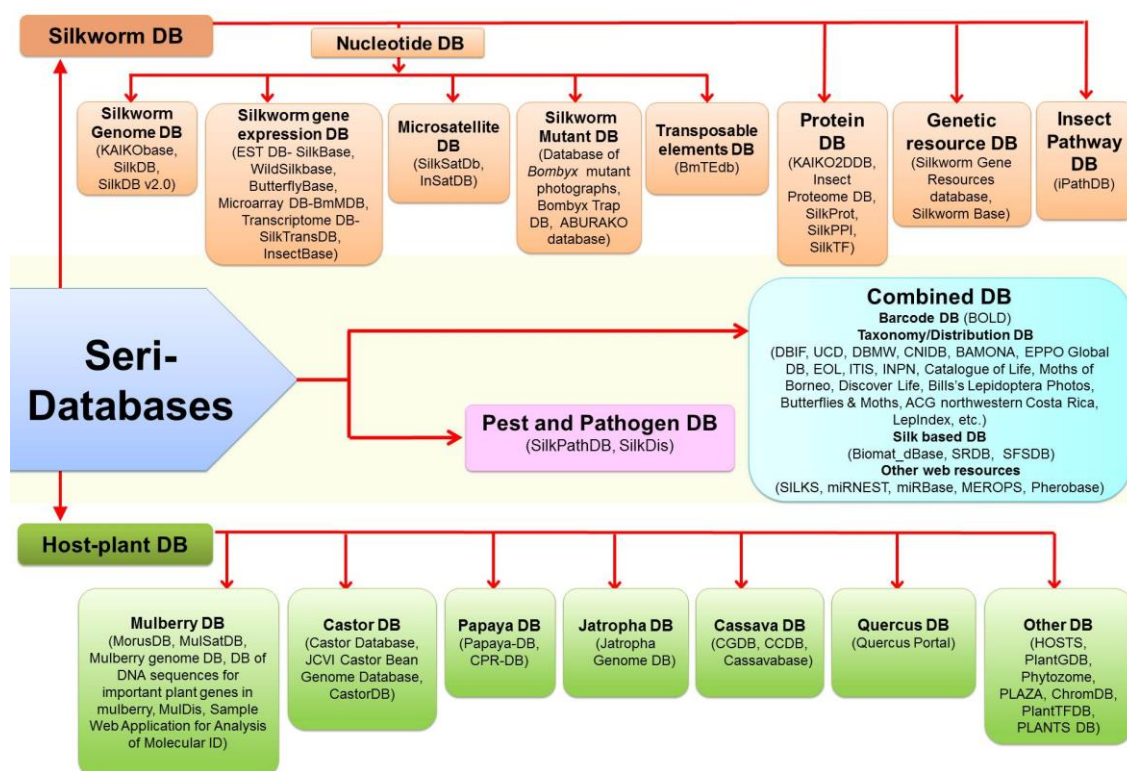


Figure 5.5 Schematic representation of Seri-databases classified into four categories- Silkworm Databases- 22 No.s, Host Plant Databases- 25 No.s, Pest and Pathogen Databases- 02 No.s, Combined Databases- 26 No.s.

(Abbreviations- **SilkDB**: Silkworm Knowledgebase, **EST DB**: Expressed Sequence Tag Database, **BmMDB**: *Bombyx mori* Microarray Database, **SilkTransDB**: Silkworm Transcriptome Database, **SilkSatDb**: Silkworm Microsatellite Database, **DBMP**: Database of *Bombyx* mutant photographs, **BmTEdb**: Transposable elements database for *B. mori*, **Insect Proteome DB**: Insect Proteome Database, **SilkProt**: Annotated protein database of silkworm, **SilkPPI**: Silkworm Protein- Protein Interaction database, **SilkTF**: Silkworm Transcription Factor Database, **SGRDB**: Silkworm Gene Resources database, **iPathDB**: Insect Pathway Database, **MorusDB**: Morus Genome Database, **MulSatDB**: Mulberry Microsatellite Database, **CastorDB**: A comprehensive knowledgebase database for *R. communis*, **Papaya-DB**: Papaya Genomic Resources Online, **CPR-DB**: Papaya Repeat Database, **CGDB**: Cassava Genome Database, **CCDB**: Chinese Cassava Genome Database, **HOSTS**: a Database of the World's Lepidopteran Host plants, **PlantGDB**: Resources for Comparative Plant Genomics, **ChromDB**: The Chromatin Database, **PlantTFDB**: Plant Transcription Factor Database, **SilkPathDB**: Silkworm Pathogen Database, **SilkDis**: Silkworm Disease Database, **BOLD**: Barcode of Life Data System, **DBIF**: Database of Insects and their Food Plants, **UCD**: Universal Chalcidoidea Database, **DBMW**: Database of Butterflies and Moth of the World, **CNIDB**: Common Names of Insects Databases, **BAMONA**: Butterflies and moths of North America, **EOL**: Encyclopedia of Life, **ITIS**: Integrated taxonomic information system, **INPN**: *Inventaire National du Patrimoine Naturel*, **ACG**: Area de

Conservación Guanacaste, **SRDB**: Spatio-temporal database of Silk Road, **SFSDB**: Silk Fabric Specification Database, **SILKS**: Sericulture Information Linkages and Knowledge System, **miRNEST**: An integrative microRNA resource, **miRBase**: The microRNA database, **MEROPS**: the peptidase database).

Seribase

For construction of Seribase, wild silk moths mainly belonging to family Saturniidae, their host plants, pests and disease causing pathogens were listed in Microsoft excel sheets and presented in the database. Common search engines used for data mining were Google, Scopus, Web of Science and ScienceDirect. The search was performed using various keywords like silkworm name (scientific or common name), silkworm+host plant, silkworm+pest, silkworm+pathogen, silkmoths of Saturniidae family, silkmoths of Bombycoidea superfamily, etc. The scientific names, common names, taxonomy and distribution were searched along with the images of respective silkmoths. The copyright of the images used in the database were obtained by respective photographer or museum. The status of barcode information of the selected organisms were obtained from BOLD; genomic, organellar (mitochondrial and chloroplast) and transcriptome information were retrieved from NCBI database and publications. Finally, the patent data related to silk, sericulture, biomedical applications of silk, textile, etc. was acquired from Google patents, Espacenet, PATENTSCOPE, etc. The data was manually curated and categorized based on the information type.

Seribase comprises of datasets categorized into six different modules (**Figure 5.6**) - “Silkworm”, “Host plants”, “Silkworm Pests”, “Silkworm Pathogens”, “Patents” and “Related Publications”. It has an easy-to-use web-interface equipped with various options like search, browse and download for easy retrieval of data. Currently, the database enlists more than 1600 wild seri-resources (501 species of wild silkmoths, 1023 host plants of the respective silkmoths, 35 disease

pathogens and 90 pests of wild silkmoths. More than 300 patents have been included in the database till now. Each entry in the database was provided with a unique identity (ID) number 'SBDB_XX0001' where the prefix remained same while the suffix was assigned as 'SW' for silkworm and pathogen, 'HP' for host plant, 'SP' for silkworm pest and 'PT' for patents.

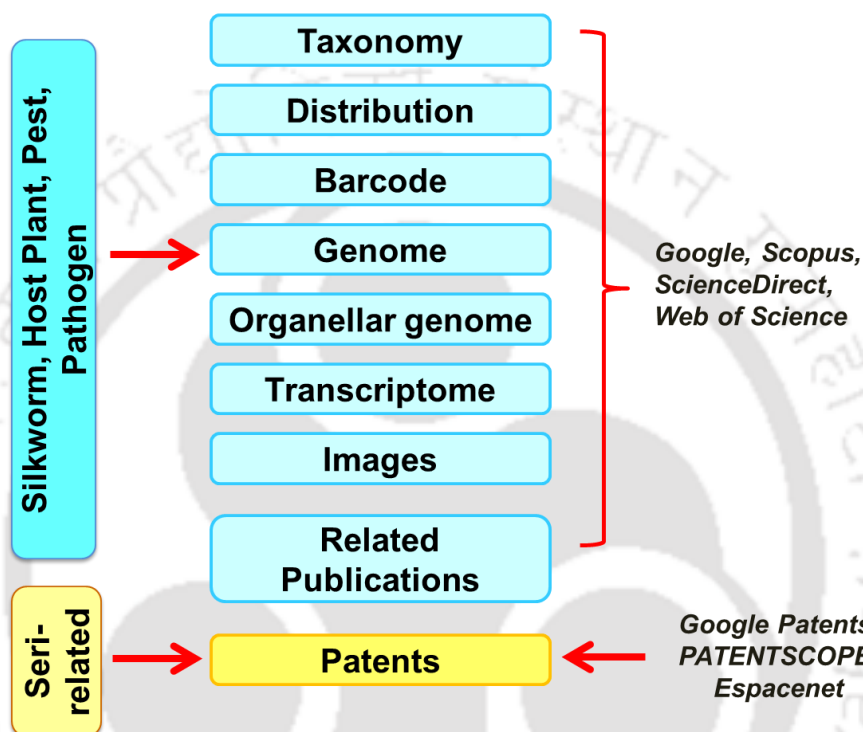


Figure 5.6 Data content in Seribase.

5.2.2 Design and implementation of SeriPort and Seribase

SeriPort

The database for SeriPort was designed using Basic HTML 5 and CSS 3 for front-end and PhpMyAdmin and MySQL 5 for the back end (**Figure 5.7**). The working language of website is PHP and HTML 5. The connectivity between front end and back end was done by using PHP. The website was implemented using a responsive design so that it performs well not only on desktop systems but also on mobile devices (Android, iOS, etc.).

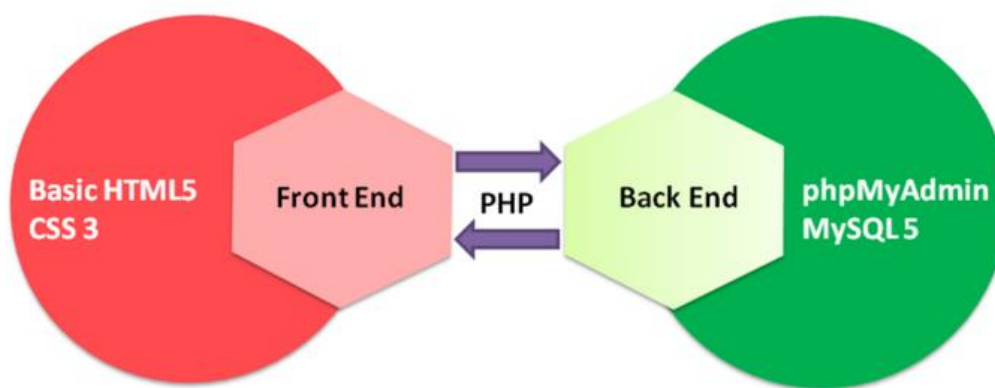


Figure 5.7 Schematic of database design for SeriPort.

Seribase

Seribase has been implemented in a PHP (Hypertext Preprocessor) module with php and HTML scripts running in the front end whereas MySQL database and Apache web server at the back end (**Figure 5.8**). The complete architecture resides upon a Linux-based server hosted at Arvixe. The web interface comprises a collection of web pages related to the data. It is noteworthy to mention that initial data consisted of spreadsheets later converted to sql format using phpMyAdmin. SeriBase is compatible with commonly used web browsers such as Google Chrome or Mozilla Firefox and also smartphone compatible. Various aesthetic aspects of the database were modulated via CSS styles.

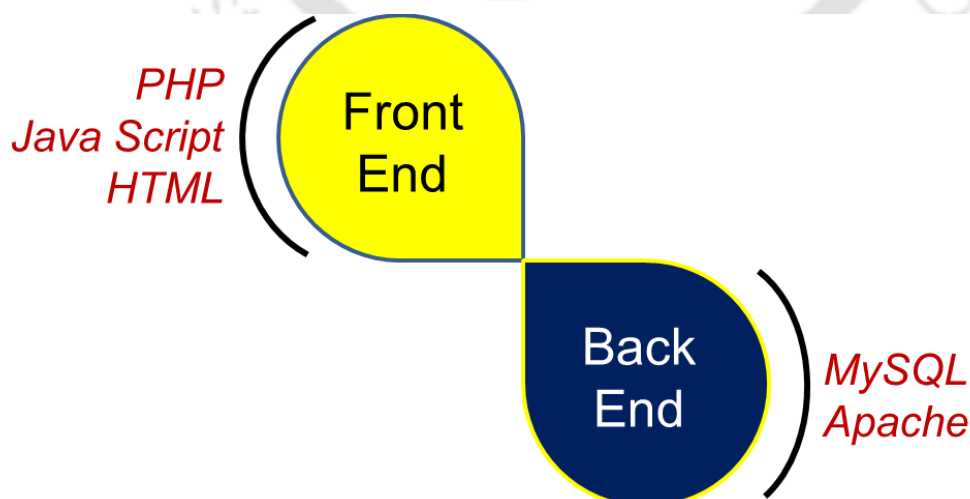


Figure 5.8 Schematic of database design for Seribase.

5.3 Salient features and accessibility of SeriPort and Seribase

SeriPort

The web interface of Seriport consists of a collection of web pages like Home page, Search Page, Download page, Publications, etc. The home page of SeriPort provides a brief view of what SeriPort is about (**Figure 5.9**). The contents of SeriPort have been hyperlinked to their respective websites. A separate webpage was created for each database which consists of a short description and its web link. 'Search' page is described in the next section. 'Download' option has been provided that enables the user to download seri-related data as discussed later. 'Publications' option directly links to the publications of respective databases. The references and links associated with the databases are included in the 'References' option separately. 'Help' tab leads a user to commonly encountered help topics. 'Feedback' option has been provided to receive suggestions from users regarding the portal. The portal is freely accessible to the users at www.seriport.in.

Seribase

Seribase consists of a user-friendly interface equipped with several necessary features. The home page of Seribase (**Figure 5.10**) provides a brief introduction on the information contained in the database and is linked to "About" page which acquaints the users with the data type (e.g. distribution, barcode, genome, etc.). The data in Seribase can be accessed through options like search, browse, download, submit data and help options as discussed in separate sections below. Seribase is freely accessible to the users at <http://www.seribase.in>.



Figure 5.9 Home Page of SeriPort.

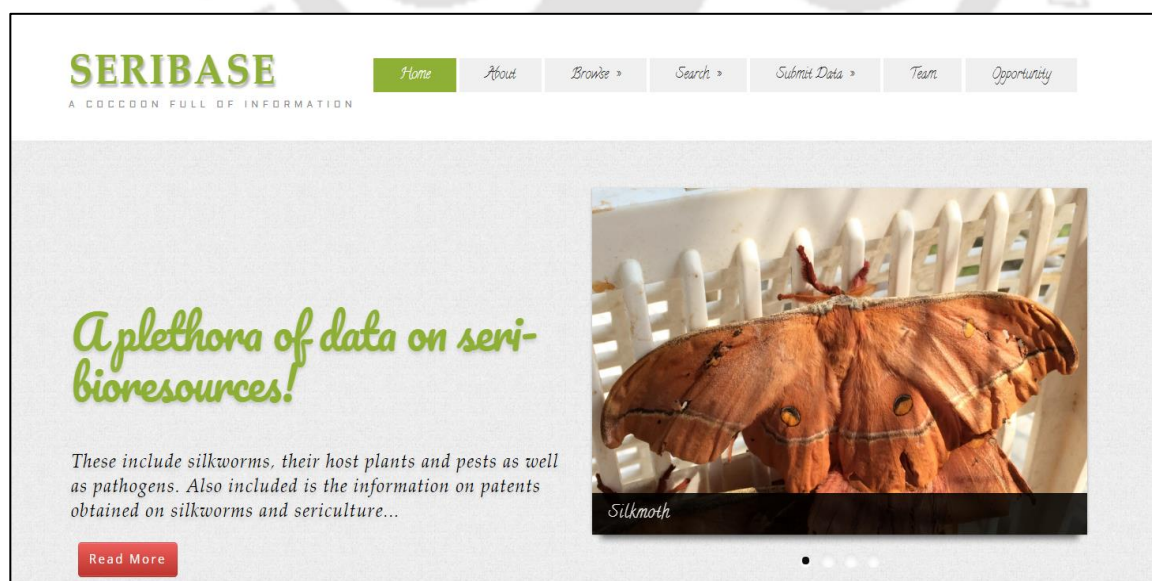


Figure 5.10 Home Page of Seribase.

5.4 Data access in the portal and database

SeriPort

Provisions have been made to search the data based on specific keywords like database name, silkworm name, host plant name, etc. A large range of keywords were selected for each database and used as tags for our search mechanism. A common search bar i.e. “Search content” has been provided at right side of each page of the portal (**Figure 5.11**). Users can search the data from the search bar by entering any keyword e.g. Silkworm, host plant, combined, pest and pathogen, WildSilkDB, SilkDB, MorusDB, MulSatDB, etc. The search will lead to a brief description of the database related to that keyword along with a web link for the respective database. The web link will further lead to the home page of original database with detailed information.

Seribase

Data access and viewing of Seribase is provided using two components- “Browse” and “Search”. The browse feature was created for flexible navigation of the database contents. It consists of HTML in the front-end and MySQL database as back-end. It enables a user to dynamically view, sort and search all the available information in database of each data type, namely, silkworms, silkworm hostplants, silkworm pests, silkworm pathogens and patents (**Figure 5.12**).

The search page of each data type described above has been individually prepared for specific querying of different data types (**Figure 5.13**). Each search page acts as a stand-alone gateway to query each sql database contents dynamically as instructed by the user. The search feature enables a user to query the database in multiple fields of the sql database. The common query for each database is SBDB_ID (a database-specific identifier for each data entry). Other

than that, a multiple number of keywords can be used which corresponds to various data fields of the sql databases. For instance- search options for 'Silkworm' database are "common or alternative or scientific name of a silkworm", "taxonomical class", "location or distribution", "Barcode ID (NCBI) or Barcode sequence", "whole genome sequence or transcriptome NCBI ID".

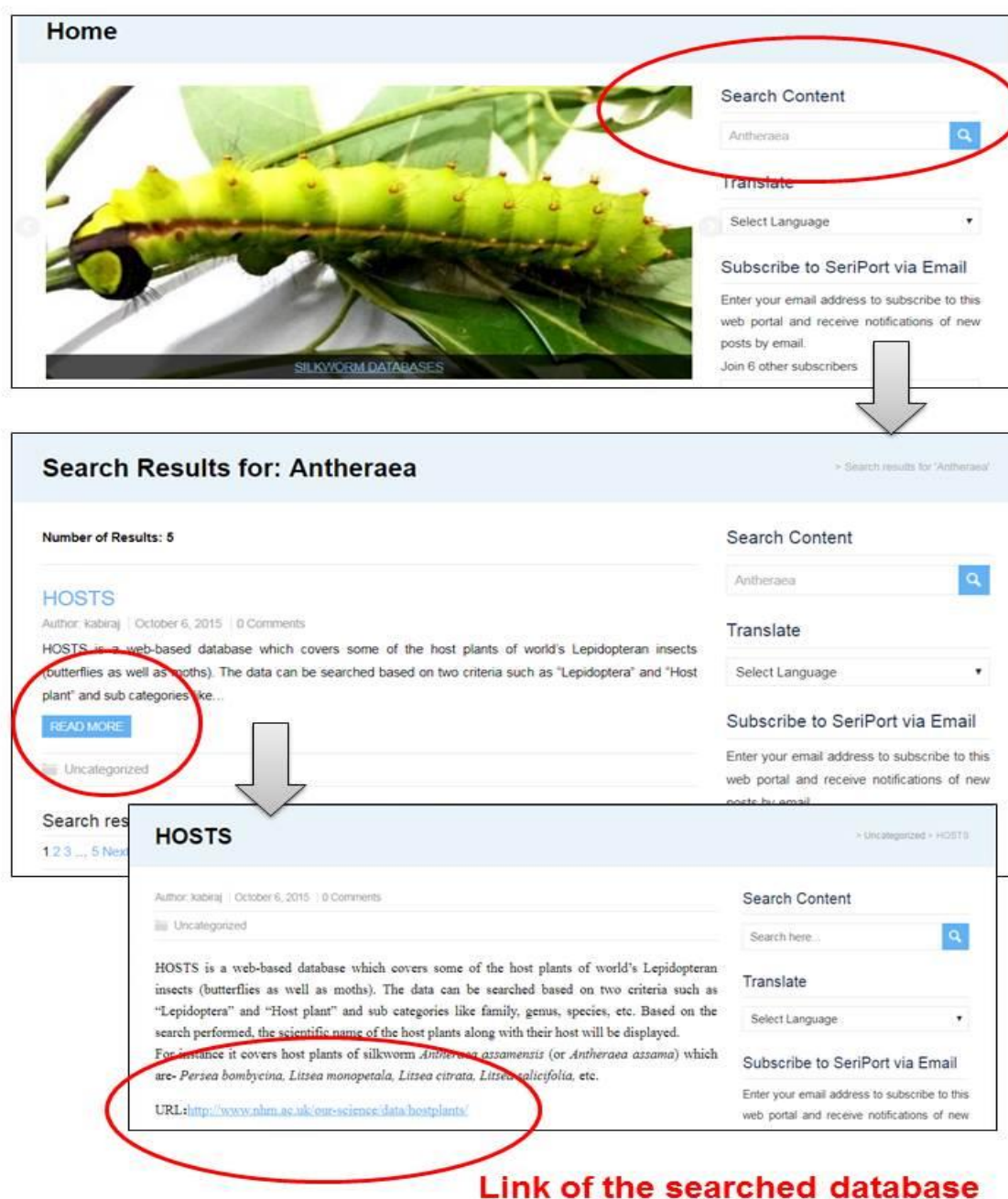


Figure 5.11 Search option in SeriPort and snapshots of search results based on keyword "Antheraea".

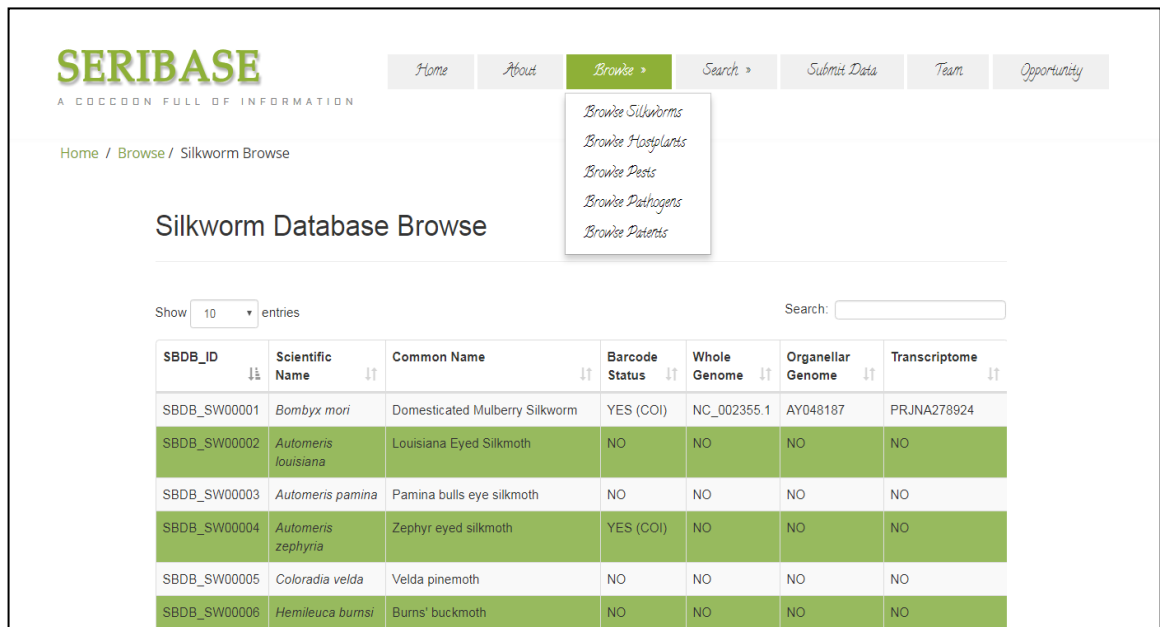


Figure 5.12 Browse Page of Seribase.

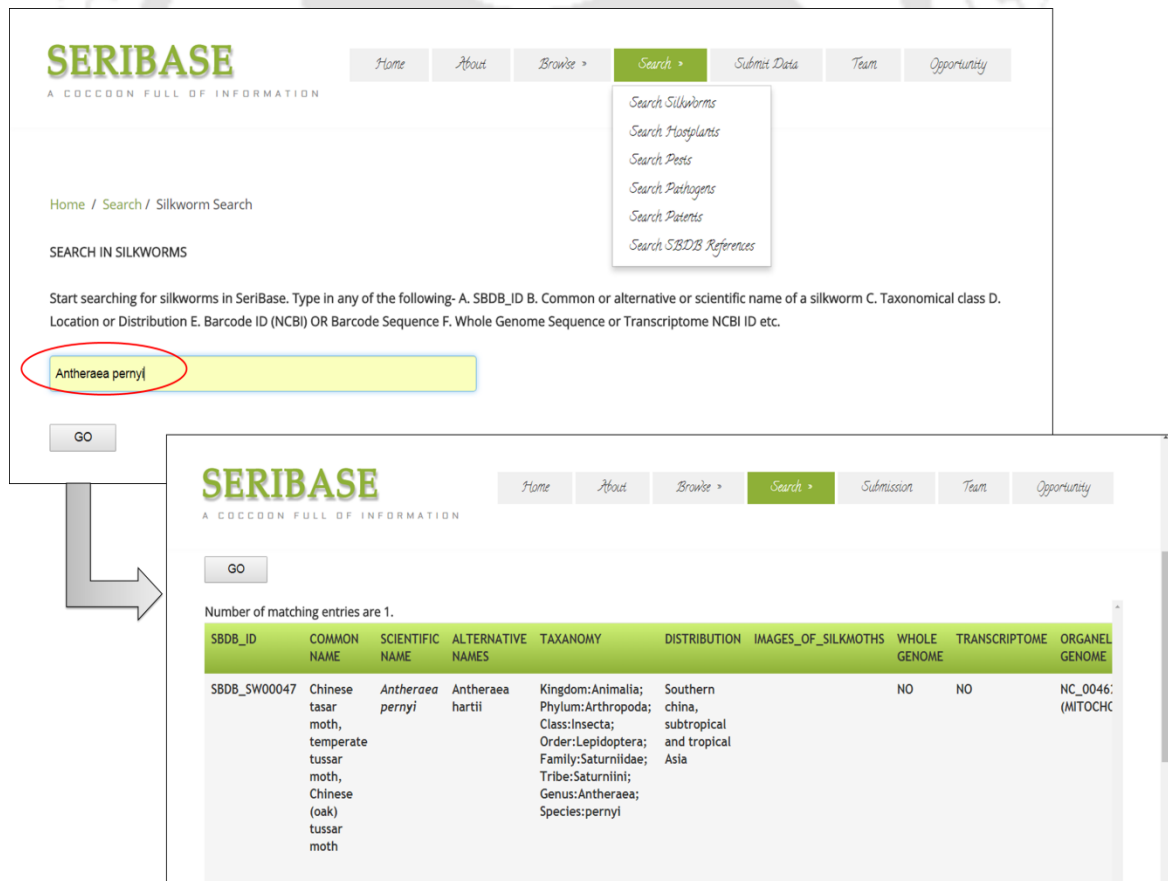


Figure 5.13 Search Page of Seribase and search results based on keyword “Antheraea pernyi”.

5.5 Data submission, download and updates

SeriPort

Users can submit the information on databases related to seri-informatics, seri-genomics, etc. which have not been included in SeriPort. A 'Submit Data' tab has been provided for users to facilitate data submission. The data submitted will be verified by the web portal administrator before it is added to SeriPort. A spreadsheet containing the available databases providing information on the genome, proteome and transcriptome of silkworms and their resources with references can be downloaded through 'Download' option. This option also allow users to download the guidelines available in Help page for using the portal. In addition, a list of sericulture related industries and organization can be downloaded from the 'Sericulture Industry' Page.

Seribase

Data can be submitted by users of the database using "submit data" form. Data related to all types of data namely, silkworms, silkworm hostplants, silkworm pests, silkworm pathogens, patents and literature can be submitted. The submitted data will be subjected to manual curation by the administrator followed by inclusion to database. Datasets will be updated periodically on the basis of new data availability while interface will be updated annually. Complete datasets available in Seribase can be downloaded as spreadsheets using "Download" page. "Help" page is also provided for navigation of the database.

Occasional updates will be done in SeriPort and Seribase whenever any new information becomes available.

5.6 Cross-referencing and pictorial data in Seribase

Data of Seribase has been cross-referenced with other external databases like BOLD, SilkBase, NCBI, etc. This will help in establishing better credibility to the data as well as helping in retrieval of further information on the same. Images of several silkworms have also been included in Seribase with sources.

5.7 Conclusions

SeriPort is a user-friendly portal that enlists and promotes seri-databases, thus helping scholars interested in data mining in the seri-related fields. Some databases are well-known while some are found rarely on the internet. SeriPort can play an important role in bringing focus to such databases. This portal also acts as a scaffold between the web resources helping diverse themes to come together under a single platform. This will pave a way for the entomological community to move towards construction of new databases in future.

On the other hand, Seribase intends to provide the current information status of wild seri-bioresources to the researchers working in entomological research. Options like browse and search will allow users for data mining of available data in the database. Based on data mining, we found that pest and pathogen information was limited as compared to silkworms and their host plants. We strongly encourage the researchers to submit relevant data to make it more enriched and updated. We hope that this database will facilitate more focus on Saturniid research which is one of the economic and ecologically important areas. In future, the search will be extended to additional sericigenous species and their resources along with more sequence and analytical information on the same.



Chapter 6

Summary and Future prospects

CHAPTER 6

Summary and Future prospects

The present study deals with the identification of few seri-bioresources of Assam, mitochondrial genome sequencing of wild silkmoths and seri-related database development.

6.1 Molecular characterization of selected Seri-bioresources of Assam through DNA Barcoding

In the present study, we have barcoded 18 indigenous seri-bioresources which is the first report on muga and eri silkworms, host plants, pests and pathogens. Different barcode markers (*COI*, *matK*, *rbcL*, *trnH-psbA*, *ITS*, *ITS2* and *16S rRNA*) were utilized for the identification of various species. We found that the selected barcode markers were efficient in discriminating the selected species based on sequence homology. Phylogenetic analysis using barcode regions of silkworms, pests, host plants, bacteria and fungal isolates revealed that selected barcode regions clustered the species into their respective groups (related species and family). The key findings of the study include identification of pests like *Odoiporus longicollis* and *Oecophylla smaragdina*, and fungal species like *Lasiodiplodia theobromae*, *Colletotrichum gloeosporioides*, *Neopestalotiopsis clavispora*, *Phomopsis phyllanthicola* which have been reported to infect silkworm host plants, thus affecting silkworm growth and silk production. All the sequences have been submitted to NCBI GenBank database with their corresponding accession numbers. We expect that this study will be useful for the researchers working in the exploration of silkworm resources and future phylogenetic and evolutionary studies.

6.2 Mitochondrial genome sequencing of *Antheraea assamensis*, a semi-domesticated silkworm endemic to Assam and North-eastern part of India

In the present work, we have reported the complete mitogenome of *A. assamensis*, an endemic silk moth species of North-eastern India, for the first-time using next-generation sequencing (NGS). This is the 12th representative organism of Saturniidae family with a completely sequenced mitogenome. Mitochondrial genome of *A. assamensis* is an AT rich (A+T content ~80.2%) circular molecule consisting of 37 genes and a control region. Its typical $tRNA^{Met}$ - $tRNA^{Ile}$ - $tRNA^{Gln}$ arrangement differed from ancestral insects ($tRNA^{Ile}$ - $tRNA^{Gln}$ - $tRNA^{Met}$). All the protein coding genes (PCGs) evolved under strong purifying selection with highest evolutionary rates observed for *atp8* gene while lowest for *cox1* gene. Comparatively higher value in *atp8* gene may be associated with energy requirement for silk production which was also found to be higher at inter-species level than at intra-species level. Transition-Transversion ratio suggests that *nad4l* gene may be the significant contributor for evolution to form *A. assamensis* within *Antheraea* genus. Codon usage analysis revealed high frequency of AT rich codons having A or T nucleotides at third codon position which can be used for future gene expression studies. All tRNAs folded into typical clover-leaf shaped secondary structures with a few exceptions in case of $tRNA^{Ser1}$ and $tRNA^{Tyr}$ where stable DHU and TΨC loop were absent. A significant number of mismatches (35) were found to spread over 19 tRNA structures. The control region of mitogenome contained a six bp (CTTAGA/G) deletion atypical of other *Antheraea* species and lacked tandem repeats. Phylogenetic position of *A. assamensis* was consistent with the traditional taxonomic classification of Saturniidae. The complete annotated mitogenome is available in GenBank (Accession No. KU379695). Our

study enriches the existing genomic data repertoire on *A. assamensis* and will facilitate a better understanding of species divergence patterns and other evolutionary facets in Lepidoptera.

6.3 Mitochondrial genome sequencing of *Cricula trifenestrata*, a wild silkmoth from Assam

This chapter describes decoding of the complete mitogenome of *C. trifenestrata*, a wild economically important silkmoth from Assam using NGS. This is the 13th representative organism of Saturniidae family with a completely sequenced mitogenome. The mitogenome of *Cricula* exhibited lowest A+T content as compared to the selected organisms of lepidoptera order. The typical gene arrangement ($tRNA^{Met}$ - $tRNA^{Ile}$ - $tRNA^{Gln}$) was different from the ancestral type. All protein coding genes exhibited strong purifying selection but *atp8* gene was comparatively higher than others, suggesting its role in energy requirement for silk production. This was concordant with gene-by-gene divergence studies which reflected higher proportion of informative sites as compared to other genes. All the tRNAs folded into typical clover leaf secondary structures except $tRNA^{Ile}$, $tRNA^{Arg}$ and $tRNA^{Ser1}$ genes which lacked TΨC loop, DHU loop and stable DHU arm, respectively. Presence of conserved regions like two conserved blocks (greater than 50 bp), poly T-stretch, microsatellite (TA/AT)_n like elements and poly-A stretch was observed in the control region like other Bombycoids. However, it differed from other organisms in terms of the presence of TTAGA motif and 33 bp tandem repeat units in the control region. The phylogenetic analysis revealed that *C. trifenestrata* is close to Saturniidae family members clustering in Saturniidae family. The annotated mitogenome sequence of *C. trifenestrata* has been submitted to GenBank under the accession number KY644697.

6.4 Development of Seri-related Web-resources

In the present work, we have created an HTML based web portal named **“SeriPort”** and a seri-bioresource database named **“Seribase”**. SeriPort acts as a common platform for all varieties of data offered on sericulture field over the internet. Currently, it comprises of 75 databases which are categorized into four classes- (i) silkworm DB (22 No.s), (ii) host plant DB (25 No.s), (iii) pest and pathogen DB (02 No.s), and (iv) combined DB (26 No.s). It also enlists sericulture related industries and organizations, genome, transcriptome and proteome data related to silkworms. The main features of SeriPort include a user-friendly and dynamic user interface, in-built search engine, data download and data submission options. Seribase is an integrated repository for more than 1600 wild seri-bioresources (501 species of wild silkmoths, 1023 host plants of the respective silkmoths, 35 disease pathogens and 90 pests of wild silkmoths) of economic or ecological significance. More than 300 patents have been included in the database till now. It comprises of datasets categorized into six different modules- “Silkworm”, “Host plants”, “Silkworm Pests”, “Silkworm Pathogens”, “Patents” and “Related Publications”. It has an easy-to-use web-interface equipped with various options like search, browse and download for easy retrieval of data. We expect that these two data resources will act as supportive platforms for the scientific community working in seri-related areas.

Future Prospects

- ✓ We have barcoded few seri-bioresources of Assam in the present study and there is a great scope of barcoding of many wild sericigenous insects as Assam and other regions of North-east India are rich in seridiversity. This will facilitate exploration and characterization of these resources for conservation of biodiversity and commercial utilization.
- ✓ We have decoded the mitochondrial genome of two economically and ecologically important silkmooths belonging to Saturniidae family. Further, studies on genome, proteome and metabolome of these organisms can be done for understanding functional part of the genome, understanding genetics of silk production, pest and disease management, interaction studies, etc.
- ✓ Mitochondrial genome studies of other wild silkmooths of Saturniidae family will facilitate elucidating evolutionary relationships among the silkmooths and development of novel molecular markers for molecular systematic studies.
- ✓ The users are encouraged to add their databases and seri-bioresource data in SeriPort and Seribase, respectively to make these a single window for the exploration of seri related web- and bio- resources. Development of these two web-resources has also brought to our attention that some fields like morphological, distribution, taxonomy, disease, pest and pathogen levels, etc. lack proper databases. This is a hindrance to better understanding of these organisms and their behaviors. So, more databases can be developed on such fields to explore this area of research.



Bibliography

- Abascal, F., Zardoya, R., & Telford, M. J. (2010). TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*, 38, W7-13.
- Ahmad, F. (1994). Biology of mango defoliator *C. trifenestrata* Helfer (Lepidoptera: Saturniidae). *Bangladesh Journal of Scientific Research*, 12(2), 151-156.
- Ahmed, S. A., & Rajan, R. K. (2011). Exploration of Vanya silk biodiversity in north eastern region of India: Sustainable livelihood and poverty alleviation. In *International conference on Management, Economics and Social sciences*. Bangkok, 485-489.
- Ahmed, S. A., Dutta, L. C., & Sarmah, M.C. (2012). Bio-Efficacy of some Insecticides against Leaf Eating Caterpillar *Cricula trifenestrata* Helfer (Lepidoptera: Saturniidae) Infesting Som *Persea bombycina* Kost. Plantation. *Academic Journal of Entomology*, 5(2), 94-98.
- Akashi, H. (1997). Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene*, 205(1), 269-278.
- Akhtar, J. (2006). Host pathogen interaction with reference to aggressiveness of the isolates of *Phomopsis vexans* to different genotypes of the brinjal. *Journal of Plant Disease Sciences*, 1, 164-167.
- Alikhan, N. F., Petty, N. K., Zakour, N. L. B., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, 12, 402.

- Anand, A. A., Vennison, S. J., Sankar, S. G., et al. (2010). Isolation and characterization of bacteria from the gut of *Bombyx mori* that degrade cellulose, xylan, pectin and starch and their impact on digestion. *Journal of Insect Science*, 10(1), 107.
- Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data* <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Aparicio, O., Geisberg, J. V., Sekinger, E., et al. (2005). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current Protocols in Molecular Biology*, 69, 21.3:21.3.1–21.3.33.
- Archak, S., Meduri, E., Kumar, P. S., & Nagaraju, J. (2007). InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Research*, 35(s1), D36-D39.
- Arfi, Y., Buée, M., Marchand, C., Levasseur, A., & Record, E. (2012). Multiple markers pyrosequencing reveals highly diverse and host-specific fungal communities on the mangrove trees *Avicennia marina* and *Rhizophora stylosa*. *FEMS Microbiology Ecology*, 79(2), 433-444.
- Arora, G. S., & Gupta, I. J. (1979). Taxonomic studies of some of the Indian non-mulberry silk moths (Lepidoptera: Saturniidae). *Memoirs of the Zoological survey of India*, 16, 1-63.
- Arumuganathan, K., & Earle, E. D. (1991). Nuclear DNA content of some important plant species. *Plant Molecular Biology Reporter*, 9(3), 208-218.
- Arunkumar, K. P., & Nagaraju, J. (2006). Unusually long palindromes are abundant in mitochondrial control regions of insects and nematodes. *Plos One*, 1(1), e110.

- Arunkumar, K. P., Metta, M., & Nagaraju, J. (2006). Molecular phylogeny of silkmoths reveals the origin of domesticated silkmoth, *Bombyx mori* from Chinese *Bombyx mandarina* and paternal inheritance of *Antheraea proylei* mitochondrial DNA. *Molecular Phylogenetics and Evolution*, 40, 419-427.
- Arunkumar, K. P., Tomar, A., Daimon, T., Shimada, T., & Nagaraju, J. (2008). WildSilkbase: an EST database of wild silkmoths. *BMC Genomics*, 9(1), 338.
- Astruc, C., Julien, J. F., Errard, C., & Lenoira, A. (2004). Phylogeny of ants (Formicidae) based on morphology and DNA sequence data. *Molecular Phylogenetics and Evolution*, 31, 880–893.
- Awasthi, A. K., Pradeep, A. R., Srivastava, P. P., et al. (2008). PCR detection of densovirus isolates in silkworm (*Bombyx mori*) from India and their nucleotide variability. *Indian Journal of Biotechnology*, 7, 56-60.
- Bafeel, S. O., Arif, I. A., Bakir, M. A., et al. (2011) Comparative evaluation of PCR success with universal primers of maturase K (*matK*) and ribulose-1, 5-bisphosphate carboxylase oxygenase large subunit (*rbcL*) for barcoding of some arid plants. *Plant Omics*, 4, 195-198.
- Ball, S. L., & Armstrong, K. F. (2006). DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Canadian Journal of Forest Research*, 36(2), 337-350.
- Bankevich, A. Nurk, S., Antipov, D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19, 455-477.
- Barr, N., Ruiz–Arce, R., Obregón, O., et al. (2013). Molecular diagnosis of populational variants of *Anthonomus grandis* (Coleoptera: Curculionidae) in North America. *Journal of Economic Entomology*, 106(1), 437-449.

- Beard, C. B., Hamm, D., & Collins, F. H. (1993). The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Molecular Biology*, 2(2), 103-124.
- Beccaloni, G., Scoble, M., Kitching, I., et al. (2003). The global Lepidoptera names index (LepIndex). *World Wide Web electronic publication*. [http://www. nhm. ac. uk/entomology/lepindex](http://www.nhm.ac.uk/entomology/lepindex) [accessed 13 August 2017].
- Beckenbach, A. T. (2011). Mitochondrial genome sequences of representatives of three families of scorpionflies (Order Mecoptera) and evolution in a major duplication of coding sequence. *Genome*, 54(5), 368-376.
- Belle, E. M., Piganeau, G., Gardner, M., & Eyre-Walker, A. (2005). An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene*, 355, 58-66.
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, 27, 573-580.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., et al. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112, 535–542.
- Bernt, M., Donath, A., Jühling, F., et al. (2013). MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, 69, 313-319.
- Bhuinya, T., Singh, P., & Mukherjee, S. K. (2010). Traditional uses of the Indian species of *Litsea* Lam. *Phytotaxonomy*, 10, 45-46.
- Bhuyan, P. M., Sandilya, S. P. & Gogoi, D. K. (2013). Phyllosphere Microflora of Muga Silkworm Host Plant *Persea bombycina* Kost (Som) Leaves in Jorhat

- District of Assam, India. *International Research Journal of Biological Sciences*, 2, 60-65.
- Bhuyan, P. M., Sandilya, S. P., Gogoi, D. K., Neog, K. & Sabramanian, S. (2014). Isolation and Characterization of gut bacteria of muga silkworm (*Antheraea assamensis* Helfer) collected from different locations of Assam. *Sericologia*, 54, 28-35.
- Bi, J., Luo, G., Wang, X., & Zhu, Z. (2014). The construction of the spatio-temporal database of the ancient Silk Road within Xinjiang province during the Han and Tang dynasties. In *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, 17(1), 012169.
- Bilimoria, K. Y., Stewart, A. K., Winchester, D. P., & Ko, C. Y. (2008). The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Annals of Surgical Oncology*, 15(3), 683-690.
- Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8), 1767-1780.
- Boore, J. L. (2000). The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In *Comparative genomics*. Springer Netherlands, 133-147.
- Brand, M. D. (1997). Regulation analysis of energy metabolism. *Journal of Experimental Biology*, 200(2), 193-202.
- Brimacombe, R., & Stiege, W. (1985). Structure and function of ribosomal RNA. *Biochemical Journal*, 229(1), 1-17.
- Bruni, I., De Mattia, F., Martellos, S., et al. (2012). DNA barcoding as an effective tool in improving a digital plant identification system: a case study for the area of Mt. Valerio, Trieste (NE Italy). *Plos One*, 7(9), e43256.

- Cameron, S. L. (2014). Insect mitochondrial genomics: implications for evolution and phylogeny. *Annual Review of Entomology*, 59, 95-117.
- Cameron, S. L., & Whiting, M. F. (2008). The complete mitochondrial genome of the tobacco hornworm, *Manduca sexta*, (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. *Gene*, 408(1), 112-123.
- Campbell, C. S., Wright, W. A., Cox, M., et al. (2005). Nuclear ribosomal DNA internal transcribed spacer 1 (ITS1) in *Picea* (Pinaceae): sequence divergence and structure. *Molecular Phylogenetics and Evolution*, 35(1), 165-185.
- Cantatore, P., Gadaleta, M. N., Roberti, M., Saccone, C., & Wilson, A. C. (1987). Duplication and remoulding of tRNA genes during the evolutionary rearrangement of mitochondrial genomes. *Nature*, 329(6142), 853-855.
- Cao, S. S., & Du, Y. Z. (2014). Characterization of the complete mitochondrial genome of *Chilo auricilius* and comparison with three other rice stem borers. *Gene*, 548(2), 270-276.
- Cao, Y. Q., Ma, C., Chen, J. Y. & Yang, D. R. (2012). The complete mitochondrial genomes of two ghost moths, *Thitarodes renzhiensis* and *Thitarodes yunnanensis*: the ancestral gene arrangement in Lepidoptera. *BMC Genomics*, 13(1), 276.
- Carapelli, A., Liò, P., Nardi, F., Van der Wath, E., & Frati, F. (2007). Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. *BMC Evolutionary Biology*, 7(2), S8.
- Case, R. J., Boucher, Y., Dahllöf, I., et al. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, 73(1), 278-288.

- CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106(31), 12794-12797.
- Cha, S. Y., Yoon, H. J., Lee, E. M., et al. (2007). The complete nucleotide sequence and gene organization of the mitochondrial genome of the bumblebee, *Bombus ignitus* (Hymenoptera: Apidae). *Gene*, 392(1), 206-220.
- Chai, H. N., & Du, Y. Z. (2012). The complete mitochondrial genome of the pink stem borer, *Sesamia inferens*, in comparison with four other noctuid moths. *International Journal of Molecular Sciences*, 13(8), 10236-10256.
- Chakravorty, R., Barah, A., Neog, K., Rahman, S. A. S., & Ghose, J. (2005). Package of practices for muga culture. In *Package of practices of Muga, Eri and Mulberry Sericulture for North Eastern Region of India*. Central Muga Eri Research & Training Institute, Central Silk Board, Lahdoigarh, Jorhat, 1-23.
- Chan, A. P., Crabtree, J., Zhao, Q., et al. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology*, 28(9), 951-956.
- Chase, M. W., Salamin, N., Wilkinson, M., et al. (2005). Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1462), 1889-1895.
- Chen, B., Teh, B. S., Sun, C., et al. (2016). Biodiversity and activity of the gut microbiota across the life history of the insect herbivore *Spodoptera littoralis*. *Scientific Reports*, 6, 29505.
- Chen, M. M., Li, Y., Chen, M., et al. (2014). Complete mitochondrial genome of the atlas moth, *Attacus atlas* (Lepidoptera: Saturniidae) and the phylogenetic relationship of Saturniidae species. *Gene*, 545 (1), 95-101.
- Chen, S., Yao, H., Han, J., et al. (2010). Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *Plos One*, 5(1), e8613.

- Choudhury, S. N. (1981). Diseases, pests and parasites. In Choudhury, S.N. (Ed.), *Muga Silk Industry, Directorate of Sericulture and Weaving*. Govt. of Assam, Guwahati (Assam), India, 74–81.
- Chowdappa, P., Reddy, G. S., Kumar, A., Rao, B. M., & Rawal, R. D. (2009). Morphological and molecular characterization of *Colletotrichum* species causing anthracnose of grape in India. *The Asian and Australas Journal of Plant Science and Biotechnology*, 3, 71-77.
- Clary, D. O., & Wolstenholme, D. R. (1985). The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, 22, 252-271.
- Coates, B. S., Sumerford, D. V., Hellmich, R. L., & Lewis, L. C. (2005). Partial mitochondrial genome sequences of *Ostrinia nubilalis* and *Ostrinia furnicalis*. *International Journal of Biological Sciences*, 1(1), 13.
- Cong, Q., & Grishin, N. V. (2016). The complete mitochondrial genome of *Lerema accius* and its phylogenetic implications. *PeerJ*, 4, e1546.
- Crozier, R. H., & Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics*, 133(1), 97-117.
- Da Silva, N. D. L., Maciel, M. R. W., Batistella, C. B., & Maciel Filho, R. (2006). Optimization of biodiesel production from castor oil. *Applied Biochemistry and Biotechnology*, 130(1-3), 405-414.
- Dai, L. S., Zhu, B. J., Qian, C., et al. (2016). The complete mitochondrial genome of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). *Mitochondrial DNA Part A*, 27(2), 1512-1513.

- Dai, L. S., Zhu, B. J., Zhao, Y., Zhang, C. F., & Liu, C. L. (2016). Comparative mitochondrial genome Analysis of *Eligma narcissus* and other Lepidopteran insects reveals conserved mitochondrial genome organization and phylogenetic relationships. *Scientific Reports*, 6, 26387.
- Dai, Q. Y., Gao, Q., Wu, C. S., et al. (2012). Phylogenetic reconstruction and DNA barcoding for closely related pine moth species (*Dendrolimus*) in China with multiple gene markers. *Plos One*, 7, e32544.
- Daimon, T., Kozaki, T., Niwa, R., et al. (2012). Precocious metamorphosis in the Juvenile Hormone–Deficient Mutant of the silkworm, *Bombyx mori*. *PLoS genetics*, 8(3), e1002486.
- Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8), 772-772.
- Das, R., Chutia, M., Das, K., & Jha, D. K. (2010). Factors affecting sporulation of *Pestalotiopsis disseminata* causing grey blight disease of *Persea bombycina* Kost., the primary food plant of muga silkworm. *Crop Protection*, 29(1), 963-968.
- Das, R., Das, K. & Rajan, R. K. (2014). Effect of plant extract on foliar blight disease of soalu caused by *Colletotrichum gloesporides* penz. *Indian Journal of Plant Sciences*, 3, 76-80.
- Das, R., Das, K. & Suryanarayana, N. (2005). Anthracnose disease in muga food plant, som (*Persea bombycina* Kost) in Assam. *Indian Journal of Sericulture*, 44, 134 –135.
- Das, R., Das, K., & Benchamin, K. V. (2003). Diseases of Muga food plant Som. *Indian Silk*, 41, 23-24.

- de Bruijn, M. H. (1983). *Drosophila melanogaster* mitochondrial DNA, a novel organization and genetic code. *Nature*, 304, 234–241.
- De Ré, F. C., Wallau, G. L., Robe, L. J., & Loreto, E. L. S. (2014). Characterization of the complete mitochondrial genome of flower-breeding *Drosophila incompta* (Diptera, Drosophilidae). *Genetica*, 142(6), 525-535.
- De, U. K., & Das, M. (2010). Economics of Sericulture in Assam: A Comparative Analysis of Three Cultivars. *South Asia Economic Journal*, 11, 309-336.
- Derks, M. F., Smit, S., Salis, L., et al. (2015). The genome of winter moth (*Operophtera brumata*) provides a genomic perspective on sexual dimorphism and phenology. *Genome Biology and Evolution*, 7(8), 2321-2332.
- DeSalle, R., Freedman, T., Prager, E. M., & Wilson, A. C. (1987). Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. *Journal of Molecular Evolution*, 26, 157-64.
- Devi, K., Ibsorani, L., Somen Singh, N., Singh, I., Dutta, K., & Chaoba Singh, K. (2011). Biodiversity of Sericigenous insects and their food plants in Manipur. *The Ecoscan*, 5(1&2), 65-68.
- Dewangan, S. K., Sahu, K. R., & Achari, K. V. (2011). Sericulture-A Tool of Eco-System Checking Through Tribal. *Journal of Environmental Research and Development*, 6, 165-173.
- Dillon, R. J., & Dillon, V. M. (2004). The gut bacteria of insects: nonpathogenic interactions. *Annual Reviews in Entomology*, 49(1), 71-92.
- Dong, Q., Schlueter, S. D., & Brendel, V. (2004). PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research*, 32(s1), D354-D359.

- Dowton, M., & Austin, A. D. (1999). Evolutionary dynamics of a mitochondrial rearrangement" hot spot" in the Hymenoptera. *Molecular Biology and Evolution*, 16, 298-309.
- Dowton, M., Cameron, S. L., Dowavic, J. I., Austin, A. D., & Whiting, M. F. (2009). Characterization of 67 mitochondrial tRNA gene rearrangements in the Hymenoptera suggests that mitochondrial tRNA gene position is selectively neutral. *Molecular Biology and Evolution*, 26(7), 1607-1617.
- Dowton, M., Castro, L. R., & Austin, A. D. (2002). Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: the examination of genome 'morphology'. *Invertebrate Systematics*, 16(3), 345-356.
- Duan, J., Li, R., Cheng, D., et al. (2009). SilkDB v2. 0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Research*, 38(s1), D453-D456.
- Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, 113.
- Eggert, L. S., Rasner, C. A., & Woodruff, D. S. (2002). The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. *Proceedings of the Royal Society of London B: Biological Sciences*, 269(1504), 1993-2006.
- Ehrenmann, F., Jacques-Gustave, A., Labbé, T., et al. (2014). Quercus portal: a web resource for genetics and genomics of oaks. In *Conference IUFRO: Genetics of Fagaceae*, Bordeaux.
- El-Sayed, A. M. (2014). The Pherobase: Extensive database of insect pheromones and semiochemicals. <http://www.pherobase.com>.

- Enan, M., Al-Deeb, M., Fawzy, N., & Amiri, K. (2012). DNA barcoding of *Ricinus communis* from different geographical origin by using chloroplast matK and internal transcribed spacers. *American Journal of Plant Sciences*, 3(9), 1304-1310.
- Engstrand, R. C., Cibrián Tovar, J., Cibrián-Jaramillo, A., & Kolokotronis, S. O. (2010). Genetic variation in avocado stem weevils *Copturus aguacatae* (Coleoptera: Curculionidae) in Mexico. *Mitochondrial DNA*, 21(s1), 38-43.
- Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, 40, D136–D143.
- Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, 338(6108), 758-767.
- Feng, W., Wang, X. Q., Zhou, W., Liu, G. Y., & Wan, Y. J. (2011). Isolation and characterization of lipase-producing bacteria in the intestine of the silkworm, *Bombyx mori*, reared on different forage. *Journal of Insect Science*, 11(1), 135.
- Floyd, R. M., Wilson, J. J., & Hebert, P. D. (2009). DNA barcodes and insect biodiversity. *Insect Biodiversity: Science and Society*, 417-431.
- Floyd, R., Abebe, E., Papert, A., & Blaxter, M. (2002). Molecular barcodes for soil nematode identification. *Molecular Ecology*, 11(4), 839-850.
- FlyBase Consortium. (2003). The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research*, 31(1), 172-175.
- Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3(5), 294-299.

- Frisvad, J. C., & Samson, R. A. (2004). Polyphasic taxonomy of *Penicillium* subgenus *Penicillium*. A guide to identification of food and air-borne terverticillate *Penicillia* and their mycotoxins. *Studies in Mycology*, 49, 1-174.
- Garesse, R. (1988). *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. *Genetics*, 118(4), 649-663.
- Gendler, K., Paulsen, T., & Napoli, C. (2008). ChromDB: the chromatin database. *Nucleic Acids Research*, 36(s1), D298-D302.
- Gisder, S., & Genersch, E. (2013). Molecular differentiation of *Nosema apis* and *Nosema ceranae* based on species-specific sequence differences in a protein coding gene. *Journal of Invertebrate Pathology*, 113(1), 1-6.
- Goldsmith, M., Shimada, T., & Abe, H. (2005). The genetics and genomics of the silkworm, *Bombyx mori*. *Annual Review of Entomology*, 50, 71-100.
- Gong, Y. J., Shi, B. C., Kang, Z. J., Zhang, F., & Wei, S. J. (2012). The complete mitochondrial genome of the oriental fruit moth *Grapholita molesta* (Busck) (Lepidoptera: Tortricidae). *Molecular Biology Reports*, 39(3), 2893-2900.
- Gong, Y. J., Wu, Q. L., & Wei, S. J. (2014). The first complete mitogenome for the superfamily Cossoidea of Lepidoptera: The seabuckthorn carpenter moth *Eogystia hippophaecolus*. *Mitochondrial DNA*, 25, 288-289.
- Goodstein, D. M., Shu, S., Howson, R., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(D1), D1178-D1186.
- Grant, J. R., & Stothard, P. (2008). The CGView Server: a comparative genomics tool for circular genomes. *Nucleic acids research*, 36(suppl_2), W181-W184.
- Gray, M. (2003). Diversity and evolution of mitochondrial RNA editing systems. *IUBMB life*, 55(4-5), 227-233.

- Gray, M. W., Burger, G., & Lang, B. F. (2001). The origin and early evolution of mitochondria. *Genome Biology*, 2(6), reviews1018.1–reviews1018.5.
- Group, C. P. W., Hollingsworth, P. M., Forrest, L. L., *et al.* (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, 106, 12794-12797.
- Gu, W., Song, J., Cao, Y., *et al.* (2013). Application of the ITS2 region for barcoding medicinal plants of Selaginellaceae in Pteridophyta. *Plos One*, 8(6), e67818.
- Gu, X. S., Ma, L., Wang, X., & Huang, G. H. (2016). Analysis on the complete mitochondrial genome of *Andraca theae* (Lepidoptera: Bombycoidea). *Journal of Insect Science*, 16(1), 105.
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696-704.
- Gusmão, D. S., Santos, A. V., Marini, D. C., *et al.* (2007). First isolation of microorganisms from the gut diverticulum of *Aedes aegypti* (Diptera: Culicidae): new perspectives for an insect-bacteria association. *Memorias do Instituto Oswaldo Cruz*, 102(8), 919-924.
- Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4), 968-971.
- Halim, M., Zuki, A. A., Mohammed, M. A., & Yaakop, S. (2017). DNA barcoding and relationships of eight ladybugs species (Coleoptera: Coccinellidae) that

- infesting several crops from Peninsular Malaysia. *Journal of Asia-Pacific Entomology*, 20(3), 814-820.
- Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, 95-98.
- Haloj, K., Kalita, M. K., Nath, R., & Devi, D. (2016). Characterization and pathogenicity assessment of gut-associated microbes of muga silkworm *Antheraea assamensis* Helfer (Lepidoptera: Saturniidae). *Journal of Invertebrate Pathology*, 138, 73-85.
- Handique, P. K., Das, S., Dutta, P., & Chakravorty, R. (2008). Leaf spot disease in clones of *Persea bombycina* caused by *Colletotrichum gloeosporioides*. *Journal of Mycology and Plant Pathology*, 38, 322-324.
- Hao, J. (2012). Complete sequence of the mitochondrial genome of the Japanese buff-tip moth, *Phalera flavesceus* (Lepidoptera: Notodontidae). *Genetics and Molecular Research*, 11(4), 4213-4225.
- Hao, J. J., Hao, J. S., Sun, X. Y., Zhang, L. L., & Yang, Q. (2014). The complete mitochondrial genomes of the Fenton's wood white, *Leptidea morsei*, and the lemon emigrant, *Catopsilia pomona*. *Journal of Insect Science*, 14(1), 130.
- He, N., Zhang, C., Qi, X., et al. (2013). Draft genome sequence of the mulberry tree *Morus notabilis*. *Nature communications*, 4, 2445.
- Hebert, P. D., Cywinska, A., & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, 270(1512), 313-321.
- Heise, W., Babik, W., Kubisz, D., & Kajtoch, Ł. (2015). A three-marker DNA barcoding approach for ecological studies of xerothermic plants and

- herbivorous insects from central Europe. *Botanical Journal of the Linnean Society*, 177(4), 576-592.
- Hillis, D. M., & Dixon, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. *The Quarterly Review of Biology*, 66(4), 411-453.
- Hirsch, J., Reineke, A., & Sprick, P. (2010). DNA barcoding for weevil larvae: a molecular diagnostic method for species identification of weevils in crop protection. *Julius-Kühn-Archiv*, 428, 398-399.
- Holliday, P. (1995). *Fungus diseases of tropical crops*. Dover Publications Inc., New York, 607.
- Hollingsworth, P. M., Graham, S. W., & Little, D. P. (2011). Choosing and using a plant DNA barcode. *Plos One*, 6(5), e19254.
- Hong, G., Jiang, S., Yu, M., *et al.* (2009). The complete nucleotide sequence of the mitochondrial genome of the cabbage butterfly, *Artogeia melete* (Lepidoptera: Pieridae). *Acta Biochim Biophys Sin*, 41(6), 446-455.
- Hong, M. Y., Lee, E. M., Jo, Y. H., *et al.* (2008). Complete nucleotide sequence and organization of the mitogenome of the silk moth *Caligula boisduvalii* (Lepidoptera: Saturniidae) and comparison with other lepidopteran insects. *Gene*, 413(1), 49-57.
- Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, 9, 868-877.
- Hwang, J. S., Lee, J. S., Goo, T. W., *et al.* (1999). The comparative molecular study between Bombycidae and Saturniidae based on mtDNA RFLP and cytochrome oxidase I gene sequences: Implication for molecular evolution. *Zeitschrift für Naturforschung C*, 54(7-8), 587-594.

- Imoulan, A., Wu, H. J., Lu, W. L., et al. (2016). *Beauveria medogensis* sp. nov., a new fungus of the entomopathogenic genus from China. *Journal of Invertebrate Pathology*, 139, 74-81.
- Inoue, S., Kanda, T., Imamura, M., et al. (2005). A fibroin secretion-deficient silkworm mutant, Nd-sD, provides an efficient system for producing recombinant proteins. *Insect Biochemistry and Molecular Biology*, 35(1), 51-59.
- Ironside, J. E. (2013). Diversity and recombination of dispersed ribosomal DNA and protein coding genes in microsporidia. *Plos One*, 8(2), e55878.
- Ishii, K., Hamamoto, H., & Sekimizu, K. (2015). Studies of host-pathogen interactions and immune-related drug development using the silkworm: interdisciplinary immunology, microbiology, and pharmacology studies. *Drug Discoveries & Therapeutics*, 9(4), 238-246.
- Iskender, N. A., Algur, O. F., Aksu, Y., & Saral, A. (2017). Isolation, identification and characterization of biotechnologically important bacteria from microflora of *Dryocosmus kuriphilus* Yasumatsu (Hymenoptera: Cynipidae). *Biotechnology & Biotechnological Equipment*, 31(3), 505-510.
- Jalali, S. K., Ojha, R., & Venkatesan, T. (2015). DNA barcoding for Identification of agriculturally important insects. In *New Horizons in Insect Science: Towards Sustainable Pest Management*. Springer India, 13-23.
- Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, 45(9), 2761-2764.

- Jeong, S. Y., Park, J. S., Kim, S. S., & Kim, I. (2016). Complete mitochondrial genome of the gelechioid *Stathmopoda auriferella* (Lepidoptera: Stathmopodidae). *Mitochondrial DNA Part B*, 1(1), 522-524.
- Jiang, S. T., Hong, G. Y., Yu, M., et al. (2009). Characterization of the complete mitochondrial genome of the giant silkworm moth, *Eriogyna pyretorum* (Lepidoptera: Saturniidae). *International Journal of Biological Sciences*, 5(4), 351-365.
- Jin, J., Zhang, H., Kong, L., Gao, G., & Luo, J. (2014). PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Research*, 42(D1), D1182-D1187.
- Joshi, R. P., & Raja, I. A. (2016). Polymerase chain reaction based detection of Grasserie virus, BmNPV in Silkworm, *Bombyx mori*. *Bioscience Biotechnology Research Communications*, 9(3), 471-474.
- Jühling, F., Pütz, J., Bernt, M., et al. (2012). Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Research*, 40(7), 2833-2845.
- Junqueira, A. C. Lessinger, A. C., Torres, T. T., et al. (2004). The mitochondrial genome of the blowfly *Chrysomya chloropyga* (Diptera: Calliphoridae). *Gene*, 339, 7-15.
- Kaito, C., Kurokawa, K., Matsumoto, Y., et al. (2005). Silkworm pathogenic bacteria infection model for identification of novel virulence genes. *Molecular Microbiology*, 56(4), 934-944.
- Kajiwara, H., Nakane, K., Piyang, J., et al. (2006). Draft of silkworm proteome database. *Journal of Electrophoresis*, 50(3,4), 39-41.

- Kakati, L. N., & Chutia, B. C. (2009). Diversity and ecology of wild sericigenous insects in Nagaland, India. *Tropical Ecology*, 50(1), 137-146.
- Kalita, T., & Dutta, K. (2014). Biodiversity of Sericigenous insects in Assam and their role in employment generation. *Journal of Entomology and Zoology Studies*, 2(5), 119-125.
- Karthika, P., Krishnaveni, N., Vadivalagan, C., et al. (2016). DNA barcoding and evolutionary lineage of 15 insect pests of horticultural crops in South India. *Karbala International Journal of Modern Science*, 2, 156-168.
- Kasoju, N., Bhonde, R. R., & Bora, U. (2009). Fabrication of a novel micro–nano fibrous nonwoven scaffold with *Antheraea assama* silk fibroin for use in tissue engineering. *Materials Letters*, 63, 2466-2469.
- Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4), 772–780.
- Kelley, S. T., & Dobler, S. (2011). Comparative analysis of microbial diversity in *Longitarsus* flea beetles (Coleoptera: Chrysomelidae). *Genetica*, 139(5), 541-550.
- Kim, C., Kim, K., Park, D., et al. (2010). An integrated database for the enhanced identification of silkworm gene resources. *Bioinformatics*, 4(10), 436.
- Kim, J. S., Park, J. S., Kim, M. J., et al. (2012). Complete nucleotide sequence and organization of the mitochondrial genome of eri-silkworm, *Samia cynthia ricini* (Lepidoptera: Saturniidae). *Journal of Asia-Pacific Entomology*, 15(1), 162-173.
- Kim, M. I., Baek, J. Y., Kim, M. J., et al. (2009). Complete nucleotide sequence and organization of the mitogenome of the red-spotted apollo butterfly,

- Parnassius bremeri* (Lepidoptera: Papilionidae) and comparison with other lepidopteran insects. *Molecules and Cells*, 28(4), 347-363.
- Kim, M. J., & Kim, I. (2016). Complete mitochondrial genome of the Mormon metalmark butterfly, *Apodemia mormo* (Lepidoptera: Riodinidae). *Mitochondrial DNA Part A*, 27(2), 789-791.
- Kim, M. J., Choi, S. W., & Kim, I. (2013). Complete mitochondrial genome of the larch hawk moth, *Sphinx morio* (Lepidoptera: Sphingidae). *Mitochondrial DNA*, 24, 622-624.
- Kim, M. J., Jeong, H. C., Kim, S. R., & Kim, I. (2011). Complete mitochondrial genome of the nerippe fritillary butterfly, *Argynnis nerippe* (Lepidoptera: Nymphalidae). *Mitochondrial DNA*, 22(4), 86-88.
- Kim, M. J., Kang, A. R., Jeong, H. C., Kim, K. G., & Kim, I. (2011). Reconstructing intraordinal relationships in Lepidoptera using mitochondrial genome data with the description of two newly sequenced lycaenids, *Spindasis takanonis* and *Protantigius superans* (Lepidoptera: Lycaenidae). *Molecular Phylogenetics and Evolution*, 61, 436-445.
- Kim, M. J., Wan, X., Kim, K., Hwang, J. S., & Kim, I. (2010). Complete nucleotide sequence and organization of the mitogenome of endangered *Eumenis autonoe* (Lepidoptera: Nymphalidae). *African Journal of Biotechnology*, 9, 735-754.
- Kim, M. J., Wang, A. R., Park, J. S., & Kim, I. (2014). Complete mitochondrial genomes of five skippers (Lepidoptera: Hesperiidae) and phylogenetic reconstruction of Lepidoptera. *Gene*, 549(1), 97-112.

- Kim, S. R., Kim, M. I., Hong, M. Y., *et al.* (2009). The complete mitogenome sequence of the Japanese oak silkmoth, *Antheraea yamamai* (Lepidoptera: Saturniidae). *Molecular Biology Reports*, 36(7), 1871-1880.
- Knight, R. D., Freeland, S. J., & Landweber, L. F. (2001). A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biology*, 2(4), research0010-1.
- Kong, W., & Yang, J. (2015). The complete mitochondrial genome of *Rondotia menciaana* (Lepidoptera: Bombycidae). *Journal of Insect Science*, 15(1), 48.
- Korsten, L., Sanders, G. M., & Grosse-Weischede, E. (1994). Isolation and pathogenicity of avocado post-harvest pathogens from Westfalia and other avocado producing areas. *South African Avocado Growers' Association Yearbook*, 17, 46-48.
- Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1), D68-D73.
- Kress, W. J., & Erickson, D. L. (2007). A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *Plos One*, 2(6), e508.
- Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23), 8369-8374.
- Krishnan, R. R., Sumathy, R., Bindroo, B. B., & Naik, V. G. (2014). MulSatDB: a first online database for mulberry microsatellites. *Trees*, 28(6), 1793-1799.

- Kroemer, G., Dallaporta, B., & Resche-Rigon, M. (1998). The mitochondrial death/life regulator in apoptosis and necrosis. *Annual Review of Physiology*, 60(1), 619-642.
- Kuhad, R. C., Gupta, R., & Singh, A. (2011). Microbial Cellulases and Their Industrial Applications. *Enzyme Research*, 2011, 280696.
- Kühlbrandt, W. (2015). Structure and function of mitochondrial membrane protein complexes. *BMC Biology*, 13(1), 89.
- Kumar, A., Chetia, H., Sharma, S., et al. (2015). Curcumin resource database. *Database*, 2015, bav070.
- Kundu, A., Sarkar, D., Bhattacharjee, A., et al. (2011). A simple ethanol wash of the tissue homogenates recovers high-quality genomic DNA from *Corchorus* species characterized by highly acidic and proteinaceous mucilages. *Electronic Journal of Biotechnology*, 14, 10-11.
- Kundu, S. C. (2012). Biomat _dBase: a database on biomaterials. *Open Tissue Engineering and Regenerative Medicine Journal*, 5, 59-67.
- Kyei-Poku, G., & Sokolova, Y. Y. (2017). The microsporidium *Nosema disstriae* (Thomson 1959): Fine structure and phylogenetic position within the *N. bombycis* clade. *Journal of Invertebrate Pathology*, 143, 90-103.
- Lahaye, R., Van der Bank, M., Bogarin, D., et al. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences*, 105(8), 2923-2928.
- Lalhmingliani, E., Gurusubramanian, G., Kumar, N. S., et al. (2015). Ethnomedicinal Uses of Host Plants of Wild Silk Moths in Mizoram. *Journal of Environmental and Social Sciences*, 2(1), 114.

- Lalhmingliani, E., Gurusubramanian, G., Lalfelpuii, R., *et al.* (2013). Wild silk moth (Lepidoptera: Saturniidae) of Mizoram University campus, Aizawl, Mizoram, Northeast India. *Bioresources and Traditional Knowledge of Northeast India*, 223-228.
- Lane, N. (2006). *Power, sex, suicide: mitochondria and the meaning of life*. Oxford University Press.
- Lavrov, D. V., Boore, J. L., & Brown, W. M. (2002). Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss. *Molecular Biology and Evolution*, 19(2), 163-169.
- Lavrov, D. V., Brown, W. M., & Boore, J. L. (2000). A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede *Lithobius forficatus*. *Proceedings of the National Academy of Sciences*, 97(25), 13738-13742.
- Lawrie, D. S., Messer, P. W., Hershberg, R., & Petrov, D. A. (2013). Strong purifying selection at synonymous sites in *D. melanogaster*. *PLoS Genetics*, 9(5), e1003527.
- Lee, E. S., Shin, K. S., Kim, M. S., *et al.* (2006). The mitochondrial genome of the smaller tea tortrix *Adoxophyes honmai* (Lepidoptera: Tortricidae). *Gene*, 373, 52-57.
- Lee, S. B., Milgroom, M. G., & Taylor, J. W. (1988). A rapid, high yield mini-prep method for isolation of total genomic DNA from fungi. *Fungal Genetics Reports*, 35, 23-24.
- Lemaire, C., & Minet, J. (1998). The Bombycoidea and their relatives. *Part*, 35, 321-353.

- Letunic, I., & Bork, P. (2006). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1), 127-128.
- Levinson, G., & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution*, 4(3), 203-221.
- Li, F. B., Wang, W., Zhang, H. X., et al. (2015). Complete mitochondrial genome of the oriental armyworm *Mythimna separata* (Walker) (Lepidoptera: Noctuidae). *Mitochondrial DNA*, 26(6), 881-882.
- Li, T., Pan, G.-Q., Vossbrinck, C. R., et al. (2017). SilkPathDB: a comprehensive resource for the study of silkworm pathogens. *Database*, 2017, bax001.
- Li, T., Qi, X., Zeng, Q., Xiang, Z., & He, N. (2014). MorusDB: a resource for mulberry genomics and genome biology. *Database*, 2014, bau054.
- Li, Y. P., Song, W., Shi, S. L., et al. (2010). Mitochondrial genome nucleotide substitution pattern between domesticated silkworm, *Bombyx mori*, and its wild ancestors, Chinese *Bombyx mandarina* and Japanese *Bombyx mandarina*. *Genetics and Molecular Biology*, 33(1), 186-189.
- Li, Y., Wang, G., Tian, J., et al. (2012). Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA sequencing. *Plos One*, 7(8), e43713.
- Liang, X., Fu, Y., Tong, L., & Liu, H. (2014). Microbial shifts of the silkworm larval gut in response to lettuce leaf feeding. *Applied Microbiology and Biotechnology*, 98, 3769-3776.
- Liang, Y., Yesuf, J., Schmitt, S., Bender, K., & Bozzola, J. (2009). Study of cellulases from a newly isolated thermophilic and cellulolytic *Brevibacillus* sp. strain JXL. *Journal of Industrial Microbiology & Biotechnology*, 36, 961-970.

- Liao, F., Wang, L., Wu, S., *et al.* (2010). The complete mitochondrial genome of the fall webworm, *Hyphantria cunea* (Lepidoptera: Arctiidae). *International Journal of Biological Sciences*, 6(2), 172-186.
- Links, M. G., Dumonceaux, T. J., Hemmingsen, S. M., & Hill, J. E. (2012). The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. *Plos One*, 7(11), e49755.
- Liu, Q. N., Chai, X. Y., Bian, D. D., *et al.* (2016). The complete mitochondrial genome of fall armyworm *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *Genes & Genomics*, 38(2), 205-216.
- Liu, Q. N., Xin, Z. Z., Bian, D. D., *et al.* (2016). The first complete mitochondrial genome for the subfamily Limacodidae and implications for the higher phylogeny of Lepidoptera. *Scientific Reports*, 6, 35878.
- Liu, Q. N., Xin, Z. Z., Zhu, X. Y. *et al.* (2017). A transfer RNA gene rearrangement in the lepidopteran mitochondrial genome. *Biochemical and Biophysical Research Communications*, 489, 149-154.
- Liu, Q. N., Zhu, B. J., Dai, L. S., Wei, G. Q., & Liu, C. L. (2012). The complete mitochondrial genome of the wild silkworm moth, *Actias selene*. *Gene*, 505(2), 291-299.
- Liu, S., Xue, D., Cheng, R., & Han, H. (2014). The complete mitogenome of *Apocheima cinerarius* (Lepidoptera: Geometridae: Ennominae) and comparison with that of other lepidopteran insects. *Gene*, 547(1), 136-144.
- Liu, Y. Q., Li, Y. P., Wang, H., *et al.* (2012). The complete mitochondrial genome of the wild type of *Antheraea pernyi* (Lepidoptera: Saturniidae). *Annals of the Entomological Society of America*, 105(3), 498-505.

- Liu, Y., Li, Y., Pan, M., *et al.* (2008). The complete mitochondrial genome of the Chinese oak silkmoth, *Antheraea pernyi* (Lepidoptera: Saturniidae). *Acta Biochimica et Biophysica Sinica*, 40(8), 693-703.
- Liu, Z. F., Ci, X. Q., Li, L., *et al.* (2017). DNA barcoding evaluation and implications for phylogenetic relationships in Lauraceae from China. *Plos One*, 12(4), e0175788.
- Lu, H. F., Su, T. J., Luo, A. R., Zhu, C. D., & Wu, C. S. (2013). Characterization of the complete mitochondrion genome of diurnal moth *Amata emma* (Butler) (Lepidoptera: Erebidae) and its phylogenetic implications. *Plos One*, 8(9), e72410.
- Luo, K., Chen, S., Chen, K., *et al.* (2010). Assessment of candidate plant DNA barcodes using the Rutaceae family. *Science China Life Sciences*, 53(6), 701-708.
- Ma, R. J., Wang, C. Y., Liu, Y. W., *et al.* (2014). Identification and characterization of a novel endoglucanase (CMCase) isolated from the larval gut of *Bombyx mori*. *Journal of Asia-Pacific Entomology*, 17(1), 67-71.
- Mandal, S. D., Chhakchhuak, L., Gurusubramanian, G., & Kumar, N. S. (2014). Mitochondrial markers for identification and phylogenetic studies in insects– A Review. *DNA Barcodes*, 2, 1-9.
- Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133-141.
- Margam, V. M., Coates, B. S., Hellmich, R. L., *et al.* (2011). Mitochondrial genome sequence and expression profiling for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). *Plos One*, 6, e16444.

- Marie-Laure, R., Emerson, L., & Robertson, B. J. (2014). *The Johns Hopkins Guide to Digital Media*. Johns Hopkins University Press, Baltimore.
- Meiklejohn, C. D., Montooth, K. L., & Rand, D. M. (2007). Positive and negative selection on the mitochondrial genome. *Trends in Genetics*, 23(6), 259-263.
- Mirande, M. (2007). The ins and outs of tRNA transport. *EMBO Reports*, 8(6), 547-549.
- Mishra, N., Hazarika, N. C., Narain, K., & Mahanta, J. (2003). Nutritive value of non-mulberry and mulberry silkworm pupae and consumption pattern in Assam, India. *Nutrition Research*, 23(10), 1303-1311.
- Mita, K., Kasahara, M., Sasaki, S., et al. (2004). The genome sequence of silkworm, *Bombyx mori*. *DNA Research*, 11(1), 27-35.
- Mita, K., Morimyo, M., Okano, K., et al. (2003). The construction of an EST database for *Bombyx mori* and its application. *Proceedings of the National Academy of Sciences*, 100(24), 14121-14126.
- Moritz, C., & Brown, W. M. (1987). Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. *Proceedings of the National Academy of Sciences*, 84(20), 7183-7187.
- Nagarajan, N., & Navajas-Pérez, R. (2014). Papaya Repeat Database. In *Plant Genetics and Genomics: Crops and Models. Genetics and Genomics of Papaya* (eds. Ming R., & Moore P.H.). Springer, New York, 225–240.
- Nässig, W. A., Lampe, R. E., Kager, S., Buchsbaum, U., & Brüggemeier, F. (1996). In *The Saturniidae of Sumatra (Lepidoptera)*. Heterocera Sumatrana Society, Gottingen, 10, 3–110.
- Negi, B. K., & Sengupta, A. K., (1993). Pest status of Insects Associated with *Litsea polyantha* Juss and *Machilus bombycina* King, the primary host plants

- of Muga Silkworm, *Antheraea assama* Ww Zool. *Journal of Pure and Applied Zoology*, 3(3), 189-193.
- Neog, K., Gogoi, S. N., & Chakravorty, R. (2005). Present status and constraints of muga silkworm host plant germplasm conservation. In *Proceedings of the Workshop on Strategies for Maintenance of Non-Mulberry Silkworm and Host Plant Germplasm*. Central Muga Eri Research & Training Institute, Lahdoigarh, Jorhat, Assam, India, 11, 1-10.
- Newmaster, S. G., Fazekas, A. J., & Ragupathy, S. (2006). DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Botany*, 84(3), 335-341.
- Nindhia, T. G. T., Knejzlik, Z., Ruml, T., & Nindhia, T. S. (2014). Tensile properties and biocompatibility of Indonesian wild silk *Cricula trifenestrata*: A Preliminary Study. *Journal of Medical and Bioengineering*, 3(2).
- O'Donnell, K., Sutton, D. A., Rinaldi, M. G., *et al.* (2010). Internet-accessible DNA sequence database for identifying fusaria from human and animal infections. *Journal of Clinical Microbiology*, 48(10), 3708-3718.
- Ojha, R., Jalali, S. K., Ali, T. M., *et al.* (2014). DNA barcoding of Indian ant species based on *cox1* gene. *Indian Journal of Biotechnology*, 13, 165-171.
- Oliveira, D. C., Raychoudhury, R., Lavrov, D. V., & Werren, J. H. (2008). Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (Hymenoptera: Pteromalidae). *Molecular Biology and Evolution*, 25(10), 2167-2180.
- Osanai-Futahashi, M., Suetsugu, Y., Mita, K., & Fujiwara, H. (2008). Genome-wide screening and characterization of transposable elements and their

- distribution analysis in the silkworm, *Bombyx mori*. *Insect Biochemistry and Molecular Biology*, 38(12), 1046-1057.
- Packer, L., Gixti, J. C., Roughley, R. E., & Hanner, R. (2009). The status of taxonomy in Canada and the impact of DNA barcoding. *Canadian Journal of Zoology*, 87(12), 1097-1110.
- Padial, J. M., & De La Riva, I. (2007). Integrative taxonomists should use and produce DNA barcodes. *Zootaxa*, 1586(1), 67-68.
- Padial, J. M., Castroviejo-Fisher, S., Köhler, J., et al. (2009). Deciphering the products of evolution at the species level: the need for an integrative taxonomy. *Zoologica Scripta*, 38(4), 431-447.
- Pal, S., & Medda, P. S. (2006). Occurrence of wild silk moth, *Cricula trifenestrata* (Helfer) on cashew, *Anacardium occidentale* L. under Terai conditions of West Bengal. *Orissa Journal of Horticulture*, 34(1), 117.
- Palidwor, G. A., Perkins, T. J., & Xia, X. (2010). A general model of codon bias due to GC mutational bias. *Plos One*, 5(10), e13431.
- Paliwal, A. K., & Paliwal, D. P. (2015). Occurrence of Diseases on Muga Silkworm Host Plant *Persea bombycina* Kost. in District Bageshwar, Uttarakhand-A Survey Report. *Journal of Functional and Environmental Botany*, 5(2), 132-136.
- Palmer, J. D., & Herbon, L. A. (1988). Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, 28(1), 87-97.
- Pan, M., Yu, Q., Xia, Y., et al. (2008). Characterization of mitochondrial genome of Chinese wild mulberry silkworm, *Bombyx mandarina* (Lepidoptera: Bombycidae). *Science in China Series C: Life Sciences*, 51(8), 693-701.

- Papanicolaou, A., Gebauer-Jung, S., Blaxter, M. L., Owen McMillan, W., & Jiggins, C. D. (2008). ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Research*, 36(s1), D582-D587.
- Park, J. S., Kim, M. J., & Kim, I. (2016). The complete mitochondrial genome of the moon moth, *Actias aliena* (Lepidoptera: Saturniidae). *Mitochondrial DNA Part A*, 27(1), 149-150.
- Park, J. S., Kim, M. J., Jeong, S. Y., Kim, S. S., & Kim, I. (2016). Complete mitochondrial genomes of two gelechioids, *Mesophleps albilinella* and *Dichomeris ustalella* (Lepidoptera: Gelechiidae), with a description of gene rearrangement in Lepidoptera. *Current Genetics*, 62(4), 809-826.
- Park, J. S., Kim, S. S., Kim, K. Y., & Kim, I. (2016). Complete mitochondrial genome of Suzuki's Promolactis moth *Promalactis suzukiella* (Lepidoptera: Oecophoridae). *Mitochondrial DNA Part A*, 27(3), 2093-2094.
- Parveen, I., Gafner, S., Tehen, N., Murch, S. J., & Khan, I. A. (2016). DNA barcoding for the identification of botanicals in herbal medicine and dietary supplements: strengths and limitations. *Planta Medica*, 82(14), 1225-1235.
- Peigler, R. S., & Naumann, S. (2003). *A revision of the silkmoth genus Samia*. University of the Incarnate Word, San Antonio, Texas.
- Peng, X. Y., Zhou, P., Qiang, Y., & Qian, Z. Q. (2016). Characterization of the complete mitochondrial genome of *Bombyx huttoni* (Lepidoptera: Bombycidae). *Mitochondrial DNA Part A*, 27(6), 4112-4113.
- Pennisi, E. (2007). Wanted: A barcode for plants. *Science*, 318, 190 –191.
- Prasad, M. D., Muthulakshmi, M., Arunkumar, K. P., et al. (2005). SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Research*, 33(s1), D403-D406.

- Price, P. W., Denno, R. F., Eubanks, M. D., Finke, D. L., & Kaplan, I. (2011). *Insect ecology: behavior, populations and communities*. Cambridge University Press.
- Proost, S., Van Bel, M., Vaneechoutte, D., et al. (2015). PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research*, 43(D1), D974-D981.
- Purty, R., & Chatterjee, S. (2016). DNA Barcoding: An effective technique in molecular taxonomy. *Austin Journal of Biotechnology & Bioengineering*, 3(1), 1059.
- Qanungo, K. R., Kundu, S. C., Mullins, J. I., & Ghosh, A. K. (2002). Molecular cloning and characterization of *Antheraea mylitta* cytoplasmic polyhedrosis virus genome segment 9. *Journal of General Virology*, 83(6), 1483-1491.
- Qin, J., Zhang, Y., Zhou, X., et al. (2015). Mitochondrial phylogenomics and genetic relationships of closely related pine moth (Lasiocampidae: *Dendrolimus*) species in China, using whole mitochondrial genomes. *BMC Genomics*, 16(1), 428.
- Quintana-Obregón, E. A., Plascencia-Jatomea, M., Burgos-Hernández, A., Figueroa-Lopez, P. & Cortez-Rocha, M. O. (2013). Isolation and identification of fungi from leaves infected with false mildew on safflower crops in the Yaqui Valley, Mexico. *REVISTA MEXICANA DE MICOLOGÍA*, 37, 19-27.
- Rahman, A., Tanti, B., Sarma, G. C., & Kalita, J. (2012). Genetic diversity of *Persea bombycina* from goalpara district of Assam, India. *Advances in Bioscience and Biotechnology*, 3, 20-24.
- Rambaut, A., Suchard, M., Xie, D., & Drummond, A. (2014). Tracer v1. 6 <http://tree.bio.ed.ac.uk/software/tracer/>.

- Ramesh, G. K., Thangamalar, A., Muthuswami, M., & Subramanian, S. (2010). Characterisation of gram negative bacterial isolates from guts of few multivoltine silkworm breeds. *Karnataka Journal of Agricultural Sciences*, 22, 517-518.
- Ramírez-Ríos, V., Franco-Sierra, N. D., Alvarez, J. C., Saldamando-Benjumea, C. I., & Villanueva-Mejía, D. F. (2016). Mitochondrial genome characterization of *Tecia solanivora* (Lepidoptera: Gelechiidae) and its phylogenetic relationship with other lepidopteran insects. *Gene*, 581(2), 107-116.
- Rath, S. S., & Sinha, B. R. R. P. (2005). Parasitization of fifth instar tasar silkworm, *Antheraea mylitta*, by the uzi fly, *Blepharipa zebina*; a host–parasitoid interaction and its effect on host's nutritional parameters and parasitoid development. *Journal of Invertebrate Pathology*, 88(1), 70-78.
- Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Molecular Ecology Resources*, 7(3), 355-364.
- Rawlings, N. D., Waller, M., Barrett, A. J., & Bateman, A. (2014). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 42(D1), D503-D509.
- Razmaraii, N., Sadegh-Eteghad, S., Babaei, H., Paykari, H., Esmaeilnia, K., & Frogly, L. (2013). Molecular identification of *Nosema* species in East Azerbaijan province, Iran. *Archives of Razi Institute*, 68(1), 23-27.
- Rebijith, K. B., Asokan, R., Kumar, N. K., Krishna, V., Chaitanya, B. N., & Ramamurthy, V. V. (2013). DNA barcoding and elucidation of cryptic aphid species (Hemiptera: Aphididae) in India. *Bulletin of entomological research*, 103(5), 601-610.

- Ren, B. Q., Xiang, X. G., & Chen, Z. D. (2010). Species identification of *Alnus* (Betulaceae) using nrDNA and cpDNA genetic markers. *Molecular Ecology Resources*, 10(4), 594-605.
- Reyes, A., Gissi, C., Pesole, G., & Saccone, C. (1998). Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution*, 15(8), 957-966.
- Robinson, G. S., Ackery, P. R., Kitching, I. J., Beccaloni, G. W., & Hernández, L. M. (2010). HOSTS—a Database of the World's Lepidopteran Hostplants. *Natural History Museum, London*.
- Ronquist, F., Teslenko, M., Van Der Mark, P., et al. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3), 539-542.
- Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., & Reyes, A. (1999). Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene*, 238(1), 195-209.
- Saikia, M., Ghosh, K., & Peigler, R. S. (2016). Factors affecting on quality muga silkworm (*Antheraea assamensis* Helfer) seed crop production: A review. *Journal of Entomology and Zoology Studies*, 4(6), 806-810.
- Saikia, M., Ghosh, K., & Peigler, R. S. (2017). Wild sericigenous insect diversity of Hoollongapar Gibbon Sanctuary, Jorhat, Assam. *Journal of Entomology and Zoology Studies*, 5(2), 973-978.
- Saito, S., Tamura, K., & Aotsuka, T. (2005). Replication origin of mitochondrial DNA in insects. *Genetics*, 171(4), 1695-1705.

- Salvato, P., Simonato, M., Battisti, A., & Negrisola, E. (2008). The complete mitochondrial genome of the bag-shelter moth *Ochrogaster lunifer* (Lepidoptera, Notodontidae). *BMC Genomics*, 9(1), 331.
- Sarkar, I. N., & Trizna, M. (2011). The Barcode of Life Data Portal: bridging the biodiversity informatics divide for DNA barcoding. *Plos One*, 6(7), e14689.
- Sarmah, M. C. (2011). Eri pupa: a delectable dish of North East India. *Current Science*, 100, 3.
- Sato, S., Hirakawa, H., Isobe, S., et al. (2011). Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Research*, 18(1), 65-76.
- Scarpa, A., & Guerici, A. (1982). Various uses of the castor oil plant (*Ricinus communis* L.) a review. *Journal of Ethnopharmacology*, 5(2), 117-137.
- Schoch, C. L., Seifert, K. A., Huhndorf, S., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, 109(16), 6241-6246.
- Schoch, C. L., Sung, G. H., López-Giráldez, F., et al. (2009). The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic Biology*, 58(2), 224-239.
- Seifert, K. A. (2009). Progress towards DNA barcoding of fungi. *Molecular Ecology Resources*, 9(s1), 83-89.
- Seitz, A. (1933). The micro Lepidoptera of the world, *Indo—Australian Bombyces and Sphinges*. Stuttgart-Verlag des seitz sehen werkes, Alfred kernen, 49, 510.

- Selvakumar, T., & Datta, R. K. (2013). Role of bacteria (*Streptococcus faecalis* & *Staphylococcus aureus*), viruses and their combined infection in causation of flacherie under different environmental conditions. *Sericologia*, 53(1), 54-62.
- Sen, S. K., Jolly, M. S., & Jammy, T. R. (1970). A mycosis in the Indian tasar silkworm, *Antheraea mylitta* Drury, caused by *Penicillium citrinum* Thom. *Journal of Invertebrate Pathology*, 15(1), 1-5.
- Servín-Garcidueñas, L. E., & Martínez-Romero, E. (2014). Complete mitochondrial genome recovered from the gut metagenome of overwintering monarch butterflies, *Danaus plexippus* (L.) (Lepidoptera: Nymphalidae, Danainae). *Mitochondrial DNA*, 427-428.
- Sethusa, M. T., Millar, I. M., Yessoufou, K., et al. (2014). DNA barcode efficacy for the identification of economically important scale insects (Hemiptera: Coccoidea) in South Africa. *African Entomology*, 22(2), 257-266.
- Shangpliang, J. W., & Hajong, S. R. (2015). The wild silk moths (Lepidoptera: Saturniidae) of Khasi Hills of Meghalaya, North East India. *International Journal of Plant, Animal and Environmental Sciences*, 5(2), 20-24.
- Sharma, J., Chakraborty, S., & Uddin, A. (2014). Comparative Analysis of Codon Usage Bias Between Two Lepidopteran Insect Species: *Bombyx mandarina* and *Ostrinia furnacalis*. *International Journal of Scientific Research*, 3(6), 47-50.
- Sharma, P., Joshi, N., & Sharma, A. (2010). Isolation of genomic DNA from medicinal plants without liquid nitrogen. *Indian Journal of Experimental Biology*, 48, 610-614.

- Sharp, P. M., Stenico, M., Peden, J. F., & Lloyd, A. T. (1993). Codon usage: mutational bias, translational selection, or both?. *Biochemical Society Transactions*, 21, 835–841.
- Shen, Q. Q., & Wang, L. (2016). The complete mitochondrial genome sequence of *Heliconius hecale* (Insecta: Lepidoptera: Nymphalidae). *Mitochondrial DNA Part A*, 27(2), 1291-1292.
- Shepherd, L. D., & McLay, T. G. (2011) Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. *Journal of Plant Research*, 124, 311-314.
- Shi, B. C., Liu, W., & Wei, S. J. (2013). The complete mitochondrial genome of the codling moth *Cydia pomonella* (Lepidoptera: Tortricidae). *Mitochondrial DNA*, 24(1), 37-39.
- Shi, P., Lu, Z., He, Y., et al. (2016). Complete mitochondrial genome of *Hepialus gonggaensis* (Lepidoptera: Hepialidae), the host insect of *Ophiocordyceps sinensis*. *Mitochondrial DNA Part A*, 27(6), 4205-4206.
- Shimomura, M., Minami, H., Suetsugu, Y., et al. (2009). KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics*, 10(1), 486.
- Shioiri, C., & Takahata, N. (2001). Skew of mononucleotide frequencies, relative abundance of dinucleotides, and DNA strand asymmetry. *Journal of Molecular Evolution*, 53(4), 364-376.
- Silvestro, D., & Michalak, I. (2012). raxmlGUI: a graphical front-end for RAXML. *Organisms Diversity & Evolution*, 12(4), 335-337.
- Simmons, R. B., & Weller, S. J. (2001). Utility and evolution of cytochrome b in insects. *Molecular Phylogenetics and Evolution*, 20(2), 196-210.

- Singh, D., Chetia, H., Kabiraj, D., *et al.* (2016). A comprehensive view of the web-resources related to sericulture. *Database*, 2016.
- Singh, H. R., Unni, B. G., Neog, K. & Wann, S. B. (2011). Isolating silkworm genomic DNA without liquid nitrogen suitable for marker studies. *African Journal of Biotechnology*, 10, 11365-11372.
- Singh, K. C. & Chakrovorty, R. (2006) Seri-biodiversity of North Eastern India-an update. In *Biodiversity Conservation and Future Concern* (eds. Handique, P., & Kalita, M. C.). Gauhati University, Guwahati, 8-19.
- Singh, P. K., Das, P. K., & Negi, B. K. (1993). Studies on the incidence of uzi fly. *Annual Report*. Regional Muga Research Station, Boko, Assam, India, 7.
- Singh, R. N., & Maheshwari, M. (2002). Biological control of pests of non-mulberry silkworms and its host plants in India. *International Journal of Industrial Entomology*, 4(2), 83-91.
- Singh, R., Kumar, M., Mittal, A., & Mehta, P. K. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech*, 6, 174.
- Skowron, M. A., Guzow-Krzemińska, B., Barańska, S., Jędrak, P., & Węgrzyn, G. (2015). A rapidly progressing, deadly disease of *Actias selene* (Indian moon moth) larvae associated with a mixed bacterial and baculoviral infection. *Journal of Biosciences*, 40(3), 487-495.
- Smith, C., Heyne, S., Richter, A. S., Will, S., & Backofen, R. (2010). Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Research*, 38, W373-W377.
- Smith, M. A., Bertrand, C., Crosby, K., *et al.* (2012). *Wolbachia* and DNA barcoding insects: patterns, potential, and problems. *Plos One*, 7(5), e36514.

- Song, S. N., Tang, P., Wei, S. J., & Chen, X. X. (2016). Comparative and phylogenetic analysis of the mitochondrial genomes in basal hymenopterans. *Scientific Reports*, 6, 20972.
- Speight, M. R., Watt, A., & Hunter, M. (2005). *Ecology of insects: Concepts and application (2nd edn)*. Blackwell Science, London.
- Srivastava, P. K., & Thangavelu, K. (2005) *Sericulture and Seri-biodiversity*. Associated Publishing Company, a division of Astral International Pvt. Limited., New Delhi, 1-10.
- Stewart, J. B., & Beckenbach, A. T. (2009). Characterization of mature mitochondrial transcripts in *Drosophila*, and the implications for the tRNA punctuation model in arthropods. *Gene*, 445, 49-57.
- Stolle, E., Kidner, J. H., & Moritz, R. F. A. (2013). Patterns of Evolutionary Conservation of Microsatellites (SSRs) Suggest a Faster Rate of Genome Evolution in Hymenoptera Than in Diptera. *Genome Biology and Evolution*, 5, 151-162.
- Subharani, S., & Jayaprakash, P. (2015). Insect Pests of Muga Silkworm, *Anthereae assamensis* (Lepidoptera: Saturniidae). *Global Journal of Biology, Agriculture and Health Sciences*, 4(2), 130-134.
- Sujata, Y., & Anand, K. (2003). New record of wild silk caterpillar *Cricula trifenestrata* Helfer on large cardamon and notes on it's biology. *Uttar Pradesh Journal of Zoology*, 23(1), 67-69.
- Sumathy, R., Rao, A. S. K., Chandrakanth, N., & Gopalakrishnan, V. K. (2014). In silico identification of protein-protein interactions in Silkworm, *Bombyx mori*. *Bioinformation*, 10(2), 56.

- Sun, X. Q., Qu, Y. Q., Yao, H., *et al.* (2015). Applying DNA barcodes for identification of economically important species in Brassicaceae. *Genetics and Molecular Research*, 14(4), 15050-15061.
- Sun, Y. B., Shen, Y. Y., Irwin, D. M., & Zhang, Y. P. (2011). Evaluating the roles of energetic functional constraints on teleost mitochondrial-encoded protein evolution. *Molecular Biology and Evolution*, 28(1), 39-44.
- Sun, Y. X., Wang, L., Wei, G. Q., *et al.* (2016). Characterization of the Complete Mitochondrial Genome of *Leucoma salicis* (Lepidoptera: Lymantriidae) and Comparison with Other Lepidopteran Insects. *Scientific reports*, 6.
- Sun, Z., Wan, D. G., Murphy, R. W., *et al.* (2009). Comparison of base composition and codon usage in insect mitochondrial genomes. *Genes & Genomics*, 31(1), 65-71.
- Suriana, S., Solihin, D. D., Noor, R. R., & Thohari, A. M. (2011). Characterization of Partial Coding Region Fibroin Gene on Wild Silkmoth *Cricula trifenestrata* Helfer (Lepidoptera: Saturniidae). *Media Peternakan*, 34(1), 23.
- Swathi, H. C., Bhaskar, R. N., & Rangaswamy, K. T. (2015). Molecular characterization of different mutualistic pathogenic bacteria causing flacherie disease of silkworm *Bombyx mori* L. *Mysore Journal of Agricultural Sciences*, 49(2), 364-367.
- Szalanski, A. L., McKern, J. A., Solorzano, C., & Austin, J. W. (2010). Genetic Diversity of Ants (Hymenoptera: Formicidae) from the Ozark-St. Francis National Forest, Arkansas, USA. *Sociobiology*, 56, 601.
- Szcześniak, M. W., & Makałowska, I. (2014). miRNEST 2.0: a database of plant and animal microRNAs. *Nucleic Acids Research*, 42(D1), D74-D77.

- Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et Biophysica Acta-Bioenergetics*, 1410(2), 103-123.
- Taberlet, P., Coissac, E., Pompanon, F., *et al.* (2007). Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, 35(3), e14-e14.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution*, 30, 2725-2729.
- Tan, H. W., Liu, G. H., Dong, X., *et al.* (2011). The complete mitochondrial genome of the Asiatic cavity-nesting honeybee *Apis cerana* (Hymenoptera: Apidae). *Plos One*, 6(8), e23008.
- Tang, M., Tan, M., Meng, G., *et al.* (2014). Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mitometagenomics. *Nucleic Acids Research*, 42(22), e166-e166.
- Tang, P., Wu, H., Wang, J., Guo, Z., & Yuan, M. (2015). Complete Mitochondrial Genome of *Plodia interpunctella* and phylogenetic analysis of Pyraloidea (Insecta: Lepidoptera). *SCIENTIA SINICA Vitae*, 45(9), 890-900.
- Taylor, H. R., & Harris, W. E. (2012). An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Molecular Ecology Resources*, 12(3), 377-388.
- Thakur, S., Jha, S., & Chattoo, B. B. (2011). CastorDB: a comprehensive knowledge base for *Ricinus communis*. *BMC Research Notes*, 4(1), 356.
- Thangavelu, K. (1991). Wild sericigenous insects of India. A need for conservation. *Wild Silkmoths*, 91, 71-77.

- Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2002). Multiple Sequence Alignment Using ClustalW and ClustalX. *Current Protocols in Bioinformatics*, 2.3.1–2.3.22.
- Tikader, A. (2011). Muga culture: the rich tradition of Assam. *Indian Silk*, 2(5), 18-20.
- Tikader, A., & Kamble, C. (2010). Seri-biodiversity with Reference to Host Plants in India. *Asian and Australasian Journal of Plant Science and Biotechnology*, 4, 1-11.
- Tikader, A., Vijayan, K., & Saratchandra, B. (2013). Improvement of host plants of Muga silkworm (*Antheraea assamensis*) for higher productivity and better adaptation-A Review. *Plant Knowledge Journal*, 2(2), 83-88.
- Tikader, A., Vijayan, K., & Saratchandra, B. (2013). Muga silkworm, *Antheraea assamensis* (Lepidoptera: Saturniidae)-an overview of distribution, biology and breeding. *European Journal of Entomology*, 110, 293-300.
- Tikader, A., Vijayan, K., & Saratchandra, B. (2014). *Cricula trifenestrata* (Helfer) (Lepidoptera: Saturniidae)- A silk producing wild insect in India. *Tropical Lepidoptera Research*, 24(1), 22-29.
- Timmermans, M. J. T. N., Dodsworth, S., Culverwell, C. L., et al. (2010). Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Research*, 38(21), e197-e197.
- Timmermans, M. J., Lees, D. C., & Simonsen, T. J. (2014). Towards a mitogenomic phylogeny of Lepidoptera. *Molecular Phylogenetics and Evolution*, 79, 169-178.
- Tongsri, V., Songkumarn, P., & Sangchote, S. (2016). Leaf Spot Characteristics of *Phomopsis durionis* on Durian (*Durio Zibethinus* Murray) and Latent Infection

- of the Pathogen. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 64, 185-193.
- Tóth, G., Gáspári, Z., & Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, 10(7), 967-981.
- Tuskes, P. M., Tuttle, J. P., & Collins, M. M. (2000). The wild silk moths of North America. *Journal of Research on the Lepidoptera*, 35, 137–140.
- Uchino, K., Sezutsu, H., Imamura, M., *et al.* (2008). Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. *Insect Biochemistry and Molecular Biology*, 38(12), 1165-1173.
- Unni, B. G., Goswami, M., Kakoty, Y., *et al.* (2009). Indigenous knowledge of silkworm cultivation and its utilization in North Eastern region of India. *Indian Journal of Traditional Knowledge*, 8(1), 70-74.
- USDA, NRCS. (2015). The PLANTS Database. *National Plant Data Team, Greensboro, NC*.
- van Asch, B., Blibech, I., Pereira-Castro, I., Rei, F. T., & da Costa, L. T. (2016). The mitochondrial genome of *Prays oleae* (Insecta: Lepidoptera: Praydidae). *Mitochondrial DNA Part A*, 27(3), 2108-2109.
- Veer, V., Negi, B. K., & Rao, K. M. (1996). Dermestid beetles and some other insect pests associated with stored silkworm cocoons in India, including a world list of dermestid species found attacking this commodity. *Journal of Stored Products Research*, 32(1), 69-89.
- Vinod, K. K. (2004). Total genomic DNA extraction, quality check and quantitation. In *Proceedings of the training programme on “Classical and modern plants breeding techniques-a hands on training*. Tamil Nadu Agricultural University, Coimbatore, 109-121.

- Wang, A. R., Jeong, H. C., Han, Y. S., & Kim, I. (2014). The complete mitochondrial genome of the mountainous duskywing, *Erynnis montanus* (Lepidoptera: HesperIIDae): a new gene arrangement in Lepidoptera. *Mitochondrial DNA*, 25(2), 93-94.
- Wang, J. P., Cao, T. W., Xuan, S. B., et al. (2013). The complete mitochondrial genome of *Sasakia funebris* (Leech) (Lepidoptera: Nymphalidae) and comparison with other Apaturinae insects. *Gene*, 526(2), 336-343.
- Wang, J., Xia, Q., He, X., et al. (2005). SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Research*, 33(s1), D399-D402.
- Wang, Q., Zhang, Z., & Tang, G. (2016). The mitochondrial genome of *Atrijuglans hetaohei* Yang (Lepidoptera: Gelechioidea) and related phylogenetic analyses. *Gene*, 581(1), 66-74.
- Wang, W., Wu, Y., Yan, Y., et al. (2010). DNA barcoding of the Lemnaceae, a family of aquatic monocots. *BMC Plant Biology*, 10(1), 205.
- Wang, Y., Liu, X., & Yang, D. (2013). The first mitochondrial genome for caddisfly (Insecta: Trichoptera) with phylogenetic implications. *International Journal of Biological Sciences*, 10(1), 53-63.
- Wang, Z., Yang, W., Sun, L., Zhu, B., Li, D., & Liu, C. (2016). Characterization of a Neutral Protease Gene of *Bacillus subtilis* Isolated from the Guts of *Bombyx mori*. *Pakistan Journal of Zoology*, 48(1), 179-185.
- Wei, L., He, J., Jia, X., et al. (2014). Analysis of codon usage bias of mitochondrial genome in *Bombyx mori* and its relation to evolution. *BMC Evolutionary Biology*, 14(1), 262.
- Wei, S. J., Shi, M., Chen, X. X., et al. (2010). New views on strand asymmetry in insect mitochondrial genomes. *Plos One*, 5(9), e12708.

- Wolstenholme, D. R. (1992). Animal mitochondrial DNA: structure and evolution. *International Review of Cytology*, 141, 173-216.
- Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., & Yuen, K. Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection*, 14(10), 908-934.
- Wu, L. W., Lees, D. C., Yen, S. H., & Hsu, Y. F. (2010). The complete mitochondrial genome of the near-threatened swallowtail, *Agehana maraho* (Lepidoptera: Papilionidae): evaluating sequence variability and suitable markers for conservation genetic studies. *Entomology News*, 121(3), 267-280.
- Wu, L., Wu, X., Xiong, X., et al. (2016). The complete mitochondrial genome of *Euthrix laeta* (Lepidoptera: Lasiocampidae). *Mitochondrial DNA Part B*, 1(1), 532-533.
- Wu, Q. L., Cui, W. X., & Wei, S. J. (2015). Characterization of the complete mitochondrial genome of the black cutworm *Agrotis ipsilon* (Lepidoptera: Noctuidae). *Mitochondrial DNA*, 26(1), 139-140.
- Wu, Y. P., Li, J., Zhao, J. L., et al. (2012). The complete mitochondrial genome of the rice moth, *Corcyra cephalonica*. *Journal of Insect Science*, 12(1), 72.
- Wu, Y. P., Zhao, J. L., Su, T. J., et al. (2012). The complete mitochondrial genome of *Leucoptera malifoliella* Costa (Lepidoptera: Lyonetiidae). *DNA and Cell Biology*, 31(10), 1508-1522.
- Wu, Y. P., Zhao, J. L., Su, T. J., et al. (2016). The complete mitochondrial genome of *Carposina sasakii* (Lepidoptera: Carposinidae). *Mitochondrial DNA Part A*, 27(2), 1432-1434.

- Xia, Q., Cheng, D., Duan, J., *et al.* (2007). Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. *Genome Biology*, 8(8), R162.
- Xia, Q., Zhou, Z., Lu, C., *et al.* (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, 306(5703), 1937-1940.
- Xu, H. E., Zhang, H. H., Xia, T., *et al.* (2013) BmTEdb: a collective database of transposable elements in the silkworm genome. *Database*, 2013, bat055.
- Xu, Y. M., Chen, S. C., Wang, X. Q., Peng, P., & Li, P. W. (2016). The complete mitogenome of *Jankowskia athleta* (Lepidoptera: Geometridae). *Mitochondrial DNA Part A*, 27(4), 3035-3036.
- Yajun, Z., Guoliang, Z., Rui, F., Jun, Y., & Jianping, Y. (2010). The complete sequence determination and analysis of *Lymantria dispar* (Lepidoptera: Lymantriidae) mitochondrial genome. *Plant Quarantine*, 4(005).
- Yang, J., Vázquez, L., Chen, X., *et al.* (2017). Development of chloroplast and nuclear DNA markers for Chinese oaks (*Quercus* subgenus *Quercus*) and assessment of their utility as DNA barcodes. *Frontiers in plant science*, 8.
- Yang, L., Wei, Z. J., Hong, G. Y., Jiang, S. T., & Wen, L. P. (2009). The complete nucleotide sequence of the mitochondrial genome of *Phthonandria atrilineata* (Lepidoptera: Geometridae). *Molecular Biology Reports*, 36(6), 1441-1449.
- Yang, X., Cameron, S. L., Lees, D. C., Xue, D., & Han, H. (2015). A mitochondrial genome phylogeny of owlet moths (Lepidoptera: Noctuoidea), and examination of the utility of mitochondrial genomes for lepidopteran phylogenetics. *Molecular Phylogenetics and Evolution*, 85, 230-237.

- Yang, X., Liu, X. M., Sun, X. K., & Feng, C. (2012). Study on selvage parameter based on silk fabric specifications database. In *Advanced Materials Research*. Trans Tech Publications, 502, 159-163.
- Yang, X., Xue, D., & Han, H. (2013). The complete mitochondrial genome of *Biston panterinaria* (Lepidoptera: Geometridae), with phylogenetic utility of mitochondrial genome in the Lepidoptera. *Gene*, 515(2), 349-358.
- Yang, Y. W., Lai, K. N., Tai, P. Y., & Li, W. H. (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *Journal of Molecular Evolution*, 48(5), 597-604.
- Yang, Z., & Yoder, A. D. (1999). Estimation of the transition/transversion rate bias and species sampling. *Journal of Molecular Evolution*, 48(3), 274-283.
- Ye, F., & You, P. (2016). The complete mitochondrial genome of *Paracymoriza distinctalis* (Lepidoptera: Crambidae). *Mitochondrial DNA Part A*, 27(1), 28-29.
- Ye, F., Yu, H. L., Li, P. F., & You, P. (2015). The complete mitochondrial genome of *Lista haraldusalis* (Lepidoptera: Pyralidae). *Mitochondrial DNA*, 26(6), 853-854.
- Yin, C., Shen, G., Guo, D., et al. (2016). InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Research*, 44(D1), D801-D807.
- Yin, J., Hong, G. Y., Wang, A. M., Cao, Y. Z., & Wei, Z. J. (2010). Mitochondrial genome of the cotton bollworm *Helicoverpa armigera* (Lepidoptera: Noctuidae) and comparison with other Lepidopterans. *Mitochondrial DNA*, 21(5), 160-169.
- Yu, J., Xue, J. H., & Zhou, S. L. (2011). New universal matK primers for DNA barcoding angiosperms. *Journal of Systematics and Evolution*, 49(3), 176-181.

- Yuan, M. L., & Zhang, Q. L. (2013). The complete mitochondrial genome of *Gynaephora menyuanensis* (Lepidoptera: Lymantriidae) from the Qinghai-Tibetan Plateau. *Mitochondrial DNA*, 24(4), 328-330.
- Yuan, Z. H., Lan, X. Q., Yang, T., Xiao, J. & Zhou, Z. Y. (2006). Investigation and analysis of the bacteria community in silkworm intestine. *Wei sheng wu xue bao= Acta microbiologica Sinica*, 46, 285-291.
- Yukuhiro, K., Sezutsu, H., Itoh, M., Shimizu, K., & Banno, Y. (2002). Significant levels of sequence divergence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkmoth, *Bombyx mandarina*, and its close relative, the domesticated silkmoth, *Bombyx mori*. *Molecular Biology and Evolution*, 19, 1385-1389.
- Yukuhiro, K., Sezutsu, H., Tamura, T., Kosegawa, E., & Kiuchi, M. (2011). Nucleotide sequence variation in mitochondrial *COI* gene among 147 silkworm (*Bombyx mori*) strains from Japanese, Chinese, European and moltinism classes. *Genes & Genetic Systems*, 86(5), 315-323.
- Zeng, Q., Chen, H., Zhang, C., et al. (2015). Definition of Eight Mulberry Species in the Genus *Morus* by Internal Transcribed Spacer-Based Phylogeny. *Plos One*, 10(8), e0135411.
- Zhang, D. X., & Hewitt, G. M. (1997). Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology*, 25, 99-120.
- Zhang, D. X., Szymura, J. M., & Hewitt, G. M. (1995). Evolution and structural conservation of the control region of insect mitochondrial DNA. *Journal of Molecular Evolution*, 40(4), 382-391.

Zhang, D., Ding, G., Ge, B., Zhang, H., & Tang, B. (2017). Molecular evolution of mitochondrial coding genes in the oxidative phosphorylation pathway in malacostraca: purifying selection or accelerated evolution?. *Mitochondrial DNA Part A*, 28(4), 593-596.

Zhang, Z., Yin, C., Liu, Y., et al. (2014). iPathCons and iPathDB: an improved insect pathway construction tool and the database. *Database*, 2014, bau105.

Zhao, F., Huang, D. Y., Sun, X. Y., et al. (2013). The first mitochondrial genome for the butterfly family Riodinidae (*Abisara fylloides*) and its systematic implications. *Dongwuxue Yanjiu*, 34(E4-5), E109.

Zhao, J. L., Zhang, Y. Y., Luo, A. R., et al. (2011). The complete mitochondrial genome of *Spilota lechriaspis* Meyrick (Lepidoptera: Tortricidae). *Molecular Biology Reports*, 38(6), 3757-3764.

Website URL:

Area de Conservación Guanacaste (ACG) northwestern Costa Rica:
<http://janzen.sas.upenn.edu/caterpillars/database.lasso>

Bill's Lepidoptera Photos: <http://www.harkphoto.com/menu.html>

Butterflies and moths of North America (BAMONA):
<http://www.butterfliesandmoths.org/>

Butterflies and Moths of the World: <http://www.nhm.ac.uk/our-science/data/butmoth/>

Butterflies and Moths: http://www.stillwaterwoods.net/butterflies_&_moths.htm

Catalogue of Life: <http://www.catalogueoflife.org/col/info/ac>

Common Names of Insects Database:
http://www.entsoc.org/pubs/common_names

Database of Insects and their Food Plants:

<http://www.brc.ac.uk/dbif/homepage.aspx>

Discover Life: <http://www.discoverlife.org/20/q>

Encyclopedia of Life (EOL) Database: <http://www.eol.org>

EPPO Global Database: <https://gd.eppo.int>

Espacenet: <https://worldwide.espacenet.com/>

Google Patents: <https://patents.google.com/>

INPN: <http://inpn.mnhn.fr/>

Integrated taxonomic information system (ITIS) database: <http://www.itis.gov/>

Moths of Borneo: <http://www.mothsofborneo.com/>

PATENTSCOPE: <http://www.wipo.int/patentscope/en/>

ScienceDirect: <http://www.sciencedirect.com/science/journals>

Scopus: <https://www.scopus.com/home.uri>

The Global Lepidoptera Names Index or LepIndex: <http://www.nhm.ac.uk/our-science/data/lepindex/index.html>

Universal Chalcidoidea Database: <http://www.nhm.ac.uk/our-science/data/chalcidoids/database/>

Web of Science:

http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=W1gh56McCwbjz98GS8v&preferencesSaved=



List of Publications

Published manuscripts

1. **Singh, D.**, Kabiraj, D., Sharma, P., Chetia, H., Mosahari, P. V., Neog, K., & Bora, U*. (2017) The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. *PLOS ONE*, 12(11), e0188077.
2. **Singh, D.**, Chetia, H., Kabiraj, D., Sharma, S., Kumar, A., Sharma, P., Deka, M., & Bora, U.* (2016) A comprehensive view of the current web-resources in sericulture and related fields. *Database*, 2016, 1-31.
3. **Singh, D.**, Sett, A., Sharma, P., Deka, M., & Bora, U.* (2016) DNA Barcoding- Current status and future prospects in '*Biotechnological tools for genetic resources*'. Ed. Prof. P. C. Deka. Daya Publishing House. Astral International Pvt. Ltd. New Delhi, pp 19-58.

Manuscripts under review/preparation

1. **Singh, D.**, Kabiraj, D., Chetia, H., Sharma, P., Neog, K., & Bora, U.* (2017) Complete mitochondrial genome of *Cricula trifenestrata* and its comparative analysis with other lepidopteran insects. (Under preparation).
2. **Singh, D.**, Sharma, P., Neog, K., & Bora, U.* (2017) DNA barcoding of few seribioresources of Assam, India. (Under preparation).

Manuscripts from collaborative work

1. Ojha, S., **Singh, D.**, Sett, A., Chetia, H., Kabiraj, D., & Bora, U.* (2017) Nanotechnology in Crop Protection in '*Nanomaterials in Plants, Algae and Micro-organism: Concepts and Controversies*'. Eds. D. K. Tripathi, P. Ahmad, S. Sharma, D. K. Chauhan, N. K. Dubey. Elsevier, pp 345-391.
2. Chetia, H., Kabiraj, D., **Singh, D.**, Mosahari, P. V., Das, S., Sharma, P., Neog, K., Sharma, S., Jayaprakash, P., & Bora, U.* (2017) De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). *Gene*, 611, 54-65.
3. Kabiraj, D., Kalita, J. J., Chetia, H., **Singh, D.**, & Bora, U.* (2015) Expanding the frontiers of rice research through omics. *Journal of Assam Science Society*, 56(2), 1-28.
4. Bora, U.*, Sett, A., & **Singh, D.** (2013) Nucleic Acid Based Biosensors for Clinical Applications. *Biosensors Journal*, 2, 104.
5. Mosahari, P. V., **Singh, D.**, Kalita, J. J., Sharma, P., Chetia, H., Kabiraj, D., Mahanta, C., & Bora, U.* Nanotoxicity: Impact on Health and Environment. CRC Taylor & Francis. (Under review).

* represents corresponding author

List of Conferences/Workshops/ Symposia

1. **Singh D**, Raut R and Kumar A (2012) 'Development of Seri-Bioresource Database (SBDB)'. International consultative meeting on Seribiotechnology (ICMS) 2012, organized by Institute of Bioresources and Sustainable Development (IBSD), Imphal, Manipur.
2. Raut R, **Singh D** and Kumar A (2012) 'Silkworm nucleotide databases'. International consultative meeting on Seribiotechnology (ICMS) 2012, organized by Institute of Bioresources and Sustainable Development (IBSD), Imphal, Manipur.
3. **Singh D** (2014) 'Seri-Bioinformatics Databases: Current status and future prospects'. *Metamorphose* souvenir in the Conference on Exploitation of Seribiodiversity for Novel Product Development, organized by IIT Guwahati.
4. **Singh D** (2014) 'Silkworm as a biomedical model'. International Conference on Disease Biology and Therapeutics (ICDBT 2014), organized by Institute of Advanced Study in Science and Technology, Guwahati, Assam.
5. **Singh D**, Khan A and Bora U (2015) 'Genetic manipulation in silkworms'. Proceedings of the National Seminar on Insect Biotechnology, organized by Nambol L. Sansoi College, Manipur.

6. **Singh D**, Kabiraj D, Chetia H and Bora U (2016) 'SeriPort'. National Seminar on Problems and Prospects of Muga and Eri Silk Sectors, organized by Central Muga Eri Research & Training Institute, Jorhat.
7. **Singh D** and Bora U (2016) 'SeriPort: Bridging Sericulture to Informatics'. 1st International Agrobiodiversity Congress (IAC), held at New Delhi.
8. **Singh D**, Kabiraj D, Sharma P, Chetia H, Mosahari PV, Neog K and Bora U (2018) 'Comparative mitogenomics and phylogenetics of *Antheraea assamensis* with other lepidopteran insects'. BIODIVERSE 2018, organized by ADNAT, Indian Institute of Technology Guwahati.
9. Bora U, Sharma P, **Singh D**, Kabiraj D, Chetia H, Mosahari PV, Nath A, Barkataki MP, Bharali B, Saikia D, Singh T (2018) 'Diversity of insect mitochondrial genomes'. BIODIVERSE 2018, organized by ADNAT, Indian Institute of Technology Guwahati.
10. Chetia H, Bharali B, Kabiraj D, Pathak R, **Singh D**, Mosahari PV, Bora U (2018) 'Banking on insects: a roadmap to conservation and sustainability'. BIODIVERSE 2018, organized by ADNAT, Indian Institute of Technology Guwahati.

Vitae

Deepika was born on 13th October 1988 in Chandigarh, India. She passed her Secondary School Examination in 2004 and Higher Secondary School Examination in 2006 both conducted by Board of School Education Haryana, Haryana. She completed her Bachelor of Science from Pt.



J.L.N. Government College, Faridabad under Maharshi Dayanand University, Rohtak in 2009. She completed her Master of Science in Biotechnology from Kurukshetra University, Kurukshetra in 2011. She qualified exams including GATE Biotechnology, Department of Biotechnology DBT JRF (Category B Merit List) and CSIR-NET in Life Sciences (eligible for lectureship) in the year 2011.

Deepika joined her Ph.D. Programme in July 2011 under the supervision of Prof. Utpal Bora at Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India. She successfully completed the course work with 9.0/10 Cumulative Point Index (CPI). She received Institute fellowship under the scheme by the Ministry of Human Resource and Development (MHRD), New Delhi, India. Her thesis work focuses on mitochondrial genome informatics of selected seri-bioresources and development of seri-related database. She had published five peer-reviewed articles and two book chapters during her thesis work.

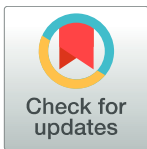
RESEARCH ARTICLE

The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects

Deepika Singh^{1,2}, Debajyoti Kabiraj¹, Pragya Sharma³, Hasnahana Chetia¹, Ponnala Vimal Mosahari², Kartik Neog⁴, Utpal Bora^{1,2*}

1 Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India, **2** Centre for the Environment, Indian Institute of Technology Guwahati, Assam, India, **3** Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati, Assam, India, **4** Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, Jorhat, Assam, India

* ubora@iitg.ernet.in, ubora@rediffmail.com



OPEN ACCESS

Citation: Singh D, Kabiraj D, Sharma P, Chetia H, Mosahari PV, Neog K, et al. (2017) The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. PLoS ONE 12(11): e0188077. <https://doi.org/10.1371/journal.pone.0188077>

Editor: Daniel Doucet, Natural Resources Canada, CANADA

Received: June 28, 2017

Accepted: October 31, 2017

Published: November 15, 2017

Copyright: © 2017 Singh et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The full annotated mitogenome sequence and SRA data of *A. assamensis* submitted to NCBI GenBank are available under the accession numbers KU379695 and SRR3948351, respectively.

Funding: The authors thank the Department of Biotechnology, Govt. of India, New Delhi for supporting the research through the UXCEL project

Abstract

Muga (*Antheraea assamensis*) is an economically important silkworm endemic to the states of Assam and Meghalaya in India and is the producer of the strongest known commercial silk. However, there is a scarcity of genomic and proteomic data for understanding the organism at a molecular level. Our present study is on decoding the complete mitochondrial genome (mitogenome) of *A. assamensis* using next generation sequencing technology and comparing it with other available lepidopteran mitogenomes. Mitogenome of *A. assamensis* is an AT rich circular molecule of 15,272 bp (A+T content ~80.2%). It contains 37 genes comprising of 13 protein coding genes (PCGs), 22 tRNA and 2 rRNA genes along with a 328 bp long control region. Its typical $tRNA^{Met}-tRNA^{Ile}-tRNA^{Gln}$ arrangement differed from ancestral insects ($tRNA^{Ile}-tRNA^{Gln}-tRNA^{Met}$). Two PCGs *cox1* and *cox2* were found to have CGA and GTG as start codons, respectively as reported in some lepidopterans. Interestingly, *nad4l* gene showed higher transversion mutations at intra-species than inter-species level. All PCGs evolved under strong purifying selection with highest evolutionary rates observed for *atp8* gene while lowest for *cox1* gene. We observed the typical clover-leaf shaped secondary structures of tRNAs with a few exceptions in case of $tRNA^{Ser1}$ and $tRNA^{Tyr}$ where stable DHU and TΨC loop were absent. A significant number of mismatches (35) were found to spread over 19 tRNA structures. The control region of mitogenome contained a six bp (CTTAGA/G) deletion atypical of other *Antheraea* species and lacked tandem repeats. Phylogenetic position of *A. assamensis* was consistent with the traditional taxonomic classification of Saturniidae. The complete annotated mitogenome is available in GenBank (Accession No. KU379695). To the best of our knowledge, this is the first report on complete mitogenome of *A. assamensis*.

(Sanction Order No: BT/411/NE/U-Excel/2013 dated 06.02.2014). The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Mitochondria are known to have descended from α -proteobacterium endosymbionts and have retained numerous bacterial features [1]. Apart from being the powerhouse of the cell, they are involved in various cellular processes like fatty acid metabolism, apoptosis and aging [2]. These functions are carried out by many nuclear-encoded genes along with extra-nuclear genes that include protein-coding genes (PCGs), transfer RNA (tRNA) and ribosomal rRNA (rRNA) genes and are present in the mitogenome which is mostly circular and self-replicative. It also consists of several non-coding regions with the longest being AT-rich control region comprising of several conserved regions and repeats. Although small in size, the mitogenome governs maternal inheritance and has several unique features like faster evolution rate, low or absence of homologous recombination, evolutionary conserved gene products and richness in genetic polymorphism. This makes it a potential marker for barcoding, phylogeographic and phylogenetic studies [3]. It plays a potential role in molecular evolutionary studies by elucidating evolutionary models and substitution patterns that vary timely and across sequences. Compared to individual genes, whole mitogenomes are more informative phylogenetic models due to its multiple genome level features like gene position, content, secondary structures of RNA and control region.

Over the past few decades, animal mitogenomes, particularly insects (~80% of the sequenced arthropods), have been widely studied for comparative genomics and molecular systematics [3–4]. More than 250 Lepidopteran insect mitogenomes have been sequenced till now using Sanger sequencing and next generation sequencing (NGS) technologies as found in GenBank (<https://www.ncbi.nlm.nih.gov/>). The mitogenome ranges from 14–16 kilo basepairs (kb) in majority of the lepidopterans and consists of 37 genes (13 PCGs, 2 rRNA genes, 22 tRNA genes) along with a control region. The PCGs encode 2 subunits of ATPase (ATP6 and ATP8), 3 subunits of cytochrome *c* oxidase (COI, COII and COIII), 1 subunit of cytochrome *b* (CYTB) and 7 subunits of NADH dehydrogenase (ND1, ND2, ND3, ND4, ND4L, ND5 and ND6). These proteins are responsible for oxidative phosphorylation (OXPHOS) as they form essential mitochondrial membrane-associated protein complex systems [3]. The control region plays a role in the replication and transcription of mitogenome. The utilities and potential of mitochondrial PCGs (*cox1* and *cox2*) as barcode markers have been well demonstrated in the order Lepidoptera [5]. The comparative mitogenome analysis also elucidated sequence divergence patterns among domesticated and non-domesticated lepidopteran mitogenomes [6–7]. Although frequency of mitogenome sequencing of lepidopterans has increased, the evolutionary relationships among many family members of the same order have been hardly investigated [8–9].

Saturniidae is the most diverse family of wild silkmoths which include giant moths, royal moths and emperor moths. Most of these silkmoths are unexplored and might have potential significance in sericulture [10]. Till date, mitogenome data of eleven species have been sequenced and made available in GenBank. However, numerous individual gene sequences of various wild species are also available whose complete mitogenome sequencing has not been done. *Antheraea assamensis*, muga silkmoth is a semi-domesticated and one of the economically important moths of the Saturniidae family. It is endemic to erstwhile undivided Assam and its adjoining hilly areas located in the North-Eastern part of India. It is a multivoltine and polyphagous moth that primarily thrives on two main host plants *Persea bombycina* and *Litsea monopetala* [11]. Its silk has wide applications in textile industry and has great potential as a biomaterial due to its unique biophysical properties like golden lustre, tenacity and absorption of UV radiation [12]. Like other *Antheraea* species, it produces reelable silk which is the most expensive silks of the world. However, its semi-domesticated nature and extraction of fibroin

directly from cocoon fibers limit its extensive rearing and prospects of global applicability as a biomaterial. The whole genome of *A. assamensis* is not yet available. However, *de novo* transcriptome data from our laboratory is available in GenBank (Accession Number- SRX1293136, SRX1293137 and SRX1293138) [13].

In the present study, we report the whole mitogenome sequence of *A. assamensis* using NGS and comparative analysis of its sequences and genome architectures with that of the other lepidopterans. The comparative study was based on several characteristics such as genome arrangement, PCGs, tRNAs, rRNAs, nucleotide composition, codon usage, evolutionary rates, gene divergence, conserved regions in control region, etc. Furthermore, phylogenetic trees inferred using datasets like nucleotide sequences of 13 PCGs and 13 PCGs+2 rRNAs were analyzed to elucidate the relationships among lepidopteran insects. This study will facilitate a better understanding of the comparative and evolutionary biology of *A. assamensis* with the other lepidopteran insects.

Materials and methods

Sample processing, DNA sequencing and assembly

The larvae of *A. assamensis* were reared on *Persea bombycina* or Som (a primary host plant) in the experimental field of Central Muga Eri Research and Training Institute, Lahdoigarh, Jorhat, Assam, India following recommended package of practices from brushing till attainment of maturity [14]. The effective rate of rearing of silkworms referred as the number of mature larvae collected from the total brushed larvae was found to be 61.63%. Other rearing parameters like body weight of mature larvae (11.84 g), cocoon weight (6.26 g) and shell weight (0.55 g) were found to be suitable for experimental work. The fifth instar larval stage of *A. assamensis* was used for the present study (Sample ID: CMERI-Aa-001). The larvae were sterilized by washing with 70% ethanol before being processed for mitogenome studies. The larvae were stored in 95% absolute ethanol at -80°C for future use. The total DNA was extracted using CTAB (Cetyl trimethylammonium bromide) based buffer and silica column by the service provider (Genotypic Technology Pvt. Ltd. Bengaluru, India). Subsequently, mitochondrial DNA was enriched from total DNA extracted once the integrity, quantity and purity of extracted DNA was confirmed by agarose gel electrophoresis, light absorbance and fluorescence spectroscopy. The complete overview of sequencing and analysis of mitochondrial genome of *A. assamensis* is represented in Fig 1.

Briefly, the preferential enrichment of mitochondrial DNA was carried out with NEBNext microbiome DNA enrichment kit (New England Biolabs, USA) which selectively removes CpG-methylated eukaryotic nuclear DNA. The enriched mitochondrial DNA was subjected to DNA sequencing at Genotypic Technology's Genomics facility where the enriched DNA was acoustically sheared to 300–500 bp using a specially designed Covaris microTube (for focused ultra-sonication). The fragmented mitochondrial DNA was cleaned up using HighPrep beads (MagBio Genomics, Inc, Gaithersburg, Maryland) and subjected to end-repair, A-tailing and ligation with multiplex adaptors using the NEXTFlex DNA Sequencing kit (Catalogue # 5140–02, Bioo Scientific). The ligated DNA was cleaned with HighPrep beads and subjected to amplification via PCR as follows: initial denaturation at 98°C for 2 min; 10 cycles of denaturation at 98°C for 30 sec, annealing at 65°C for 30 sec and extension at 72°C for 60 sec; and a final extension at 72°C for 4 min) using the primers provided by NEXTFlex DNA Sequencing kit. Finally, the PCR product was purified with HighPrep beads, quantified and fragment range was assessed using Qubit fluorometer and the Agilent D1000 Tape (Agilent Technologies), respectively prior to sequencing. The sequencing was carried out in Illumina NextSeq 500 sequencer (Illumina Inc, Sandiego, USA) through 2 × 150 bp paired-end chemistry. The

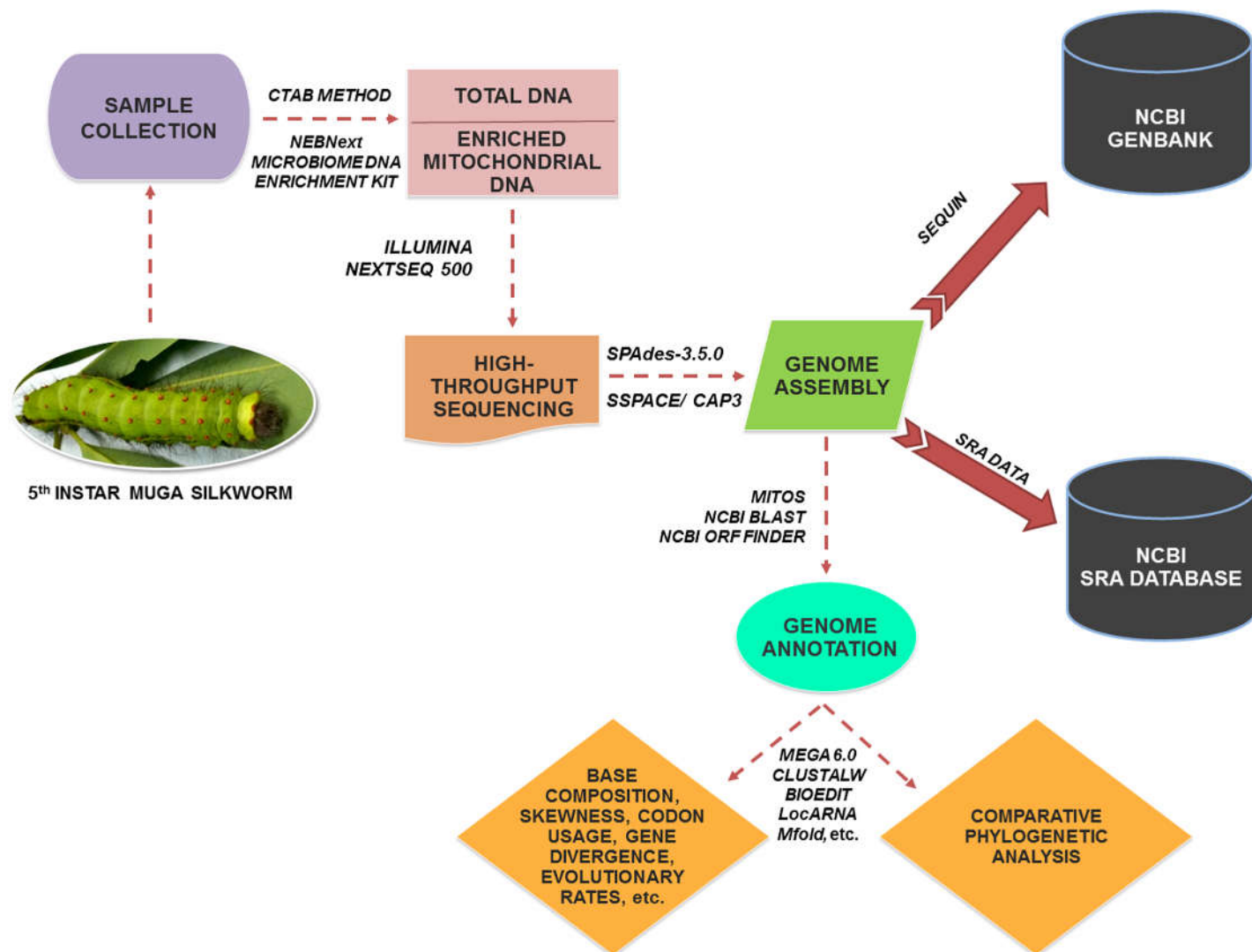


Fig 1. A pictorial overview of the methodologies used for sequencing and analysis of *A. assamensis* mitogenome.

<https://doi.org/10.1371/journal.pone.0188077.g001>

raw paired end data obtained was de-multiplexed using Bcl2fastQ, assessed with FastQC tool [15] to remove low quality bases ($Q < 30$) and preprocessed using ABLT-Scripts (Genotypic technology, Bangalore India). The SPAdes-3.5.0 was used for sequence assembly followed by the filling of gaps. Finally, scaffolding of assembled contigs and clustering were carried out with SSPACE and CAP3 programs, respectively [16–17].

Genome annotation, visualization and comparative analysis

The assembled scaffolds were annotated using MITOS web-server to predict the location of protein coding regions/genes, tRNAs, rRNAs and secondary structures of tRNAs. MITOS is a widely used webserver for the annotation of metazoan mitochondrial genomes due to its advanced annotation methodology [18]. The location of PCGs, tRNAs and rRNAs was further confirmed using tools like NCBI (National Center for Biotechnology Information)–BLAST, BioEdit and Clustal Omega by comparing the sequences of *A. assamensis* with the respective sequences published for other lepidopteran insects [19–20]. The boundaries of PCGs i.e.

initiation and termination codons were determined using NCBI ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>) by specifying the invertebrate mitochondrial genetic code. Further, the number of overlapping or spacer regions between the genes were visualized and calculated manually. The control region was validated by comparing with the available sequences in GenBank and the tandem repeats in this region were searched through Tandem Repeat Finder [20–21]. Then, the whole mito-map was constructed and visualized using Blast Ring Image Generator (BRIG) tool [22]. Finally, the complete annotated file of mitochondrial genome was prepared using NCBI Sequin tool (<http://www.ncbi.nlm.nih.gov/Sequin/>) and the sequin file along with SRA data were submitted to NCBI GenBank.

For understanding evolutionary relationships of *A. assamensis* with various lepidopterans as listed in Table 1, the sequences of whole mitogenome, coding regions, tRNAs, rRNAs and

Table 1. List of species considered for comparative mitogenome study. (Highlighted ones are the organisms of Bombycoidea superfamily used for comparative mitogenome analysis with respect to *A. assamensis*).

Superfamily	Family	Species	Size (bp)	Accession No.	References
Bombycoidea	Saturniidae	<i>Antheraea assamensis</i>	15,272	KU379695	This study
		<i>Attacus atlas</i>	15,282	KF006326	[23]
		<i>Samia cynthia ricini</i>	15,384	JN215366	[24]
		<i>Antheraea pernyi</i>	15,575	AY242996	[25]
		<i>Antheraea yamamai</i>	15,338	EU726630	[26]
		<i>Antheraea frithi</i>	15,338	NC_027071	GenBank
		<i>Actias selene</i>	15,236	NC_018133	[27]
		<i>Saturnia boisduvalii</i>	15,360	NC_010613	[28]
		<i>Eriogyna pyretorum</i>	15,327	FJ685653	[29]
	Bombycidae	<i>Bombyx mandarina</i>	15,682	AY301620	[30]
		<i>Bombyx mori</i>	15,656	AB070264	[31]
	Sphingidae	<i>Rondotia menciiana</i>	15,301	KC881286	[32]
		<i>Manduca sexta</i>	15,516	EU286785	[33]
		<i>Sphinx morio</i>	15,299	KC470083	[34]
Geometroidea	Geometridae	<i>Biston panterinaria</i>	15,517	NC_020004	[35]
		<i>Phthonandria atrilineata</i>	15,499	NC_010522	[36]
Cossoidea	Cossidae	<i>Eogystia hippophaecolus</i>	15,431	KC831443	[37]
Papilionoidea	Riodinidae	<i>Abisara fylloides</i>	15,301	HQ259069	[38]
	Lycaenidae	<i>Spindasis takanonis</i>	15,349	HQ184266	[39]
		<i>Protantigius superans</i>	15,248	HQ184265	[39]
		<i>Papilio maraho</i>	16,094	FJ810212	[40]
	Papilionidae	<i>Parnassius bremeri</i>	15,389	NC_014053	[41]
		<i>Danaus plexippus</i>	15,314	NC_021452	GenBank
		<i>Eumenis autonoe</i>	15,489	GQ868707	[42]
Noctuoidea	Lymantriidae	<i>Gynaephora menyuanensis</i>	15,770	KC185412	[43]
		<i>Lymantria dispar</i>	15,569	NC_012893	GenBank
Yponomeutoidea	Plutellidae	<i>Plutella xylostella</i>	16,014	KM023645	[44]
Tortricoidea	Tortricidae	<i>Adoxophyes honmai</i>	15,680	DQ073916	[45]
		<i>Cydia pomonella</i>	15,253	JX407107	[46]
Pyraloidea	Crambidae	<i>Ostrinia furnacalis</i>	14,536	NC_003368	[47]
		<i>Ostrinia nubilalis</i>	14,535	NC_003367	[47]
Diptera (Outgroup)		<i>Drosophila melanogaster</i>	19,517	NC_001709	[48]
		<i>Drosophila yakuba</i>	16,019	NC_001322	[49]
		<i>Anopheles gambiae</i>	15,363	NC_002084	[50]

<https://doi.org/10.1371/journal.pone.0188077.t001>

control region were retrieved from the NCBI GenBank database and a comparative analysis was performed.

Comparative mitogenome analysis. Comparative mitogenome analysis was carried out in order to find out similarities and differences of *A. assamensis* with other insects in terms of characteristics like genome size, organization, structure, gene content and rearrangement. The sequence homology of *A. assamensis* mitogenome with selected insects was also performed using Clustal Omega [20].

Comparative analysis among protein coding genes (PCGs). The PCGs of *A. assamensis* were compared with selected organisms based on their number, length, initiation and termination codons. The sequence similarity among the PCGs was determined by aligning the sequences of *A. assamensis* with selected Bombycoids using Clustal Omega. MEGA 6.0 tool was used for determining the gene-by-gene divergences in 13 PCGs in terms of phylogenetically informative sites, conserved sites as well as variable sites [51]. The same tool was used for estimating nucleotide substitution and evolutionary rates among the genes in terms of transitions (ts), transversions (tv), rate of synonymous substitutions (Ks), rate of non-synonymous substitutions (Ka) and their ratios. Further, the correlation between GC content and Ka/Ks ratio was studied in order to predict the effect of GC content on evolutionary rates of PCGs.

Comparative analysis among transfer RNAs (tRNAs). The length, arrangement, secondary structures and variation in the structures were studied among the selected organisms. Homologous sites in their secondary structures were identified by aligning the tRNA sequences using LocARNA tool and the type/number of mismatches were calculated [52].

Comparative analysis among ribosomal RNAs (rRNAs). The number, type, location and length of rRNAs in *A. assamensis* were compared with the selected organisms in order to study the conservation pattern. In addition, sequence homology of rRNAs of *A. assamensis* with the selected organisms was determined and gene-by-gene divergences among the rRNAs were studied using MEGA 6.0 tool [51].

Comparative analysis among control region (A+T rich region). The control region of *A. assamensis* was compared with selected organisms based on the length and location. Multiple sequence alignment of control region was then carried out using Clustal Omega and their conserved regions, repeats and indels were visualized using BioEdit tool [19].

Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions. The OS and IGS regions of *A. assamensis* mitogenome were compared with selected organisms (Table 1) in terms of number, length and location. The sequence homology of OS and IGS was determined through sequence alignment using Clustal Omega and BioEdit tools to find conserved motifs.

Comparative analysis with respect to nucleotide composition, skewness and codon usage. The nucleotide composition of sequences of whole mitochondrial genome, concatenated and individual PCGs, tRNAs, rRNAs, spacers and control region was calculated using MEGA 6.0 software [51]. The base composition skewness was also calculated for all the regions of mitogenome using the formula (Eq 1 and Eq 2) described by Junqueira et al., 2004 [53].

$$AT \text{ skew } = \frac{1}{4} \frac{A - T}{A + T} \quad \text{---} \quad \frac{A}{A + T} - \frac{T}{A + T}$$

$$GC \text{ skew } = \frac{1}{4} \frac{G - C}{G + C} \quad \text{---} \quad \frac{G}{G + C} - \frac{C}{G + C}$$

Where A, T, G and C denote the frequencies of respective bases.

The codon usage and relative synonymous codon usage (RSCU) values of PCGs were determined using MEGA 6.0.

Comparative phylogenetic analysis

About 34 organisms representing 13 different families of 8 super families within the order Lepidoptera were considered for phylogenetic analysis along with three organisms belonging to the order Diptera as outgroups (Table 1). The phylogenetic relationship of *A. assamensis* was elucidated using two criteria's: (i) concatenated nucleotide sequences of 13 PCGs and (ii) concatenated nucleotide sequences of 13 PCGs and 2 rRNAs.

The nucleotide sequences of each PCG were translated via amino acid alignment using MAFFT algorithm in TranslatorX server which were then back translated into nucleotide alignment with GBlocks [54]. The individual nucleotide sequences obtained from GBlocks and rRNA sequences were then concatenated for further phylogenetic analysis. Substitution model optimization for each dataset was performed in jModelTest 2.1.7 [55–56]. The GTR+G+I model was selected as the best substitution model for both the datasets to conduct maximum likelihood (ML) and bayesian inference (BI) analysis. Phylogenetic trees for both the datasets were inferred using maximum likelihood (ML) method implemented in RaxmlGUI v1.3 with 1000 bootstrap replicates [57]. The BI analysis was conducted using Markov chain Monte Carlo (MCMC) method in MrBayes v3.2.6 for 1,150,000 and 1,320,000 generations for PCG and PCG+rRNA datasets, respectively [58–59]. Trees were sampled every 1000 generations and the first 25% were discarded as burn-in. The stationary was considered to be reached when the average standard deviation of split frequency reached below 0.01. Other parameters like effective sample size (ESS) and potential scale reduction factor (PSRF) were evaluated for stationary using Tracer v1.6 [60]. The consensus phylogenetic trees obtained for both the datasets were visualized using iTOL v3.6.1 tool [61].

Results and discussion

Sample processing, DNA sequencing and assembly

Total DNA extracted from fifth instar larvae of *A. assamensis* was found to be optimal (1363.23 ng/μl concentration by NanoDrop spectrophotometer and 234.36 ng/μl concentration by Qubit fluorometer) for mitochondrial DNA enrichment. The mitochondrial DNA was enriched from total DNA for library preparation. The library profile showed that size of mitogenome fragments were in the range of 180 to 880 bp. However, total insert size distribution was from 300 to 1000 bp as it involved ~120 bp adapters along with mitogenome fragments. In addition to the proper distribution of fragments, their concentrations (~27.2 ng/μl) were also found to be adequate for sequencing (S1 Fig). The sequencing resulted in 32,12,599 total number of raw reads, out of which around 28,69,153 high quality reads were processed for scaffold preparation. These reads were used for mitogenome scaffold preparation through *de novo* assembly of sequenced contigs using SPAdes-3.5.0, SSPACE and CAP3 program. Finally, a scaffold of 15,272 bp length was obtained which represents the whole mitogenome sequence of *A. assamensis*.

Genome annotation, visualization and comparative analysis

The annotated mitochondrial genome of *A. assamensis* appeared to be a closed circular structure comprising of 37 genes (13 PCGs, 22 tRNAs and 2 rRNAs) along with a non-coding control region over its 15,272 bp length. 23 genes were found to be encoded by the majority strand or J-strand (+) and the remaining 14 genes by the minority strand or N-strand (-). The whole mito-map constructed using BRIG along with its characteristics like gene location and arrangement are shown in Fig 2. The full annotated mitogenome sequence and SRA data of *A. assamensis* submitted to NCBI GenBank are available under the accession numbers KU379695 and SRR3948351, respectively.

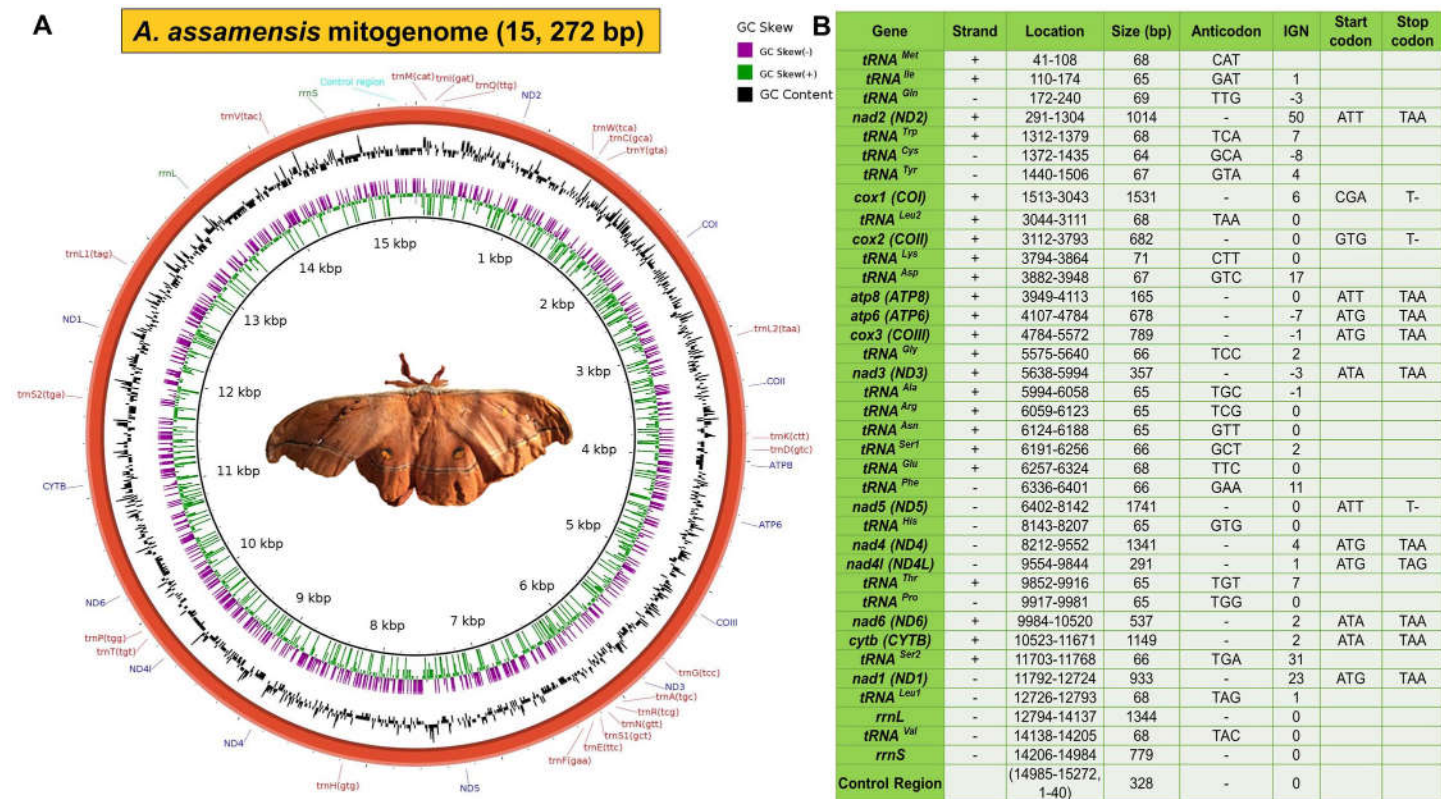


Fig 2. *A. assamensis* mitochondrial genome. (A) Circular map (B) Characteristic features of the mitogenome. *tRNAs* are represented as *trn* followed by the IUPAC-IUB single letter amino acid codes e.g. *trnM* denote *tRNA^{Met}*. IGN represents (+) values as intergenic nucleotides and (-) values as overlapping regions.

<https://doi.org/10.1371/journal.pone.0188077.g002>

Comparative mitogenome analysis. In order to study the conservation pattern, features like genome size, gene content, genome organization, structure, rearrangement and sequence of *A. assamensis* mitogenome were compared with other lepidopterans belonging to various levels of taxonomic classification (Table 1).

We found that the length of *A. assamensis* mitogenome was smaller than that of other sequenced Bombycoids and just bigger than that of *Actias selene* (15,236 bp) and *A. artemis* (15,243 bp) [27, 62]. Notably, it lies within the characteristic size of most of the insects (14 to 20 kb). In insects, mitogenome size variation may be attributed to variation in non-coding regions especially the control region that shows great differences in length as well as pattern. It collectively leads to higher degree of gene rearrangements. In contrast, PCGs are quite stable in the mitogenome [23–24]. Smaller mitogenome size in *A. assamensis* may be favored to reduce accumulation of slightly deleterious mutations during its evolution.

The existence of a typical gene content i.e. 37 genes and a control region was evident in *A. assamensis* mitogenome as observed in other sequenced mitogenomes of lepidopteran insects. The arrangement of mitochondrial genes was found to be in the order identical to the selected Bombycoid insects (Fig 3). Three typical rearrangements were observed in the mitogenome of *A. assamensis* similar to most of the other lepidopteran insects [35, 37, 44, 45] as shown in Fig 3. These were (i) *tRNA^{Met-Ile-Gln}* (*trnM/trnI/trnQ* or *MIQ*) cluster, (ii) *tRNA^{Lys-Asp}* (*K-D*) cluster and (iii) *tRNA^{Ala-Arg-Asn-Ser1-Glu-Phe}* (*ARNS1EF*) cluster. The typical *MIQ* cluster rearrangement located between the control region and *nad2* gene differed from the ancestral pattern *tRNA^{Ile-Gln-Met}* (*trnI/trnQ/trnM* or *IQM*) as reported in two ghost moths (Fig 3) belonging

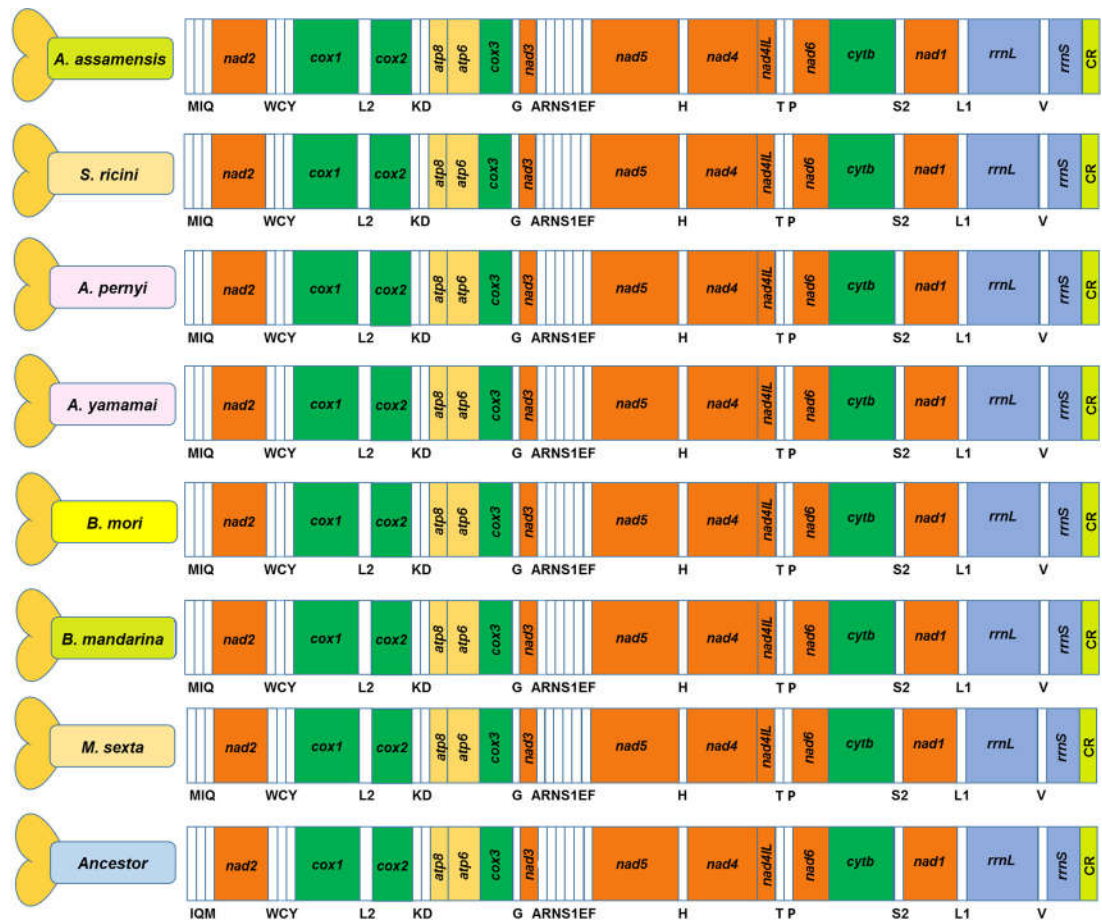


Fig 3. Mitogenome arrangement of *A. assamensis* with respect to other selected species. Ancestor here denotes *Thitarodes renzhienensis*.

<https://doi.org/10.1371/journal.pone.0188077.g003>

to Hepialidae family—a non-ditrysian lineage (Lepidoptera: Exoporia: Hepialoidea) [63–64]. This suggests that *A. assamensis* along with other lepidopteran insects have evolved with a typical gene arrangement after splitting from its stem lineage during the course of time. A typical *ARNS1EF* arrangement is found in most of the lepidopteran insects. However, some rearrangement in this cluster has been reported in few lepidopteran insects different from *A. assamensis* [65–66]. Further, sequence similarity performed among selected Bombycoidea showed that *A. assamensis* shared highest similarity with *Antheraea* species (*A. yamamai*—92.1%, *A. pernyi*—91.7%) followed by the other Bombycoidea like *S. ricini* (88.9%), *M. sexta* (85.5%), *B. mori* (83.4%) and *B. mandarina* (82.9%) indicating a significant level of homology among Bombycoidea superfamily of Lepidoptera.

Comparative analysis among protein coding genes (PCGs). The mitogenome of *A. assamensis* comprised of a total of 13 PCGs which spanned 11,208 bp constituting around 73.38% of the total mitogenome. The total size of PCGs and their individual sizes were found to be similar to the other Bombycoidea species (S1 Table). The PCGs were distributed over both the strands of double stranded mitogenome, 9 genes in the J-strand and 4 genes in the N-strand.

The ATN sequence was found to be the start codon for all the genes in *A. assamensis* similar to other Bombycoidea insects [24, 25, 26, 31, 33]. These codons include ATT (*nad2*, *atp8*, *nad5*), ATA (*nad3*, *nad6*, *cytb*) and ATG (*atp6*, *cox3*, *nad4*, *nad4l*, *nad1*) (Fig 2B). Interestingly, start

codons of *cox1* and *cox2* genes in *A. assamensis* were found to deviate from the canonical ATN initiation codon pattern. Similarly, *A. pernyi* [25], *A. yamamai* [26], *S. boisduvalii* [28], *B. mandarina* [30], *B. mori* [31] and *M. sexta* [33] were also reported to diverge from this pattern. In addition to the gene sequence similarity analysis, comparative amino acid sequence alignment of the reported lepidopteran insects confirmed CGA and GTG as the initiation codons for both *cox1* and *cox2* genes of *A. assamensis*, respectively. CGA is also reported as the start codon for *cox1* gene in some of the members of Lepidoptera such as *A. atlas* [23], *E. pyretorum* [29], *M. sexta* [33], *A. honmai* [45], *Maruca vitrata* [67] and dipteran insect such as *Drosophila melanogaster* [68]. On the other hand, GTG is the start codon for *cox2* gene in *S. ricini* [24], *S. boisduvalii* [28] and *E. pyretorum* [29].

Further, three genes *cox1*, *cox2* and *nad5* showed incomplete termination codon (only T) instead of canonical termination codon pattern TAA or TAG (Fig 2B). This is similar to *A. pernyi* which also shows incomplete termination codon in *cox1*, *cox2* and *nad5* genes [25]. The existence of incomplete stop codon has been observed as a common phenomenon of mitochondrial genes of metazoans like lepidopterans in order to minimize IGS and OS [25, 26, 31, 33, 49]. Several studies reveal that it could be a recognition site for an endonuclease that truncates the polycistronic pre-mRNA. These incomplete codons are rectified by the post-transcriptional polyadenylation to yield functional stop codon with TAA termini [33].

The PCGs of *A. assamensis* were also compared with seven different organisms belonging to various levels of taxonomic classification in Insecta class at nucleotide level and corresponding amino acid level in order to determine the conserved, variable and phylogenetically informative sites (S2 Table). Various species of *Antheraea* showed 84–90% conserved sites at nucleotide level while at amino acid level it was comparatively higher (86–98%). This might be due to synonymous substitution where a non-beneficial mutation is purified by same amino acid encoding codons. However, *atp8* gene exhibited lower values (76.8%) at amino acid level which may suggest that mutation in the gene was needed to help the organism in its evolution. Various comparative analyses were further carried out for the organisms of different genus, family, superfamily and order. The conserved sites were found to decrease with higher hierarchical organisms while variable and phylogenetically informative sites increased in number. Among all PCGs, *cox1* and *cox2* genes showed maximal conservation both at nucleotide and amino acid levels across the organisms of various classification levels. Similarly, *M. sexta* and *P. maraho* were also reported to exhibit higher conservation for *cox1* and *cox2* genes [33, 40].

The study of nucleotide variation pattern (transition-transversion ratio) among the PCGs of *A. assamensis* with respect to various members (*A. pernyi*, *A. yamamai*, *S. ricini*, *M. sexta*, *B. mori* and *B. mandarina*) of Bombycoidea superfamily revealed that the transition-transversion ratio decreased with distant organisms in comparison with closer ones across all the PCGs [69]. For example, ratio for *atp8* gene varied from 1.7 to 0.1 when *A. assamensis* was compared with its close relative *A. pernyi* (same genus) to distant organism *M. sexta* (different family). However in case of *nad4l* gene, a lower ratio was observed for *A. pernyi* (0.24) than distant ones. Further comparative analysis of all the sequenced species of *Antheraea* genus was carried out in order to find out variation in *nad4l* gene within the genus. The *nad4l* gene of *A. assamensis* showed lower ratio across the species whereas higher ratio was observed in comparison with other species. Interestingly, it exhibited higher ratio with *S. ricini* which belongs to different genus of same family. This apparently depicts that *nad4l* gene is a decisive marker for evolution to form *A. assamensis* within *Antheraea* genus. Further, we may hypothesize that *A. assamensis* might have acquired gene from *S. ricini* during its evolution as both the organism are found in Northeast India and share the same evolutionary space [70].

Different evolutionary rates among the genes reflect different functional constraints affecting the mitogenome that varies among species. Therefore, in order to determine the

evolutionary rates in *A. assamensis*, the PCGs were compared with that of seven different organisms belonging to various hierarchy levels in Insecta class. The average rate of synonymous substitutions (Ks) and the average rate of non-synonymous substitutions (Ka) along with their ratios (Ka/Ks) were calculated for all 13 PCGs. The ratio Ka/Ks less than, greater than and equal to 1 indicates that genes are under negative (purifying) selection, positive (adaptative) selection and neutral evolution, respectively [71]. Among 13 PCGs, *atp8* gene encoding ATPase subunit 8 exhibited 1.5 ratio with reference to that of *D. melanogaster* indicating positive/ relaxed selection acting on this gene (Fig 4). This indicates that mutation in *atp8* of *A. assamensis* was due to a necessity to re-organize its structure. These changes may be due to the requirement of energy for the production and secretion of silk fibers which are not observed in *D. melanogaster*. It depicts that functional change in an organism needs commitment mutation in responsible genes in order to support survival of the fittest. Notably, the ratio for remaining PCGs of *A. assamensis* with all organisms was found to be less than one which suggests that mutation was against the requirement and hence mutation was replaced by synonymous nucleotides. This suggests that all the protein coding genes evolved under strong purifying (negative) selection as is evident from the number of conserved and variable sites. The Ka/Ks ratio in *atp8* with reference to *B. mandarina*, *B. mori* and *M. sexta* was found to be significantly higher. These organisms belong to different family with respect to *A. assamensis* and hence significant variation in *atp8* gene can be expected due to certain changes in the mitogenome kinetics. As *S. ricini*, *A. yamamai*, *A. pernyi* belong to the same family of *A. assamensis*, significant Ka/Ks ratio was not observed across the PCGs. Overall comparison of PCGs exhibited that mutation in *atp8* gene was essential in insects to yield several strains with different kinetics as other PCGs showed lower ratios. The lowest evolutionary rates were observed for *cox1* gene indicating that it is least susceptible to variation in protein sequence and hence shows potential as a barcode for evolutionary studies in silkworms. In addition, other genes like *cox2*, *cox3*, *cytb* and *nad* genes with slightly low evolutionary rates next to *cox1* can also serve as barcode markers. Furthermore, the effect of GC content on Ka/Ks ratio was studied and we found a significant negative correlation between them (S2 Fig). This indicates

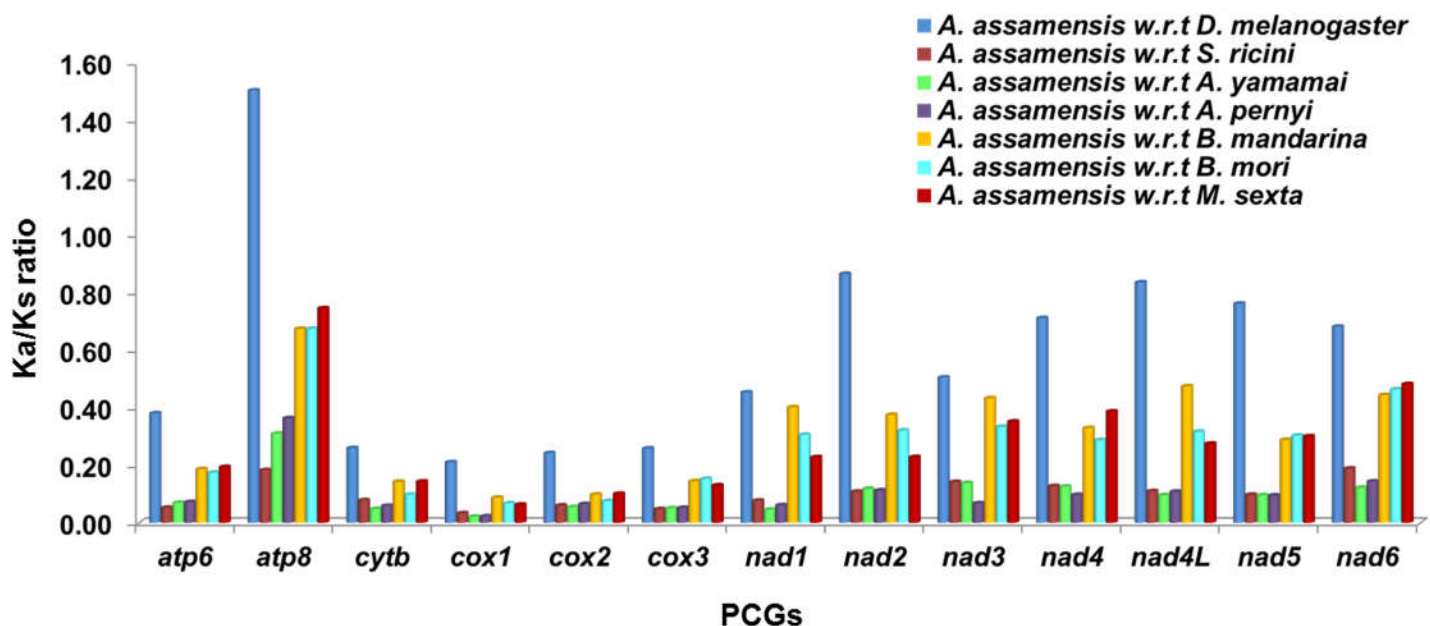


Fig 4. Evolutionary rates (Ka/Ks) of individual PCG of *A. assamensis* with selected species. w.r.t. here denotes with reference to-.

<https://doi.org/10.1371/journal.pone.0188077.g004>

that change in GC content may result in the variation in nucleotide substitution pattern or evolutionary pattern among the PCGs.

Comparative analysis among transfer RNAs (tRNAs). Twenty two tRNA genes with a total length of 1465 bp were found to be present in the mitogenome of *A. assamensis*. 14 tRNA genes were encoded by the J strand while 8 tRNA genes were encoded by the N-strand (Fig 2B). The individual length of each gene varied from 64 to 71 bp as reported for other lepidopterans (S1 Table).

The secondary structures of tRNAs of *A. assamensis* exhibited a typical clover-leaf structure similar to other lepidopteran species (S3 and S4 Figs). This depicts that arrangement of all the tRNA genes is mostly conserved among insects. However, some variation in the structures of *A. assamensis* was observed such as aminoacyl acceptor stem of $tRNA^{Met}$, TCC loop of $tRNA^{Ile}$, $tRNA^{Cys}$, $tRNA^{Tyr}$, $tRNA^{Arg}$, $tRNA^{Asn}$, $tRNA^{Glu}$, $tRNA^{Phe}$, $tRNA^{Trp}$, $tRNA^{Ser2}$, dihydrouridine (DHU) arm of $tRNA^{Ser1}$, etc. The lack of stable DHU arm in $tRNA^{Ser1}$ and TCC loop in $tRNA^{Tyr}$ has also been reported in lepidopterans such as *S. ricini*, *A. yamamai*, *A. pernyi*, *B. mori*, *B. mandarina*, *M. sexta*, *E. pyretorum*, *A. fylloides*, *B. panterinaria*, *E. autonoe*, etc. [24, 25, 26, 29, 30, 31, 33, 35, 38, 42]. These structural variations might be attributed to variation in the number of base pairs responsible for the formation of aminoacyl acceptor (AA) stem, DHU arm, TCC arm and anticodon (AC) stem-loop (S5 and S6 Figs). Among various structural parts of tRNAs, AA stem of *A. assamensis* was found to be more consistent in size across all the tRNAs and in other lepidopterans as observed from comparative analysis. AC stem, AC loop and DHU stem also showed consistency in the number of base pairs across lepidopterans. However, the consistency of these stems and loops varied with the type of tRNAs. In addition, anticodons of tRNAs were found to be identical to the selected Bombycoids (S3 Table) and majority of other lepidopteran insects [26, 28, 30, 35]. On the contrary, T Ψ C stem, T Ψ C loop and DHU loop exhibited randomness in the number of base pairs across the organisms and tRNAs. For instance, T Ψ C loop was absent in $tRNA^{Tyr}$ of *A. assamensis* while $tRNA^{Trp}$ of *A. assamensis* and $tRNA^{Tyr}$ of *B. mori* had 7 bp and 5 bp loops, respectively. The consistency in the number of base pairs of AA stem, AC stem, AC loop and DHU stem was probably to avoid dysfunctioning of tRNAs actively participating in protein synthesis and thus critical for the survival in the evolutionary milieu.

We identified a total of 35 mismatched base pairs in tRNAs of *A. assamensis* scattered over the AA stem, TCC stem, AC and DHU regions of 19 tRNAs. 28 mismatches were observed in G-U combinations while 6 and 1 mismatches were found to be in U-U and G-A combinations, respectively (S3 Table). These mismatches lead to changes in tRNA structures; for instance, two U-U mismatches in $tRNA^{Ser2}$ resulted in the formation of an extra loop in AC stem.

From the comparative analysis, we found that the number of total mismatches was higher in *A. assamensis* (35 No.s) than the selected Bombycoids like *S. ricini* (32 No.s), *B. mori* (29 No.s), *M. sexta* (25 No.s), etc. Numerically, this mismatch was higher than many Saturniids like *E. pyretorum* (24 No.s), *S. cynthia* (28 No.s), etc. Similarly, mismatches such as G-U, U-U, G-A, A-A, A-C and C-U have been previously reported for many lepidopterans like *E. autonoe* (26 No.s), *A. fylloides* (27 No.s), etc. [38, 42]. Among various mismatches in *A. assamensis*, G-U mismatches were found to be the highest (28 No.s) than the selected Bombycoids and other lepidopterans e.g. *A. atlas* (25 No.s) and *E. autonoe* (20 No.s). The details of mismatches found in various stem-loop structures of different tRNAs of Bombycoid members are compared in S3 Table. The non-canonical pairing might occur due to the insertion/deletion of nucleotides or nucleotide substitution of bases within the tRNA genes as a form of ancient insertional editing. The presence of high number of mismatches indicates higher nucleotide substitution in the tRNA sequences of *A. assamensis* as compared to the selected Bombycoids which may result in genome structure and sequence evolution. These mismatches are likely to be corrected by RNA editing mechanism as observed in other arthropods [72].

Comparative analysis among ribosomal RNAs (rRNAs). *A. assamensis* comprised of two genes for rRNA coded by (–) strand of the mitogenome. *rrnL* (*lrRNA*) is located between *tRNA^{Leu1}* and *tRNA^{Val}* genes, whereas *rrnS* (*srRNA*) is accommodated between *tRNA^{Val}* and control region as reported for other lepidopterans. The lengths of *rrnL* and *rrnS* were 1344 bp and 779 bp, respectively (S1 Table) and thus similar to many reported lepidopteran insects [38, 43]. The sequence homology of two rRNA genes of *A. assamensis* with the selected Bombycoids revealed that the sequence identity of *rrnS* gene (84 to 95%) was higher than that of *rrnL* (80 to 89%). The divergence study in rRNA genes (*rrnL* and *rrnS*) also exhibited low variability sites which suggest that these are highly conserved sites and may have potential application in molecular systematics. This may be due to the natural effort to preserve the structure and function of rRNA as protein synthesis depends on it.

Comparative analysis among control region (A+T rich region). The control region is the longest non-coding region of *A. assamensis* mitogenome and with a length of 328 bp, is similar to other Bombycoids. This region is located between *rrnS* and *MIQ* cluster. The comparative sequence homology of control region displayed higher similarity with its close relatives such as *A. yamamai* (89.2%) and *A. pernyi* (82.7%) while showed lower similarity with distant organisms such as *B. mori* (~68%). It was found that two sequence blocks of more than 50 bp in length displayed 96–100% similarity with other *Antheraea* species while *S. ricini* (78–96%), *M. sexta* (68–88%) and *B. mori* (73–80%) showed low similarity. Additionally, some conserved structures and repeats were observed in the control region of *A. assamensis* (Fig 5). These were ‘ATAGA’ motif, poly T stretch (T)₁₉, microsatellite (TA)₉, repeat sequence, poly-A region and a 6 bp deletion. ‘ATAGA’ motif and (T)₁₉ are known to be important for gene regulation and serve as a recognition site for replication initiation of minor or light strand. The poly A tail has been proposed to be required for RNA maturation and serve as a sequence for controlling transcription or replication initiation in insects [73]. The microsatellites (TA)₉ also known as simple sequence repeats (SSR) are proposed to be used as molecular markers due to their abundance and highly polymorphic nature [24]. A microsatellite (AT)_n element is a well conserved site and can be used for evolutionary and conservation genetic studies [25, 26, 74]. The control region of *A. assamensis* lacked tandem repeat elements similar to other completely sequenced *Antheraea* species except *A. pernyi* which harbored six 38 bp tandem repeats. The significance of these tandem repeats is unclear and needs further taxon sampling and mitogenome characterization. Presence of highly conserved yet distinct consensus sequences in the control region thus shows its high potential as a phylogenetic marker at lower taxonomic levels.

Comparative analysis among overlapping sequence (OS) and intergenic spacer (IGS) regions. The overlapping sequences (OS) and intergenic spacers (IGS) are commonly found in the mitogenome of lepidopterans. These regions may vary in length and location from species to species during their evolution. In case of *A. assamensis*, six OS and seventeen IGS were found to be present in the mitogenome. The size of OS and IGS were in the range of 1–8 bp (cumulative length of 23 bp) and 1–50 bp (cumulative length of 171 bp), respectively. The presence of OS, IGS and their corresponding lengths were similar to many of the reported lepidopterans [23, 27, 32, 43, 45] as given in S4 Table. A characteristic conserved overlapping junction (ATGATAA) was observed in between *atp8* and *atp6* genes as reported for most of the lepidopterans [38, 42, 45].

The spacer between the genes *tRNA^{Gln}* and *nad2* (*tRNA^{Gln}*—*nad2* spacer) was found to be the largest (50 bp) in *A. assamensis* but slightly shorter than that observed in *A. yamamai* (53 bp), *S. cynthia* (54 bp), etc. Similarly, it has been detected as the largest spacer in many other lepidopterans like *A. atlas*, *A. selene*, *S. boisduvali*, *E. pyretorum*, *P. bremeri* and *A. honmai* [23, 27, 28, 29, 41, 45]. Further, the spacers *cytb*-*tRNA^{Ser2}* (31 bp), *tRNA^{Ser2}*-*nad1* (23 bp), *tRNA^{Lys}*-

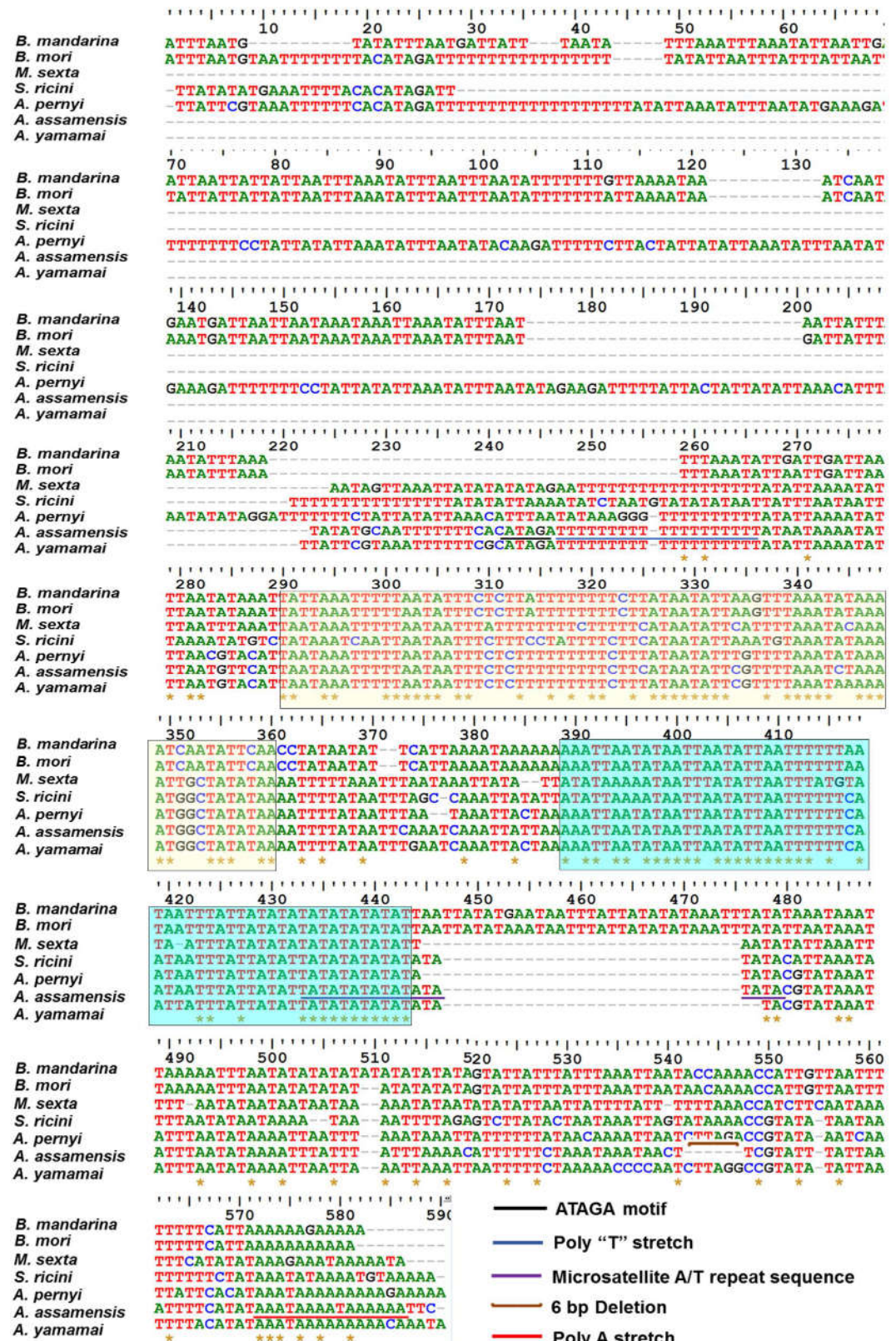


Fig 5. Multiple sequence alignment of control region of *A. assamensis* mitogenome showing conserved structural elements. Yellow and green highlighted blocks represent 1st and 2nd conserved blocks, respectively.

<https://doi.org/10.1371/journal.pone.0188077.g005>

tRNA^{Asp} (17 bp), etc. were also found to exist in *A. assamensis* mitogenome (Fig 2B). The spacer *cytb-tRNA^{Ser2}* may serve as a binding site for the mtTERM, which is a transcription termination peptide [24, 26, 33]. A conserved motif 'ATACTAA' was found in the spacer *tRNA^{Ser2}-nad1* on the comparison of *A. assamensis* with selected Bombycoid species. This spacer is implicated in mitochondrial transcription termination where the 'ATACTAA' motif is essential as a recognition site for mtTERM [24, 26, 33].

The spacers in *A. assamensis* mitogenome also exhibited sequence homology with their adjacent genes as observed in many of the lepidopteran insects [24–26]. For instance, spacer *tRNA^{Gln}-nad2* showed 68% similarity with its adjacent *nad2* gene. This might be because of the partial duplication of *nad2* gene in order to provide another origin of replication [26, 28, 33]. This also suggests that *A. assamensis* may have undergone rapid sequence divergence attributed to the non-coding nature of IGS as seen in other organisms [26]. Apart from adjacent genes, IGS also showed sequence homology with their distant genes in *A. assamensis*. For example, spacer *tRNA^{Lys}-tRNA^{Asp}* showed homology with *rrnS* (82%), *nad5* (76%) and *atp8* (82%) genes. Similarly, the same spacer in *S. ricini* has also been reported to exhibit homology with *rrnS* (78%) and *nad5* (72%) genes [24]. This might be attributed to gene duplication and degeneration.

Comparative analysis with respect to nucleotide composition, skewness and codon usage. The nucleotide composition in mitogenome of *A. assamensis* was analyzed in terms of specific nucleotide content, AT and GC skewness. Whole mitogenome nucleotide composition revealed 80.2% AT content (40.8% T + 39.4% A) and remaining 19.8% GC content (12.1% C + 7.7% G). Similar to *A. assamensis*, many lepidopteran insects also showed enhanced AT content [37, 39]. This AT content was found to vary with different regions of the mitogenome. The AT content was maximal (~91.2%) in the control region followed by rRNAs genes (84.3%), tRNAs genes (80.8%) and PCGs (78%). The individual PCGs also exhibited variation in the corresponding AT content. For instance, *atp8*, *nad4l* and *nad6* genes had the highest AT content while *cox1* and *cox3* genes had lowest values. This variation pattern was also detected in the other lepidopteran insects like *E. pyretorum*, *P. atrilineata*, *G. menyuanensis*, etc. [29, 36, 43]. A close analysis of the third codon position in all 13 PCGs of *A. assamensis* exhibited AT biasness which was 93% than the first (73%) and second (70%) codon positions similar to the other Bombycoids [24, 25, 29]. This may be attributed to more relaxed constraints on A+T content at 3rd position than at 1st and 2nd position due to the phenomena of degeneracy in genetic code.

Nucleotide skewness was determined to measure the relative number of Gs to Cs and As to Ts. The AT skewness in *A. assamensis* mitogenome was found to be negative (-0.02) which indicates the occurrence of more Ts than As and lies within the range of Saturniid moths [25–26]. This value differed from the selected Bombycidae species having positive AT skew values e.g. 0.06 in *B. mori*, *B. mandarina*, etc. [30–31]. Similarly, mitogenome GC skewness also exhibited negative value (-0.22) as observed in many Bombycoids such as *M. sexta* (-0.18) and *S. ricini* (-0.23). The GC skewness varied with major and minor strands of PCGs, however, no significant change was observed in AT skewness. Minor strand was G-skewed i.e. rich in G nucleotides while more 'C's were encoded by major strand (S5 Table). The strand asymmetry is common in insects and has been reported in all the lepidopteran insects which occur due to asymmetrical mutation pressure on mitogenome. These skewness patterns were also observed in rRNAs, tRNAs, control region and are similar to the other sequenced lepidopteran insects.

Nucleotide biasness was also reflected in the codon usage of PCGs that determines the frequency of synonymous codon usage (SCU) in various genomes at inter and intra levels. Based on the codon usage and relative SCU results, codons with A or T nucleotides at the 3rd codon position were highly used in comparison with G or C nucleotides as discussed above. The five

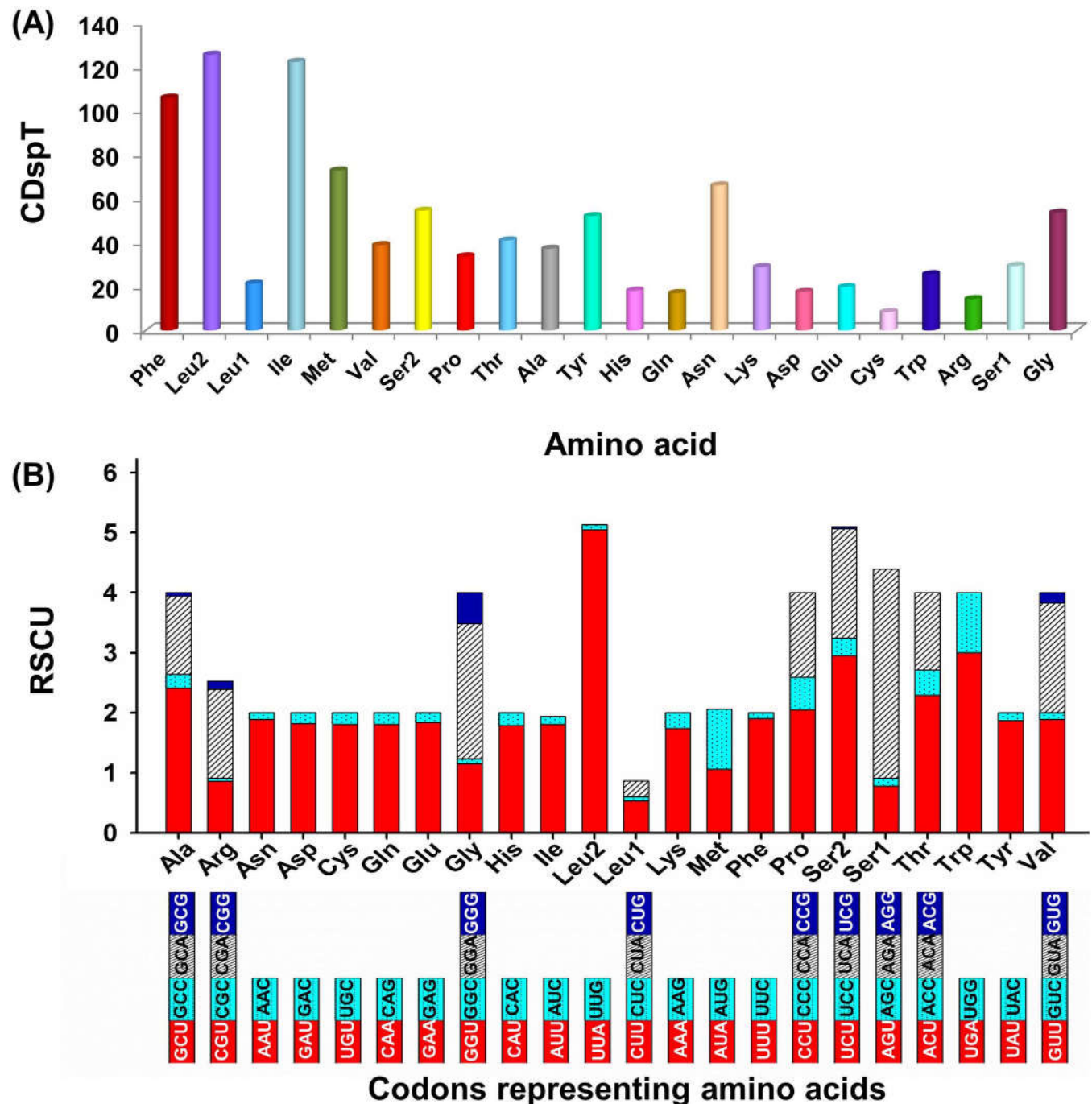


Fig 6. Codon usage in *A. assamensis* mitogenome. (A) Codon distribution where CDspT = codons per thousand codons (B) Relative synonymous codon usage (RSCU). Termination codons were excluded in the study.

<https://doi.org/10.1371/journal.pone.0188077.g006>

most frequently used amino acids were Leu, Ile, Phe, Met and Asn (Fig 6) and the most prevalent codons for these corresponding amino acids were UUA, AUU, UUU, AUA and AAU with a usage frequency of 45.53% to 48.95% (S2 Fig). Cys was found to be the least used amino acid as detected in other selected Bombycoidea. Biasness in SCU has been observed in many other lepidopterans to avoid errors due to mis-incorporation of amino acids which depends

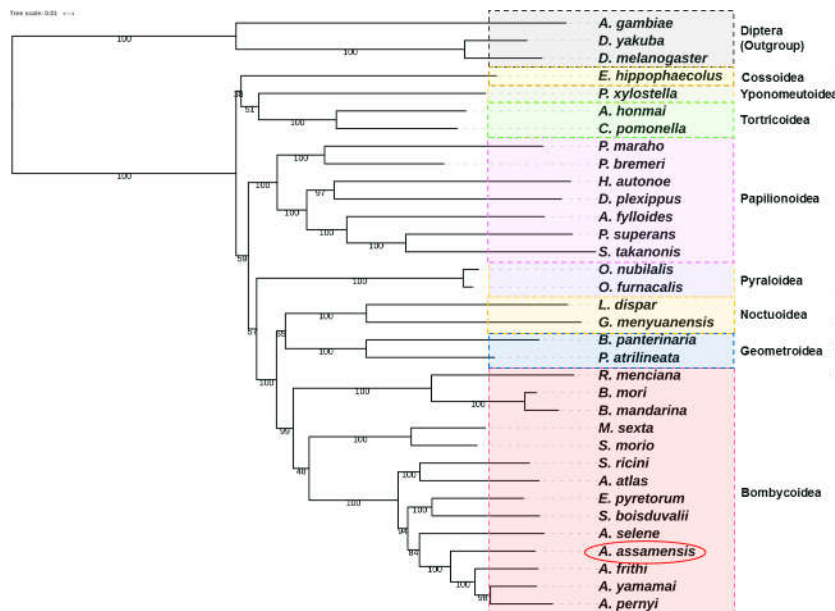


Fig 7. Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms using maximum likelihood (ML) method in RaxmlGUI v1.3 (1000 bootstrap replicates). The tree is drawn to scale with bootstrap values indicated along with the branches.

<https://doi.org/10.1371/journal.pone.0188077.g007>

upon various factors like natural selection (gene length and function), mutation bias (base mutation and GC content), etc. [24, 35]. This A+T biasness may result in the change in amino acid composition similar to the other lepidopterans. Therefore, the codon usage analysis will have great significance in studying gene expression and evolutionary studies in silkworms.

Comparative phylogenetic analysis. Maximum likelihood (ML) and Bayesian Inference (BI) phylogenetic trees were constructed based on nucleotide sequences of 13 PCGs and 13 PCGs+2 rRNAs datasets, respectively (Figs 7 and 8, S7 and S8 Figs). The phylogenetic trees obtained using both the methods were found to be of similar topologies. The tree clustering showed that *A. assamensis* belongs to Saturniidae family which is clustered in the same branch with other families (Bombycidae+Sphingidae) of Bombycoidea superfamily within the order Lepidoptera. Within the Saturniidae family, *A. assamensis* has clustered around with other *Antheraea* species and was also close to *A. selene*. Further, our study supported the triplet superfamily (Bombycoidea+Geometroidea+Noctuoidea) relationship under Macroheterocera which along with Pyraloidea and Papilionoidea formed Obtectomera, a group within Apoditrysia [8–9]. These results were consistent with the previously reported studies [26, 32, 35, 75]. More mitogenome sequence information on Bombycoidea superfamily is needed to gain deep insights into the families/superfamilies relationships within Lepidoptera. The present study supports the fact that intra and inter family level relationships are well resolved within the superfamily Bombycoidea providing a better evolutionary relationship among the silk producing insects.

Conclusion

Our study deciphered the complete mitogenome of *A. assamensis* which shares similarity with majority of the lepidopterans, particularly Saturniids, in several characteristics such as genome organization and content, PCG size and structure, AT/GC skewness, tRNA structure and anti-codons, OS and IGS regions. Our analysis indicates that *A. assamensis* PCGs evolved under

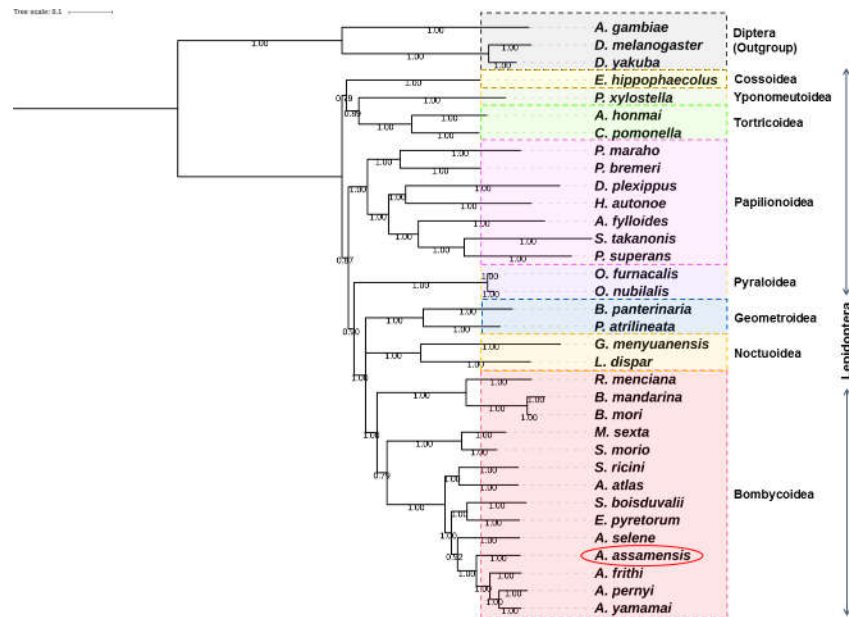


Fig 8. Phylogenetic tree inferred from nucleotide sequences of 13 PCGs of 34 organisms using bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

<https://doi.org/10.1371/journal.pone.0188077.g008>

strong purifying selection and tRNAs genes showed high base substitution/mismatches. Transition-transversion ratio suggests that *nad4l* gene may be the significant contributor for evolution to form *A. assamensis* within *Antheraea* genus. A six-bp indel (deletion) and two highly conserved sequence blocks in control region suggest its potential application as a marker for delineating other closely related *Antheraea* sp. and subsequent usage in phylogenetic studies at lower taxonomic levels. The largest *tRNA^{Gln}-nad2* spacer found in the *A. assamensis* mitogenome may serve as another origin of replication. Our study assigns the taxonomic status of *A. assamensis* using an optimal model based phylogenetic tree construction. This is the 12th representative organism of Saturniidae family with a completely sequenced mitogenome. As *A. assamensis* is the sole producer of unique golden Muga silk which is the backbone of Assam's sericulture industry, our study on its mitogenomic landscape is an important addition to the existing genome informatics resources on silkworms.

Supporting information

S1 Fig. Bioanalyzer profiles after sonication of enriched mitochondrial DNA sample and libraries.

(TIF)

S2 Fig. (A) G+C content versus Ka/Ks ratios of 13 concatenated PCGs in *A. assamensis* mitogenome (B) The usage rate of five frequently used codons in Bombycoidea members.

(TIF)

S3 Fig. Secondary structures of tRNAs of *A. assamensis* from *trnM* to *trnR*. *trnY* lacked TCC loop. *tRNA* here are represented as *trn*. AA denotes amino acyl arm and AC denotes anticodon arm.

(TIF)

S4 Fig. Secondary structures of tRNAs of *A. assamensis* from trnN to trnV. *trnS1* lacked DHU loop. tRNA here are represented as *trn*. AA denotes amino acyl arm and AC denotes anticodon arm.

(TIF)

S5 Fig. Variation in the secondary structures of trnM, trnI, trnC, trnY, trnR and trnN in *A. assamensis* and selected Bombycoids. AA denotes Aminoacyl arm and AC denotes Anticodon arm. S-L denotes Stem-Loop of tRNA gene. tRNA is represented as *trn*.

(TIF)

S6 Fig. Variation in the secondary structures of trnS1, trnE, trnF and trnT in *A. assamensis* and selected Bombycoids. AA denotes Aminoacyl arm and AC denotes Anticodon arm. S-L denotes Stem-Loop of tRNA gene. tRNA is represented as *trn*.

(TIF)

S7 Fig. Phylogenetic tree inferred from 13PCGs+2rRNAs of 34 organisms using maximum likelihood (ML) method in RaxmlGUI v1.3 (1000 bootstrap replicates). The tree is drawn to scale with bootstrap values indicated along with the branches.

(TIF)

S8 Fig. Phylogenetic tree inferred from 13PCGs+2rRNAs of 34 organisms using Bayesian inference (BI) method in MrBayes v3.2.6. The tree is drawn to scale with bayesian posterior probability values indicated along with the branches.

(TIF)

S1 Table. Comparative gene lengths in Bombycoids.

(PDF)

S2 Table. Gene by gene divergences in the protein coding genes at different taxonomic levels, both at nucleotide level and amino acid levels. C here denotes conserved sites, I denotes phylogenetically informative sites, V denotes variable sites, N denotes nucleotide level, A denotes amino acid level. The number of sites is represented in terms of percentage values.

(PDF)

S3 Table. Anticodons (AC) of tRNAs and number of mismatches present in Amino-acyl (AA) stem, TCC-stem, DHU-stem and Anticodon (AC) stem of mitochondrial tRNAs of *A. assamensis* with respect to the selected Bombycoids.

(PDF)

S4 Table. Intergenic spacers (IGS) and overlapping sequence (OS) between different genes in *A. assamensis* with respect to the selected Bombycoid species.

(PDF)

S5 Table. AT skewness and GC skewness in the whole mitogenome sequence, protein coding genes (PCGs), tRNAs, rRNAs and control region of *A. assamensis* and the selected Bombycoid species.

(PDF)

Author Contributions

Conceptualization: Deepika Singh, Utpal Bora.

Data curation: Deepika Singh.

Formal analysis: Deepika Singh, Debajyoti Kabiraj.

Investigation: Deepika Singh.

Methodology: Pragya Sharma, Kartik Neog.

Project administration: Utpal Bora.

Resources: Kartik Neog, Utpal Bora.

Supervision: Utpal Bora.

Writing—original draft: Deepika Singh.

Writing—review & editing: Debajyoti Kabiraj, Pragya Sharma, Hasnahana Chetia, Ponnala Vimal Mosahari, Utpal Bora.

References

1. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UC, Podowski RM, et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 1998; 396: 133–140. <https://doi.org/10.1038/24094> PMID: 9823893
2. Salvioli S, Bonafè M, Capri M, Monti D, Franceschi C. Mitochondria, aging and longevity—a new perspective. *FEBS Lett.* 2001; 492: 9–13. [https://doi.org/10.1016/S0014-5793\(01\)02199-8](https://doi.org/10.1016/S0014-5793(01)02199-8) PMID: 11248228
3. Cameron SL. Insect mitochondrial genomics: implications for evolution and phylogeny. *Annu Rev Entomol.* 2014; 59: 95–117. <https://doi.org/10.1146/annurev-ento-011613-162007> PMID: 24160435
4. Boore JL. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol.* 2006; 21: 439–446. <https://doi.org/10.1016/j.tree.2006.05.009> PMID: 16762445
5. Mandal SD, Chhakchhuak L, Gurusubramanian G, Kumar NS. Mitochondrial markers for identification and phylogenetic studies in insects—A Review. *DNA Barcodes* 2014; 2: 1–9. <https://doi.org/10.2478/dna-2014-0001>
6. Arunkumar KP, Metta M, Nagaraju J. Molecular phylogeny of silkworms reveals the origin of domesticated silkworm, *Bombyx mori* from Chinese *Bombyx mandarina* and paternal inheritance of *Antheraea proylei* mitochondrial DNA. *Mol Phylogenet Evol.* 2006; 40: 419–427. <https://doi.org/10.1016/j.ympev.2006.02.023> PMID: 16644243
7. Liu YQ, Li YP, Wang H, Xia RX, Chai CL, Pan MH, et al. The complete mitochondrial genome of the wild type of *Antheraea pernyi* (Lepidoptera: Saturniidae). *Ann Entomol Soc Am.* 2012; 105: 498–505. <https://doi.org/10.1603/AN11156>
8. Timmermans MJTN, Lees DC, Simonsen TJ. Towards a mitogenomic phylogeny of Lepidoptera. *Mol Phylogenet Evol.* 2014; 79: 169–178. <https://doi.org/10.1016/j.ympev.2014.05.031> PMID: 24910155
9. Mitter C, Davis DR, Cummings MP. Phylogeny and Evolution of Lepidoptera. *Annu Rev Entomol.* 2017; 62: 265–283. <https://doi.org/10.1146/annurev-ento-031616-035125> PMID: 27860521
10. Nässig WA, Lampe REJ, Kager S. Heterocera Sumatrana: The Saturniidae of Sumatra (Lepidoptera). Vol. 10. Gottingen: Heterocera Sumatrana Society; 1996.
11. Tikader A, Vijayan K, Saratchandra B. Muga silkworm, *Antheraea assamensis* (Lepidoptera: Saturniidae)—an overview of distribution, biology and breeding. *Eur J Entomol.* 2013; 110: 293. <https://doi.org/10.14411/eje.2013.096>
12. Kasoju N, Bhonde RR, Bora U. Fabrication of a novel micro–nano fibrous nonwoven scaffold with *Antheraea assama* silk fibroin for use in tissue engineering. *Mater Lett.* 2009; 63: 2466–2469. <https://doi.org/10.1016/j.matlet.2009.08.037>
13. Chetia H, Kabiraj D, Singh D, Mosahari PV, Das S, Sharma P, et al. De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). *Gene* 2017; 611: 54–65. <https://doi.org/10.1016/j.gene.2017.02.021> PMID: 28216038
14. Chakravorty R, Barah A, Neog K, Rahman SAS, Ghose J. Package of practices for muga culture. Package of practices of Muga, Eri and Mulberry Sericulture for North Eastern Region of India, Central Muga Eri Research & Training Institute, Central Silk Board, Lahdoigarh, Jorhat. 2005: 1–23.
15. Andrews S. FastQC: A quality control tool for high throughput sequence data. 2010. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
16. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012; 19: 455–477. <https://doi.org/10.1089/cmb.2012.0021> PMID: 22506599

17. Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res.* 1999; 9: 868–877. <https://doi.org/10.1101/gr.9.9.868> PMID: 10508846
18. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G. MITOS: Improved *de novo* meta-zoan mitochondrial genome annotation. *Mol Phylogenet Evol.* 2013; 69: 313–319. <https://doi.org/10.1016/j.ympev.2012.08.023> PMID: 22982435
19. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser.* 1999; 41: 95–98. <https://doi.org/10.1021/bk-1999-0734.ch008>
20. Thompson JD, Gibson TJ, Higgins DG. Multiple Sequence Alignment Using ClustalW and ClustalX. *Curr Protoc Bioinformatics* 2002; 2.3.1–2.3.22. <https://doi.org/10.1002/0471250953.bi0203s00> PMID: 18792934
21. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999; 27: 573. <https://doi.org/10.1093/nar/27.2.573> PMID: 9862982
22. Alikhan NF, Petty NK, Zakour NLB, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011; 12: 402. <https://doi.org/10.1186/1471-2164-12-402> PMID: 21824423
23. Chen MM, Li Y, Chen M, Wang H, Li Q, Xia RX, et al. Complete mitochondrial genome of the atlas moth, *Attacus atlas* (Lepidoptera: Saturniidae) and the phylogenetic relationship of Saturniidae species. *Gene* 2014; 545: 95–101. <https://doi.org/10.1016/j.gene.2014.05.002> PMID: 24797615
24. Kim JS, Park JS, Kim MJ, Kang PD, Kim SG, Jin BR, et al. Complete nucleotide sequence and organization of the mitochondrial genome of eri-silkworm, *Samia cynthia ricini* (Lepidoptera: Saturniidae). *J Asia Pac Entomol.* 2012; 15: 162–173. <https://doi.org/10.1016/j.aspen.2011.10.002>
25. Liu Y, Li Y, Pan M, Dai F, Zhu X, Lu C, et al. The complete mitochondrial genome of the Chinese oak silkworm, *Antheraea pernyi* (Lepidoptera: Saturniidae). *Acta Biochim Biophys Sin (Shanghai)* 2008; 40: 693–703. <https://doi.org/10.1111/j.1745-7270.2008.00449.x>
26. Kim SR, Kim MI, Hong MY, Kim KY, Kang PD, Hwang JS, et al. The complete mitogenome sequence of the Japanese oak silkworm, *Antheraea yamamai* (Lepidoptera: Saturniidae). *Mol Biol Rep.* 2009; 36: 1871–1880. <https://doi.org/10.1007/s11033-008-9393-2> PMID: 18979227
27. Liu QN, Zhu BJ, Dai LS, Wei GQ, Liu CL. The complete mitochondrial genome of the wild silkworm moth, *Actias selene*. *Gene* 2012; 505: 291–299. <https://doi.org/10.1016/j.gene.2012.06.003> PMID: 22688122
28. Hong MY, Lee EM, Jo YH, Park HC, Kim SR, Hwang JS, et al. Complete nucleotide sequence and organization of the mitogenome of the silk moth *Caligula boisduvalii* (Lepidoptera: Saturniidae) and comparison with other lepidopteran insects. *Gene* 2008; 413: 49–57. <https://doi.org/10.1016/j.gene.2008.01.019> PMID: 18337026
29. Jiang ST, Hong GY, Yu M, Li N, Yang Y, Liu YQ, et al. Characterization of the complete mitochondrial genome of the giant silkworm moth, *Eriogyna pyretorum* (Lepidoptera: Saturniidae). *Int J Biol Sci.* 2009; 5: 351–365. <https://doi.org/10.7150/ijbs.5.351> PMID: 19471586
30. Pan M, Yu Q, Xia Y, Dai F, Liu Y, Lu C, et al. Characterization of mitochondrial genome of Chinese wild mulberry silkworm, *Bombyx mandarina* (Lepidoptera: Bombycidae). *Sci China C Life Sci.* 2008; 51: 693–701. <https://doi.org/10.1007/s11427-008-0097-6> PMID: 18677597
31. Yukuhiro K, Sezutsu H, Itoh M, Shimizu K, Banno Y. Significant levels of sequence divergence and gene rearrangements have occurred between the mitochondrial genomes of the wild mulberry silkworm, *Bombyx mandarina*, and its close relative, the domesticated silkworm, *Bombyx mori*. *Mol Biol Evol.* 2002; 19: 1385–1389. <https://doi.org/10.1093/oxfordjournals.molbev.a004200> PMID: 12140251
32. Kong W, Yang J. The complete mitochondrial genome of *Rondotia menciiana* (Lepidoptera: Bombycidae). *J Insect Sci.* 2015; 15: 48. <https://doi.org/10.1093/jisesa/iev032> PMID: 25888706
33. Cameron SL, Whiting MF. The complete mitochondrial genome of the tobacco hornworm, *Manduca sexta*, (Insecta: Lepidoptera: Sphingidae), and an examination of mitochondrial gene variability within butterflies and moths. *Gene* 2008; 408: 112–123. <https://doi.org/10.1016/j.gene.2007.10.023> PMID: 18065166
34. Kim MJ, Choi SW, Kim I. Complete mitochondrial genome of the larch hawk moth, *Sphinx morio* (Lepidoptera: Sphingidae). *Mitochondrial DNA* 2013; 24: 622–624. <https://doi.org/10.3109/19401736.2013.772155> PMID: 23452242
35. Yang X, Xue D, Han H. The complete mitochondrial genome of *Biston panterinaria* (Lepidoptera: Geometridae), with phylogenetic utility of mitochondrial genome in the Lepidoptera. *Gene* 2013; 515: 349–358. <https://doi.org/10.1016/j.gene.2012.11.031> PMID: 23220020
36. Yang L, Wei ZJ, Hong GY, Jiang ST, Wen LP. The complete nucleotide sequence of the mitochondrial genome of *Phthonandria atrilineata* (Lepidoptera: Geometridae). *Mol Biol Rep.* 2009; 36: 1441–1449. <https://doi.org/10.1007/s11033-008-9334-0> PMID: 18696255

37. Gong YJ, Wu QL, Wei SJ. The first complete mitogenome for the superfamily Cossioidea of Lepidoptera: The seabuckthorn carpenter moth *Eogystia hippophaecolus*. Mitochondrial DNA 2014; 25: 288–289. <https://doi.org/10.3109/19401736.2013.792071> PMID: 23841606
38. Zhao F, Huang DY, Sun XY, Shi QH, Hao JS, Zhang LL, et al. The first mitochondrial genome for the butterfly family Riodinidae (*Abisara fylloides*) and its systematic implications. Dongwuxue Yanjiu 2013; 34: E109. <https://doi.org/10.11813/j.issn.0254-5853.2013.E4-5.E109> PMID: 24115668
39. Kim MJ, Kang AR, Jeong HC, Kim KG, Kim I. Reconstructing intraordinal relationships in Lepidoptera using mitochondrial genome data with the description of two newly sequenced lycaenids, *Spindasis takanonis* and *Protantigius superans* (Lepidoptera: Lycaenidae). Mol Phylogenet Evol. 2011; 61: 436–445. <https://doi.org/10.1016/j.ympev.2011.07.013> PMID: 21816227
40. Wu LW, Lees DC, Yen SH, Hsu YF. The complete mitochondrial genome of the near-threatened swallowtail, *Agehana maraho* (Lepidoptera: Papilionidae): evaluating sequence variability and suitable markers for conservation genetic studies. Entomol News 2010; 121: 267–280. <https://doi.org/10.3157/021.121.0308>
41. Kim MI, Baek JY, Kim MJ, Jeong HC, Kim KG, Bae CH, et al. Complete nucleotide sequence and organization of the mitogenome of the red-spotted apollo butterfly, *Parnassius bremeri* (Lepidoptera: Papilionidae) and comparison with other lepidopteran insects. Mol Cells 2009; 28: 347–363. <https://doi.org/10.1007/s10059-009-0129-5> PMID: 19823774
42. Kim MJ, Wan X, Kim K, Hwang JS, Kim I. Complete nucleotide sequence and organization of the mitogenome of endangered *Eumenis autonoe* (Lepidoptera: Nymphalidae). Afr J Biotechnol. 2010; 9. <https://doi.org/10.5897/AJB09.1486>
43. Yuan ML, Zhang QL. The complete mitochondrial genome of *Gynaephora menyuanensis* (Lepidoptera: Lymantriidae) from the Qinghai-Tibetan Plateau. Mitochondrial DNA 2013; 24: 328–330. <https://doi.org/10.3109/19401736.2012.760077> PMID: 23350687
44. Dai LS, Zhu BJ, Qian C, Zhang CF, Li J, Wang L, et al. The complete mitochondrial genome of the diamondback moth, *Plutella xylostella* (Lepidoptera: Plutellidae). Mitochondrial DNA Part A 2016; 27: 1512–1513. <https://doi.org/10.3109/19401736.2014.953116> PMID: 25187437
45. Lee ES, Shin KS, Kim MS, Park H, Cho S, Kim CB. The mitochondrial genome of the smaller tea tortrix *Adoxophyes honmai* (Lepidoptera: Tortricidae). Gene 2006; 373: 52–57. <https://doi.org/10.1016/j.gene.2006.01.003> PMID: 16488090
46. Shi BC, Liu W, Wei SJ. The complete mitochondrial genome of the codling moth *Cydia pomonella* (Lepidoptera: Tortricidae). Mitochondrial DNA 2013; 24: 37–39. <https://doi.org/10.3109/19401736.2012.716054> PMID: 22978261
47. Coates BS, Sumerford DV, Hellmich RL, Lewis LC. Partial mitochondrial genome sequences of *Ostrinia nubilalis* and *Ostrinia furnicalis*. Int J Biol Sci. 2005; 1: 13. <https://doi.org/10.7150/ijbs.1.13> PMID: 15951845
48. de Bruijn MH. *Drosophila melanogaster* mitochondrial DNA, a novel organization and genetic code. Nature 1983; 304: 234–241. <https://doi.org/10.1038/304234a0> PMID: 6408489
49. Clary DO, Wolstenholme DR. The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. J Mol Evol. 1985; 22: 252–271. <https://doi.org/10.1007/BF02099755> PMID: 3001325
50. Beard CB, Hamm D, Collins FH. The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. Insect Mol Biol. 1993; 2: 103–124. <https://doi.org/10.1111/j.1365-2583.1993.tb00131.x> PMID: 9087549
51. Tamura K, Stecher G, Peterson D, Filipinski A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 2013; 30: 2725–2729. <https://doi.org/10.1093/molbev/mst197> PMID: 24132122
52. Smith C, Heyne S, Richter AS, Will S, Backofen R, Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. Nucleic Acids Res. 2010; 38: W373–W377. <https://doi.org/10.1093/nar/gkq316> PMID: 20444875
53. Junqueira AC, Lessinger AC, Torres TT, da Silva FR, Vettore AL, Arruda P, et al. The mitochondrial genome of the blowfly *Chrysomya chloropyga* (Diptera: Calliphoridae). Gene 2004; 339: 7–15. <https://doi.org/10.1016/j.gene.2004.06.031> PMID: 15363841
54. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010; 38: W7–13. <https://doi.org/10.1093/nar/gkq291> PMID: 20435676
55. Guindon S, Gascuel O. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. Syst Biol. 2003; 52: 696–704. <https://doi.org/10.1080/10635150390235520> PMID: 14530136

56. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012; 9: 772–772. <https://doi.org/10.1038/nmeth.2109> PMID: 22847109
57. Silvestro D, Michalak I. RaxmlGUI: a graphical front-end for RAxML. *Org Divers Evol*. 2012; 12: 335–337. <https://doi.org/10.1007/s13127-011-0056-0>
58. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 2001; 17: 754–755. <https://doi.org/10.1093/bioinformatics/17.8.754> PMID: 11524383
59. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012; 61: 539–42. <https://doi.org/10.1093/sysbio/sys029> PMID: 22357727
60. Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer v1.6. 2014. Available from: <http://tree.bio.ed.ac.uk/software/tracer/>.
61. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016; 44: W242–245. <https://doi.org/10.1093/nar/gkw290> PMID: 27095192
62. Park JS, Kim MJ, Kim I. The complete mitochondrial genome of the moon moth, *Actias aliena* (Lepidoptera: Saturniidae). *Mitochondrial DNA A DNA Mapp Seq Anal*. 2016; 27: 149–50. <https://doi.org/10.3109/19401736.2013.878918> PMID: 24450726
63. Cao YQ, Ma C, Chen JY, Yang DR. The complete mitochondrial genomes of two ghost moths, *Thitarodes renzhiensis* and *Thitarodes yunnanensis*: the ancestral gene arrangement in Lepidoptera. *BMC Genomics* 2012; 13: 276. <https://doi.org/10.1186/1471-2164-13-276> PMID: 22726496
64. Downton M, Austin AD. Evolutionary dynamics of a mitochondrial rearrangement" hot spot" in the Hymenoptera. *Mol Biol Evol*. 1999; 16: 298–309. <https://doi.org/10.1093/oxfordjournals.molbev.a026111> PMID: 10028295
65. Park JS, Kim MJ, Jeong SY, Kim SS, Kim I. Complete mitochondrial genomes of two gelechioids, *Mesophleps albilinella* and *Dichomeris ustalella* (Lepidoptera: Gelechiidae), with a description of gene rearrangement in Lepidoptera. *Curr Genet*. 2016; 62: 809–826. <https://doi.org/10.1007/s00294-016-0585-3> PMID: 26952721
66. Liu QN, Xin ZZ, Zhu XY, Chai XY, Zhao XM, Zhou CL, et al. A transfer RNA gene rearrangement in the lepidopteran mitochondrial genome. *Biochem Biophys Res Commun*. 2017; 489: 149–154. <https://doi.org/10.1016/j.bbrc.2017.05.115> PMID: 28546004
67. Margam VM, Coates BS, Hellmich RL, Agunbiade T, Seufferheld MJ, Sun W, et al. Mitochondrial genome sequence and expression profiling for the legume pod borer *Maruca vitrata* (Lepidoptera: Crambidae). *PloS One* 2011; 6: e16444. <https://doi.org/10.1371/journal.pone.0016444> PMID: 21311752
68. Stewart JB, Beckenbach AT. Characterization of mature mitochondrial transcripts in *Drosophila*, and the implications for the tRNA punctuation model in arthropods. *Gene* 2009; 445: 49–57. <https://doi.org/10.1016/j.gene.2009.06.006> PMID: 19540318
69. DeSalle R, Freedman T, Prager EM, Wilson AC. Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. *J Mol Evol*. 1987; 26: 157–64. <https://doi.org/10.1007/BF02111289> PMID: 3125333
70. Sarmah MC. Eri pupa: a delectable dish of North East India. *Curr Sci*. 2011; 100: 279.
71. Meiklejohn CD, Montooth KL, Rand DM. Positive and negative selection on the mitochondrial genome. *Trends Genet*. 2007; 23: 259–263. <https://doi.org/10.1016/j.tig.2007.03.008> PMID: 17418445
72. Lavrov DV, Brown WM, Boore JL. A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede *Lithobius forficatus*. *Proc Natl Acad Sci USA* 2000; 97: 13738–13742. <https://doi.org/10.1073/pnas.250402997> PMID: 11095730
73. Zhang DX, Hewitt GM. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochem Syst Ecol*. 1997; 25: 99–120. [https://doi.org/10.1016/S0305-1978\(96\)00042-7](https://doi.org/10.1016/S0305-1978(96)00042-7)
74. Stolle E, Kidner JH, Moritz RFA. Patterns of Evolutionary Conservation of Microsatellites (SSRs) Suggest a Faster Rate of Genome Evolution in Hymenoptera Than in Diptera. *Genome Biol Evol*. 2013; 5: 151–162. <https://doi.org/10.1093/gbe/evs133> PMID: 23292136
75. Liu QN, Xin ZZ, Bian DD, Chai XY, Zhou CL, Tang BP. The first complete mitochondrial genome for the subfamily Limacodidae and implications for the higher phylogeny of Lepidoptera. *Sci Rep*. 2016; 6: 35878. <https://doi.org/10.1038/srep35878> PMID: 27767191



Review

A comprehensive view of the web-resources related to sericulture

Deepika Singh¹, Hasnahana Chetia¹, Debajyoti Kabiraj¹,
Swagata Sharma¹, Anil Kumar², Pragya Sharma³, Manab Deka³ and
Utpal Bora^{1,4,5,*}

¹Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India, ²Centre for Biological Sciences (Bioinformatics), Central University of South Bihar (CUSB), Patna 800014, India, ³Department of Bioengineering & Technology, Gauhati University Institute of Science & Technology, Gauhati University, Guwahati, Assam 781014, India, ⁴Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India and ⁵Mugagen Laboratories Pvt. Ltd, Technology Incubation Centre, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

*Corresponding Author: Tel: +913612582215; Fax: +913612582249; Email: ubora@iitg.ernet.in; ubora@rediffmail.com

Citation details: Singh,D., Chetia,H., Kabiraj,D. *et al.* A comprehensive view of the current web-resources in sericulture and related fields. *Database* (2016) Vol. 2016: article ID baw086; doi:10.1093/database/baw086

Received 21 January 2016; Revised 25 April 2016; Accepted 2 May 2016

Abstract

Recent progress in the field of sequencing and analysis has led to a tremendous spike in data and the development of data science tools. One of the outcomes of this scientific progress is development of numerous databases which are gaining popularity in all disciplines of biology including sericulture. As economically important organism, silkworms are studied extensively for their numerous applications in the field of textiles, biomaterials, biomimetics, etc. Similarly, host plants, pests, pathogens, etc. are also being probed to understand the seri-resources more efficiently. These studies have led to the generation of numerous seri-related databases which are extremely helpful for the scientific community. In this article, we have reviewed all the available online resources on silkworm and its related organisms, including databases as well as informative websites. We have studied their basic features and impact on research through citation count analysis, finally discussing the role of emerging sequencing and analysis technologies in the field of seri-data science. As an outcome of this review, a web portal named SeriPort, has been created which will act as an index for the various sericulture-related databases and web resources available in cyberspace.

Database URL: <http://www.seriport.in/>

Introduction

More than 50 years have passed since the time when the term ‘database’ was coined. However, it was only during the massive digitalization of many resources like archives of music, books, etc. in the 1990s that the same term started reflecting its primary usage in today’s world as a data organizational model (1). During these years, the databases have been empowered to retrieve and filter data in various ways. Integration of these databases with biology has brought digital revolution to life science. The amalgamation of biology with information technology for data dissemination and statistics for data analytics has led to the development of some highly successful databases like GenBank, RCSB Protein Data Bank (PDB), etc. (2). Now, there are databases in almost every field of biology ranging from diseases, whole organisms, taxonomy, bio-active products, etc. making them indispensable for the researchers (3–6). One of the research fields in which databases are being constructed actively is ‘Sericulture’.

Silkworms and their respective host plants are the key players of sericulture and silk is its prime yield. Sericulture has been in practice much prior to the Silk Road era in ancient Indian and Chinese civilizations and helped in the enrichment of human endeavors in art and culture (7). *Bombyx mori*, *Antheraea assamensis*, *A. mylitta* and many other silkworms are responsible for the production of silk varieties like mulberry silk, muga silk, tasar silk, etc. for traditional and commercial usage. Researchers also developed mutants of these organisms for improving silk quality and quantity, understanding their physiology and exploiting them as bioreactors for recombinant proteins (8–10). Similarly, the host plants of silkworms are studied not only due to their importance as a survival requisite for silkworms but also for several unconventional benefits like production of biodiesel, medicinal applications, etc. (11, 12). Apart from these, other members of a silkworm’s ecosystem like pests and pathogens which threaten the existence of the silkworms are also researched for the development of treatment or pest-control strategies, host–pathogen interaction studies, etc. (13, 14).

The need of databases in sericulture field cannot be emphasized more. First, numerous organisms are involved in this field and scientists have uncovered minuscule information about most of them while some are yet to be identified. Second, the data that are generated in this field are of dissimilar type. Each data ranging from nucleotide and protein sequence to gene maps, expression profiles and biomaterials, is unique and vital. Third, the amount of data generated is huge due to the fast-evolving techniques of sequencing, analysis, imaging, etc. Especially, the sequencing techniques have progressed beyond

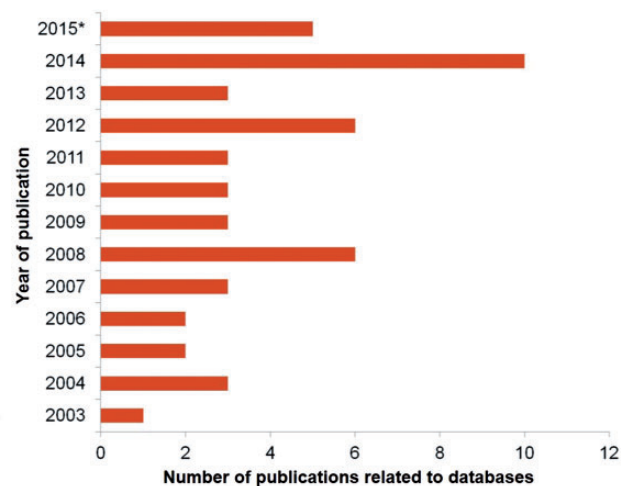


Figure 1. Number of publications on seri-related databases from the year 2003 to 2015* where (*) represents 2015-continued year.

shotgun-sequencing to more quick and efficient next generation sequencing (NGS), chromatin immuno-precipitation sequencing or ChIP-Seq, etc. which produce millions of sequence data at a go (15, 16). Till now (2003–present year), the total number of published databases in this serbioresource field is 50 out of which 27 were created in the last five years (Figure 1).

Therefore, in order to boost the research in this field, each type of data generated by scientists must be cross-checked for reliability and then archived in separate digital databases to create a helpful online space for users. These databases must be made openly accessible to others and equipped with analytical tools. This will promote better research and facilitate development of improved scientific strategies in this field.

In this review, we have collated the available databases on sericulture from 2003 till now (Figure 2) and categorized them based on the type of datasets (Supplementary Table S1). Our search attempt has led to identification of 61 databases which comprise of genome, proteome, transcriptome and other data of silkworms, host plants, pest and pathogens, etc. The databases have been briefly discussed here and schematically depicted in Figure 3. While the prime requisite of a database is to provide good quality data, it must also have an optimal web interface with integral features like search, browse, data download, etc. Quality of data can be maintained by proper data deduction methods. For example, the reliability of a transcriptome dataset can usually be depicted in the depth of sequencing. The quality can also be enhanced by regular data update and cross-referencing, simultaneously removing redundancy in the datasets. Also for a web interface, its navigation features like browser allows thorough scanning

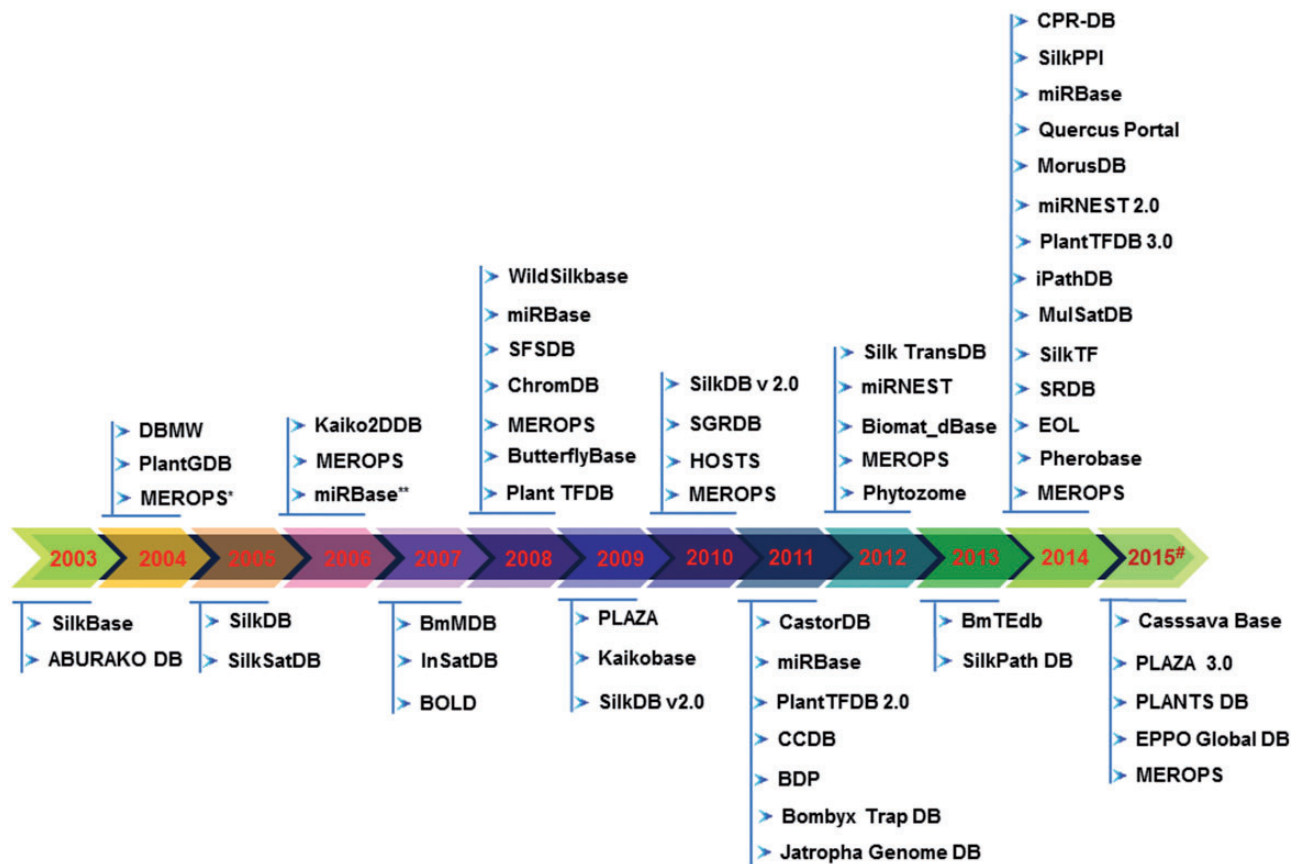


Figure 2. Timeline of the existing seridatabases from the year 2003 to 2015# generated using respective publication in the literature and database creation year from websites, where (#) represents 2015-continued year; (*) indicates database first published in 1999 and its updated versions considered from period 2003–2015; (**) indicates the same database with updated information.

of complete datasets and search engine helps a user to find the data of interest without the hassles of browsing the whole dataset. Another integral part of a database is the data download/upload option. Sometimes, huge datasets like genome or transcriptome require analyses that are not possible over the internet. In such instances, data download feature becomes really helpful. Similarly, data upload feature enables a researcher to upload their findings into a database (submitted data should always be subjected to curation by the database administrator), concurrently increasing the quantity of data. Again, user registration is not a necessary feature, but can be a useful addition to any database. Depending on the design of the registered user's interface, this feature can help a user to keep track of his or her submitted data or data of interest. Taking these features into account (Figure 4), a comparative table has been created, depicting their presence or absence (Table 1). Towards the end of the review, we have discussed potential scope and impact of these databases as well as contribution of technology to the field of sericulture and related areas. Furthermore, we have designed a user-friendly and dynamic web portal named 'SeriPort' to accommodate all the

available databases as well as web-resources related to sericulture field. The portal can be accessed at <http://www.seriport.in/>. This review will be helpful for the researchers and other enthusiasts in the field of sericulture as well as broader area of entomology.

For the ease of writing the manuscript, abbreviations of the databases were used within the text. Of all the abbreviations, some were predefined by the database creators while some were defined by authors of this manuscript.

Silkworm databases

Silkworms form the backbone of seri-ecosystem and extensive research has been done on it. Currently, there are about 20 databases available which comprise of silkworm specific information (Figure 3). According to the data type, these databases can be broadly categorized as nucleotide (13 numbers), protein (04 numbers), genetic resource (02 numbers) and pathway (01 number) databases which are briefly described and compared here.

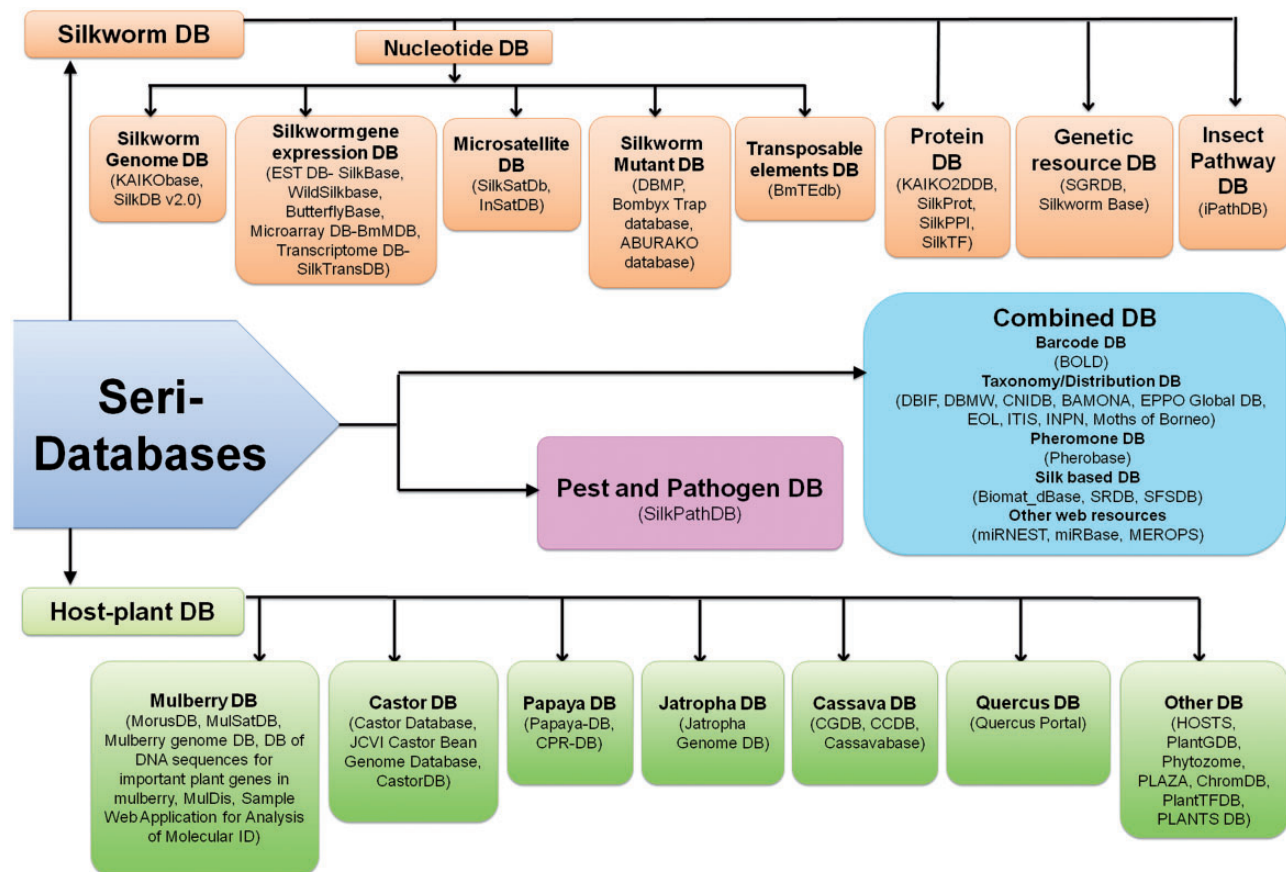


Figure 3. Schematic representation of Seri-databases classified into four categories- Silkworm Databases-20 No.s, Host Plant Databases-23 No.s, Pest and Pathogen Databases- 01 No., Combined Databases-17 No.s.

(Abbreviations- SilkDB: Silkworm Knowledgebase, EST DB: Expressed Sequence Tag Database, BmMDB: Bombyx mori Microarray Database, SilkTransDB: Silkworm Transcriptome Database, SilkSatDb: Silkworm Microsatellite Database, DBMP: Database of Bombyx mutant photographs, BmTEdb: Transposable elements database for *B. mori*, SilkProt: Annotated protein database of silkworm, SilkPPI: Silkworm Protein-Protein Interaction database, SilkTF: Silkworm Transcription Factor Database, SGRDB: Silkworm Gene Resources database, iPathDB: Insect Pathway Database, MorusDB: Morus Genome Database, MulSatDB: Mulberry Microsatellite Database, CastorDB: A comprehensive knowledgebase database for *R. communis*, Papaya-DB: Papaya Genomic Resources Online, CPR-DB: Papaya Repeat Database, CGDB: Cassava Genome Database, CCDB: Chinese Cassava Genome Database, HOSTS: a Database of the World's Lepidopteran Host plants, PlantGDB: Resources for Comparative Plant Genomics, ChromDB: The Chromatin Database, PlantTFDB: Plant Transcription Factor Database, SilkPathDB: Silkworm Pathogen Database, BOLD: Barcode of Life Data System, DBIF: Database of Insects and their Food Plants, DBMW: Database of Butterflies and Moth of the World, CNIDB: Common Names of Insects Databases, BAMONA: Butterflies and moths of North America, EOL: Encyclopedia of Life, ITIS: Integrated taxonomic information system, SRDB: Spatio-temporal database of Silk Road, SFSDb: Silk Fabric Specification Database, miRNEST: An integrative microRNA resource, miRBase: The microRNA database, MEROPS: the peptidase database).

Nucleotide databases

Nucleotide databases constitute diverse nucleotide information like genomic sequences; expressed sequence tags (ESTs), microarray, microsatellites, transcriptomic data, etc. They provide fast and easy accessibility of sequence information for biological, functional, comparative genomics and phylogenetic studies. The available nucleotide databases are briefly described as below.

Silkworm genome databases

The first draft genome of the lepidopteran model organism, *B. mori* was published in 2004 with 3× and 6× coverage by two independent research groups from Japan and

China respectively (17, 18). In the same year, as an outcome of this research, 'SilkDB' had been published from China, containing 6× draft genome sequence and integrated information on chromosomal maps, cDNAs, ESTs, transposable elements (TEs), annotation, orthologous groups in the form of genes, etc. (19). 'SilkDB' also known as 'Silkworm Knowledgebase' is the first comprehensive genomic database of *B. mori* that has been developed by *Beijing Genomics Institute (BGI), China*. The entire data in SilkDB has been organized into three modules: (i) scaffold, (ii) gene and (iii) TE linked together by the MapView program which can be accessed through keyword or BLAST search (Supplementary Table S1). The scaffold module comprises of 23 156 scaffolds for 28

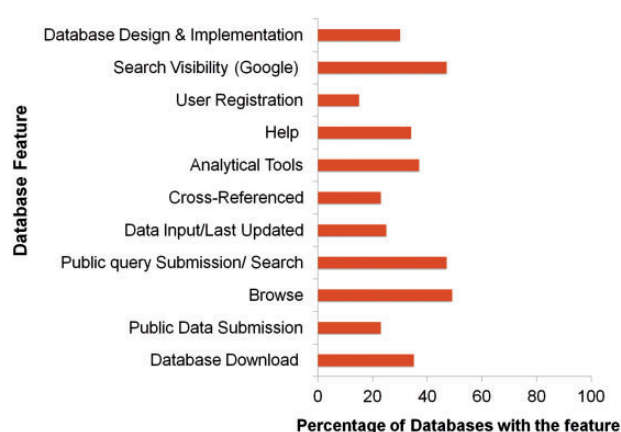


Figure 4. Percentage distribution of various features across seridatabases.

chromosomes; the gene module consists of 18 510 annotated gene sequences and full cDNA sequences of 212 known silkworm genes; and the TE module hosts around 601 225 TEs (19). This database was designed and implemented in *Oracle9i relational database management system (RDBMS)* using *JSP scripts* under *TomCat web server* and accessible at <http://silkworm.big.ac.cn/index.jsp>.

The 3× and 6× genomes carried insufficient genome sequence data due to low coverage as compared to that of other species like *Drosophila melanogaster* and *Anopheles gambiae* (20, 21). Therefore, after 3 years of publication of the draft genome, both these datasets were merged and reassembled by the ‘International Silkworm Genome Sequencing Consortium’ (ISGSC, 2007) to generate the same genome (432 Mb) with a remarkable coverage of 8.5× (22). In order to accommodate the new integrated and comprehensive genomic information, ‘KAIKObase’ and ‘SilkDB v2.0’ (upgraded version of SilkDB) were published in 2009 and 2010 respectively (23, 24). KAIKObase was developed by *National Institute of Agrobiological Sciences (NIAS), Japan* under Silkworm Genome Research Program (SGRP); while SilkDB v2.0 was developed at *Institute of Sericulture and Systems Biology, Southwest University (SWU), China*. KAIKObase harbors genome data for functional studies, BAC-end sequences, fosmid sequences, physical/genetic maps and EST sequences; while SilkDB v2.0 includes whole genome assembly, gene annotation, chromosome mapping, microarray expression, ESTs, etc. (Supplementary Table S1). KAIKObase was constructed using *PostgreSQL version 8.2.1* and implemented in *JavaScript* (23). The current version of ‘KAIKObase v3.2.2’ consists of four map browsers (PGMap, UnifiedMap, GBrowse, UTGB), one gene viewer (GeneViewer) and five independent databases (Bombyx Trap Database, KAIKO2DDB, KAIKOGAAS, full length cDNA database and EST database). The database can be accessed through

advanced three-way data mining approach-‘Chromosome Overview’, ‘Keyword and Position search’ and ‘Scaffold Sequence Search’. KAIKObase is accessible at <http://sgp.dna.affrc.go.jp/KAIKObase/>. On the other hand, SilkDB v2.0 is equipped with several user-friendly tools like Genome Browser, WEGO, ClustalW, CAP3, SilkMap, etc. in addition to BLAST (24). This version is implemented in *MySQL* (<http://www.mysql.org/>) *database management system* and navigated by *GBrowse* similar to KAIKObase. The database can be accessed at <http://www.silkdb.org>. Both KAIKObase v3.2.2 and SilkDB v2.0 are well developed and user-friendly databases with various inbuilt analysis tools. Users can upload and download map specific data from GBrowse option in KAIKObase v3.2.2. It is more advantageous for the users to use SilkDB v2.0 because it has a dedicated download page linked to an ftp server, while the download process in KAIKObase is quite complicated (Table 1). Additionally, both of these databases (KAIKObase was last updated in 2013 and SilkDBv2.0 in 2009) can be updated for their better usability. Although well-developed genome databases of *B. mori* are available as discussed above, there is a scarcity of genomic data related to other silkworms, particularly the wild silkworms. This may be attributed to the issues related to their domestication issue due to which they are still unexplored. Genome information is necessary for various downstream studies like mutation, mapping studies, etc. Development of new databases or integration of such data in the available databases will help in the analysis of their genome function and evolution.

Silkworm gene expression databases

Studies on genes using approaches like microarrays, NGS, etc. can help us in understanding gene expression and regulation under variant conditions. ESTs, microarrays and transcriptome aid in functional genomics by providing information required for genome annotation, detection of aberrant transcription, high-throughput (HT) genotyping of large populations, tissue specificity, pathogen infection-dependent gene expression, sex specificity, etc. (25–28). Gene expression studies have been applied on *B. mori* and other silkworms, leading to the development of databases which include three EST databases (‘SilkBase’, ‘WildSilkbase’ and ‘ButterflyBase’), one microarray database—‘*Bombyx mori* Microarray Database’ (BmMDB) and one transcriptome database (‘SilkTransDB’).

Among the three EST databases, SilkBase hosts EST sequences of five lepidopteran insects (*B. mori*, *B. mandarina*, *S. cynthia*, *Ernolatia moore* and *Triloca varians*), WildSilkbase hosts EST data of three economically important silkmoths (*A. assamensis*, *A. mylitta* and *S. cynthia*) of Saturniidae family and ButterflyBase integrates the EST

Table 1. Comparative features of Seri-databases (Data Search on 11 March 2016)

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
Silkworm databases													
Silkworm genome databases													
SilkDB (v2.0)	-MySQL -Navigated by GBrowse	✓ (via ftp)	✗	✓ (GBrowse, SCB tool)	✓ (Quick Search, etc.)	✗ (2009)	✓	✓ (SilkMap, BLAST, Wego, Clustal W, Cap3, BL2SEQ, EMBOS, WISE2)	✓	✗	Fair	(24)	http://www.silkdb.org
Kaikobase	-Written in Javascript -Search: PostgreSQL version 8.2.1 -Gene Viewer: HMMER version 2.1.1, ProfileScan version 2.2, PSORT version 6.4, SOSUI version 1.0, MOTIF, and InterProScan version 4.3.1 (data version 14.0) - GBrowse	✓ (Through GBrowse and in different formats- PNG, GFF, FASTA, zip, text)	✓ (can upload and share custom tracks in GBrowse)	✓ (GBrowse, UTGB and GeneViewer)	✓ (Keyword and Position Search, Scaffold Sequence Search)	✗ (07/06/2013)	✗	✓ (KAIOBLAST, KAIOGAAS, GBrowse)	✓	✗	Fair	(23)	http://sgp.dna.affrc.go.jp/KAIObase/ (version 3.2.2: 13/05/2013)
Silkworm gene expression databases													
SilkBase	NA	✓ (cDNA and EST for <i>Bombyx mori</i> and <i>Samia cynthia ricini</i> in FASTA)	✗	✓ (Libraries tab)	✓ (Different Search Bar- Keyword, gene model, genome position, ESTs, etc.)	✓ (16/03/2015)	✓	✓ (All variations of BLAST)	✗	✗	Fair	(29)	http://silkbase.ah.a.u-tokyo.ac.jp/cgi-bin/index.cgi
WildSilkbase	MySQL and PHP, Apache web server, Fedora Linux system	✓ (EST Sequence data, EST annotation data)	✗	✓	✓ (3 searches- Keyword Search, Homolog Finder and SSR Finder)	✗	✗	✓ (BLAST, Homolog Finder, SSR Finder)	✓	✗	Fair	(30)	http://www.cdf.org.in/wildsilkbase/
Butterfly Base	PostgreSQL with a customized version of the PartGene schema	✓ (FASTA)	✓	✓ (Genome Browse Under Development)	✓ (Text search, BLAST search, etc.)	✓	✗	✓ (NCBI-BLAST, PSI-BLAST and WU-BLAST-driven MS-BLAST, prot4EST)	✓	✗	Poor	(31)	http://www.butterflybase.org/ Currently link is not working (features are based on publication)
BmMDB	MySQL, Php scripts: database query, generate HTML or heatmap picture to display the query results.	✗	✗	✓ (Browse Raw data, Browse Tissue specific genes)	✓ (Search by Probe ID, Search by BLAST)	✗	✓	✓ (BLAST)	✗	✗	Fair	(33)	http://www.silkdb.org/microarray/

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/ search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
SilkTransDB	Gbrowse	✓ (PNG, SVG, FASTA, GFF)	✓ (can upload file in Gbrowse option)	✓ Gbrowse	✓ (BLAST Search)	✓ Last Updated	✓	✓ (BLAST-blastn, tblastn, tblastx)	✓	✗	Fair	(34)	http://124.17.27.136/gbrowse2/
Microsatellite database													
SilkSatDB	MySQL, PHP, Apache web server	✗	✗	✓	✓	13/10/2004	✗	✓ (SSRF, AutoPrimer)	✗	✗	Fair	(38)	http://www.cdfid.org.in/silksatdb
InSatDB	MySQL, PHP, Apache web server	✓ (.csv files)	✗	✓	✓ (multi-option query sheet)	✗	✗	✓ (Primer3)	✓ (Tutorial)	✗	Fair	(40)	www.cdfid.org.in/insatdb
Silkworm mutant databases													
DBMP	NA	✗ (Photographs of mutants can be accessed)	✗	✓	✗	✗	✗	✗ (BLAST associated with full length cDNA DB but currently not functional)	✗	✗	Fair	No publication	http://papilio.ah.a.u-tokyo.ac.jp/genome/ (Connected with Silkbase and Full length cDNA database)
ABURA-KO	NA	✗ (Photographs of mutants can be accessed)	✗	✓	✗	07/03/2012	✗	✗	✗	✗	Fair	No Publication	http://esc.nias.afrc.go.jp/natuo/en/aburako_top_en.htm
Bombyx Trap Database	Integrated with Kaikobase	✗ (Photographs of strains can be accessed)	✗	✓	✓ (Word search and Pictorial search options)	✗	✓	✗	✓	✗	Fair	No Publication for database but there is a reference publication (42)	http://sgp.dna.afrc.go.jp/ETDB/ DB is integrated with Kaikobase (23)
Transposable elements databases													
BmTEdb	NA	✓ (FASTA)	✗	✓ (Browse BmTEdb)	✓ (Keyword search)	✗	✗	✓ (BLAST, HMMER, CetoRF)	✓	✗	Fair	(43)	http://gene.cqu.edu.cn/BmTEdb/
Silkworm protein databases													
Kaiko2DDB	-Make-2DDB II software, HTML -Navigated by Graphical viewer	✗ (2D-PAGE Images can be retrieved)	✗	✓	✓ (simple search queries, combined fields)	✗	✗	✗	✗	✗	Fair	(50)	http://kaiko2ddb.dna.afrc.go.jp/ Integrated with Kaikobase, (23)
SilkProt	NA	✗	✗	✗	✓ (Submit Query)	✗	✗	✗	✗	✗	Poor	No Publication	http://www.bisomyore.in/silkprot/ (Lacks proper home page)

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/ search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
SilkPPI	NA	X	X	✓	✓	X	✓	X	✓	X	Poor	No publication for DB SilkPPI/ (Currently link is not functional) But the data available is referred in Ref (51)	http://210.212.197.30/
SilkTF	NA	X	X	✓	(Search by id)	✓	✓	✓	✓	✓	Fair	No publication	http://www.bisomyore.in/SilkTF/
Silkworm genetic resource database													
SGRDB	MYSQL, JAVA, Oracle relational database management system (RDBMS)	-	-	-	-	-	-	-	-	-	Poor	(53)	http://www.naasgo.kr/silkworm/english (Link is not functional)
Silkworm Base	NA	X	✓	✓	(3 Search options-strain, gene & references)	✓	✓	X	X	✓	Fair	No publication for DB	http://www.shigen.nig.ac.jp/silkwormbase/about_kaiko.jsp
Insect pathway databases													
iPathDB	HTML, PHP, CSS and JavaScript, Apache HTTP server, Linux operating system (Redhat 5.6, Raleigh, NC, USA)	✓ (via FTP, Data Sorted by species, Phylogenetic tree, Software: iPathCons, raw data)	✓	✓	(Search by species name, pathway ID or name, Pathway Search)	✓	✓	✓ (iPathCons)	✓	✓	Fair	(54)	http://ento.njau.edu.cn/iphath/
Silkworm host plant databases													
Mulberry databases													
MorusDB	Server: Linux Ubuntu Sever 12.04, Apache 2, MySQL Server 5.5, PHP 5.3. Common gateway interface: Perl, PHP, C, JavaScriptContent management system: Drupal CMS	✓ (via FTP-FTP Downloads, File Browser Downloads)	X	✓	GBrowse (Search, Fetch data)	✓	✓	✓ (BLAST, WEGO, Browse GO, Search released (29/08/2014)	✓	✓	Fair	(59, 60)	http://morus.swu.edu.cn/morusdb

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
MulSatDB	MySQL relational database, Apache 2.2 web server, Linux system, PHP, HTML and JavaScript	X	✓	✓	✓ (Search markers- whole genome, EST)	✓ 05/02/2014	✓	✓ (CMap, Primer3Plus, BLAST)	X	X	Fair	(61)	http://bismysore.in/mulsatdb/index.html
Castor databases													
Castor Database	PHP	X	X	✓	✓	X 04/2014	X	X	X	X	Poor	No publication	http://www.tnauagronomics.com/castor/index.php (Not accessible on 11-03-2016)
Other web resources on Castor													
JCVI Castor Bean Genome DB	NA	✓ (via FTP)	X	✓ Gbrowse (Error)	✓ (sequence BLAST search)	X	X	✓ (BLAST-blastn, blastp, blastx, tblastn, tblastx)	X	X	Fair	No Publication	http://castorbean.jcvi.org/index.shtml (Datasets are out of date)
CastorDB	MySQL, Perl API, Perl, Java, CGI, HTML, Javascript	✓ (text, jnlp file)	X	✓	✓ (simple, advanced and similarity search)	X	X	✓ (BLAST)	✓	X	Poor	(64)	http://CastorDB.msu-biotech.ac.in Currently link is not functional (11/03/2016)
Papaya databases													
Papaya-DB	NA	X	X	✓ (Gbrowse is not function-al)	✓ (Search not-functional)	X	✓	X	X	X	Fair	Draft genome published in 2008 (120)No Publication for DB	http://www.plantgenome.uga.edu/papaya/ (Not well developed. Most of the links are not working)
CPR-DB	Hosted via ftp	✓ (via FTP-TE, TR, protein sequences can be downloaded)	X	X	X	X	X	X	X	X	Fair	(66)	ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB (DB is not accessible. Only download sequence link is available)
Jatropha databases													
Jatropha genome DB	NA	✓ (via FTP)	X	✓	✓ (Keyword Search, Similarity search)	✓	X	✓ (BLASTN, BLASTP, TBLASTN)	X	X	Fair	(67)	http://www.kazusa.or.jp/jatropha/ (Available in Versions 3.0 and 4.5)

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
Cassava databases													
CGDB	NA	✓ (BAC End Sequences, Cassava ESTs, Assembled Cassava ESTs, SNPs from physical map and genes) - Right click on dataset name, & select "Save Link As"	✓	✓ (Gbrowse is not-function-al)	✓	✓	✓	(BLAST-blast, blastn, blastp, tblastn, tblastx)	✓	✓	Fair	No Publication	http://cassava.igs.umaryland.edu/cgi-bin/index.cgi
CCDB	NA	✓ (FASTA, GFF)	✓	✓ (Genome Browser)	✓	01/08/2011	✓	(BLAST)	✓ (Under construction)	✓ (Not functional)	Fair	No Publication	http://www.cassava-genome.cn/
Cassavabase	solGS: a web-based tool for genomic selection Database design: NA	✓ (via FTP)	✓	✓ GenomeBrowser ([browse])	✓ (Multi- optional search)	21/02/2016	✓	(Sequence Analysis: BLAST, VIGS Tool, Alignment Analyzer, Tree Browser Mapping: Comparative Map Viewer, CAPS Designer, solQTL: QTL Mapping Molecular Biology: In Silico PCR Systems Biology: SolCyc Biochemical Pathways, Coffee Interactomic Data, SGN Ontology Breeder Breeder Tools: Breeder Home)	✓	✓	Fair	(121)	http://www.cassava-base.org/ (Manual is provided for users to access the database)
Quercus (Oak) databases													
Quercus portal	NA	✓	✓	✓	✓ (Integrated search-Global search)	19/05/2015	✓	(BLAST)	✓ (Overall navigation guide need for easy access)	✓	Fair	-	https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=fmap (Integrated database-Features based on its all databases)

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
Other generalized plant databases (Not specific for only silkworm host plants but contains some of their information)													
HOSTS	NA	✓	✓	✓	✓	Search & Drill down search- Lepidoptera, Hostplant	✓	✓	✓	✓	Fair	(72)	http://www.nhm.ac.uk/our-science/data/hostplants/
PlantGDB	MySQL-PHP-Perl-Apache TableMakers for query and retrieval	✓ (via ftp, FASTA format, GenBank, GFF3 or EMBL format, bzip2 files and MySQL tables)	✓ (Users can submit annotations)	✓ (Genome browser)	✓ (Sequence ID, Sequence Search)	23/07/2012	✓	(BLAST, BioExtract Server, DAS, FindPrimers, GeneSeqer, GenomeThreader, MuSeqBox, PatternSearch, ProbeMatch, TE_Nest, Tracemblem, yrGATE)	✓	✓	Fair	(73)	http://www.plantgdb.org/Website is no longer updated (01/07/2015)
Phytozome	LAMPJ stack (Linux, Apache, MySQL, PHP/Perl and Java)	✓ (via the JGI Genome Portal after log in- OBO format, HTML table and tab delimited text)	✓	✓ (Gbrowse)	✓ (Keyword Search)	23/02/2016	✓	(JBrowse, BLAST, BLAT, PhytoMine, BioMart)	✓	✓	Fair	(74)	http://www.phytozome.net/ (DB is well developed and being updated as Versions 9.1, 10.0, 10.1, 10.2, 10.3, 11 etc.)
PLAZA	NA	✓ (via ftp- directory-different formats-csv, tfa),	✓	✓ (Arno) and Genome View)	✓ (Multi-optional search)	26/06/2015	✓	(BLAST and various tools to explore/identify gene families, etc.)	✓	✓	Fair	(75, 76)	http://bioinformatics.psb.ugent.be/plaza/ (DB is well developed and being updated as Versions 1, 2, 2.5, 3.0 etc.)
ChromDB	MySQL, HTML, Mason, Perl, HTML, UNIX	✓	✓	✓ (Gbrowse)	✓ (Quick Search & Advanced Search)	✓	✓	(BLAST, GMOD, GBrowse)	✓	✓	Fair	(77)	http://www.chromdb.org/index.html (Publication is available but link is not functional)
PlantTFDB	MySQL	✓ (FASTA, etc.)	✓	✓ (Browse by Species, families)	✓ (Quick Search & Advanced Search)	23/08/2013	✓	(BLAST)	✓	✓	Fair	(78-80)	http://plantfdb.cbi.pku.edu.cn/ (PTFDB v1.0, 2.0 and 3.0 available)

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
PLANTS Database	NA	✓ (uncompressed ASCII text)	✗	✓	✓ (Name Search, State Search, Advanced Search)	✓ 19/10/2015	✗	✓ (Crop Nutrient Tool, etc.)	✓	✗	Fair	(81)	http://plants.usda.gov
Pest and pathogen databases													
SilkPathDB	NA	✓ (FileServer- PNG, SVG, FASTA, GFF,	✓	✓ (Browse GO)	✓ (Keywords, sequence IDs, Locations)	✓ 08/07/2015	✓	✓ (WEGO, BLAST, EuSecPred, ProSecPred)	✓	✗	Fair	-	http://silkipathdb.swu.edu.cn/
Combined databases													
Barcode database													
BOLD	PostgreSQL, Java, C++, PHP, Linux system	✓ (Specimen data: XML or TSV, ABI sequencers) Trace files: FASTA Trace files: .ab1 or .scf. Both specimen details & sequences: XML or TSV)	✓ (trace files from ABI sequencers)	✓ (Taxonomy Browser)	✓ (Keyword search)	✓	✓	✓ (Distribution Map Analysis, Taxon ID Tree, Distance Summary, sequence composition tools etc.)	✓	✓	Fair	(86)	www.barcodinglife.org
Taxonomy/distribution related databases													
DBIF	NA	✗	✗	✓	✓ 3 options - (Search for ...:- invertebrates, host plant, source)	✗	✗	✗	✓	✗	Fair	No publication for the DB but related publications are available	http://www.brc.ac.uk/dbif/
DBMW	NA	✓ (from Aim & Scope Page)	✓	✓ (catalogue)	✓ (catalogue, bibliography, images)	✗ 06/2013	✓	✗	✓	✗	Fair	No Publication	http://www.nhm.ac.uk/research-curation/research/projects/butmoth/introduction.html
CNIDB	NA	✓ (pdf)	✓	✓	✓ (Common Names of Insects and Related Organisms)	✓ 02/03/2016	✗	✗	✗	✓	Fair	No publication	http://www.entsoc.org/pubs/common_names

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
BAMONA	NA	✓ (Images are available)	✓	✓ (Browse All Species)	✓ (Search for Species Profiles-Under Image gallery section)	✓ 08/03/2016	✗	✓ (Identification Tools)	✓	✓	Fair	No publication	http://www.butterflysandmoths.org/
EPPO Global DB	NA	✓ (Documents-pdf, Image-jpg)	✓	✓	✓ (Advanced Search)	✓ 01/2016	✗	✓ (Fast text and Batch processing tools)	✓	✓	Fair	EPPO Global Database (2015)	https://gd.eppo.int
EOL	NA	✓	✓	✓	✓ (EOL Search)	✓ 23/02/2016	✗	✗	✓	✓	Fair	(122)	http://www.eol.org
ITIS	MySQL, PostgreSQL	✓ (ITIS Downloads-Informix 7, MS SQL Server, MySQL bulk load, MySQL by table, PostgreSQL, SQLite), csv format	✓	✓	✓ (Quick Search, advanced search)	✓ 01/03/2016	✗	✓ (Compare Taxonomy/ Nomenclature, The Taxonomic Workbench tool)	✓ (Guidelines provided)	✗	Fair	No publication	http://www.itis.gov/
INPN	NA	✓	✓	✓	✓ (Search data on program, species & habitat)	✓ 09/03/2016	✗	✓ (Data & Tools option)	✗	✓	Fair	No publication	http://inpn.mnhn.fr/
Pherobase	NA	✗	✓	✓ (Browse data in various categories)	✓ (Search Bar)	✓	✓	✓ (Kovats Calculator, Formula Generator)	✓	✓	Fair	(88)	http://www.pherobase.com/
Biomat_dBase	PHP, HTML/CSS, Javascript, Red Hat Enterprise Linux 4	-	-	-	-	-	-	-	-	-	Poor	(89)	http://dbbiomat.iitkgp.ernet.in/ (But Link is not functional)
SRDB	Oracle-MySQL-PHP associated with technologies like GIS, RS, Flex and C#	-	-	-	-	-	-	-	-	-	Poor	(90)	No link available

(Continued)

Table 1. Continued

Name	Website design & implementation	Database download enabled	Public data submission	Browse	Public query submission/search	Data input	Cross-referenced	Associated with online analytical tools	Help	User registration	Search visibility (Google)	References	Comments
SFSDDB	SQL server 2000, Visual Basic.NET	-	-	-	-	-	-	-	-	-	-	(91, 92)	No link available
Other web resources													
miRNEST v2.0	HTML, CSS, PHP 5.2.11 and MySQL 4.0.31	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	(93, 94)	Integrated database available at http://lemur.amu.edu.pl/share/php/mirnest_1.0/home.php and http://mirnest.amu.edu.pl
miRBase	MySQL	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	(95-98)	http://www.mirbase.org/ miRBase 21 is currently available (released on 26/06/2014)
MEROPS	MySQL, Distributed annotation system (DAS) server	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	(99, 100)	http://merops.sanger.ac.uk Version 9.13 is currently available (released on 06/07/2015)

sequences as well as their annotation data of butterflies and other lepidopteran insects (Supplementary Table S1). SilkBase was constructed via collaboration among various institutes like *NIAS Japan, University of Tokyo*, etc. and published in 2003 (29). WildSilkbase and ButterflyBase were developed by *Centre for DNA Fingerprinting and Diagnostics (CDFD), India* and *Max Planck Institute for Chemical Ecology, Germany* and published in 2008 (30, 31). The EST data of all three databases were derived via sequencing of cDNA libraries. SilkBase comprises 35 000 ESTs obtained from 36 cDNA libraries while WildSilkbase contains 57 113 ESTs generated from 14 cDNA libraries. Both databases possess datasets extracted from different tissues and developmental stages of silkworms. ButterflyBase, being a secondary database, hosts information on 273 077 ESTs obtained from primary sequence databases like EMBL/GenBank/DBJ database for 30 diverse species and their protein translations (31). A unique feature of SilkBase is that it has cross-referenced data ensuring reliability of its data. WildSilkbase operates with a *MySQL-PHP* based interface on an *Apache web server* whereas ButterflyBase uses *PostgreSQL* with a customized version of the *PartiGene* schema (a tool used for developing partial genomes) (30–32). Database and web interface development data on SilkBase could not be discussed due to lack of information in its related publication. Other technical aspects of WildSilkbase and SilkBase like search visibility, data download, analytical tools, etc. are comparably fair. Inbuilt search options of SilkBase include all variations of BLAST, search with variable options like keywords, gene model, genome position, EST, etc. for three organisms (*B. mori*, *B. mandarina* and *S. cynthia ricini*) as well as browsing options available under ‘Libraries’ tab. Search options in WildSilkbase are slightly similar to SilkBase with options like ‘keyword search’, ‘Homology Finder’, ‘SSR Finder’ (Simple Sequence Repeat Finder) as well as inbuilt BLAST tool (blastn, tblastn and tblastx). However, there are limited options for search in ButterflyBase (simple text queries for pre-computed BLAST results). Other analytical tools in at least one of the three databases include ‘GO (Gene Ontology) Viewer’, ‘cSNP’ (SNP prediction tool), etc. Two features unique to ButterflyBase are the presence of a protein prediction tool, prot4EST and a scheme to provide services like EST and mRNA data processing prior to submission to dbEST of NCBI. This feature is absent in the other two databases. However, ButterflyBase developers should address the issues with its basic accessibility and perform search engine optimization to increase its web visibility (Table 1).

Till now, SilkBase (*Bombyx* EST Database) has the most number of citations (~225) for its related publication among the three EST databases as well as other silkworm databases

(Supplementary Table S2). Some features that account for its widespread popularity are the spectrum of search options discussed above, presence of numerous analytical tools, ease of navigation, etc. Also, SilkBase was the first database reported exclusively for a silkworm, so its early implementation has a role to play in its popularity. ‘SilkBase’, ‘WildSilkbase’ and ‘ButterflyBase’ can be accessed at <http://silkbases.ab.a.u-tokyo.ac.jp/cgi-bin/index.cgi>, <http://www.cdfd.org.in/wildsilkbases/home.php> and <http://www.butterflybase.org/> (currently not functional) respectively.

Other gene expression data available on silkworms are microarray data from ‘BmMDB’ and transcriptome data from “SilkTransDB”. “BmMDB” is the first and only microarray database of *B. mori* developed by the *Institute of Sericulture and Systems Biology (ISSB), Southwest University, China* (33). It offers tissue-specific gene expression profile obtained through genome-wide (22 987 no. of 70-mer oligonucleotides) microarray analyses in a fifth instar silkworm (Supplementary Table S1). Published in 2007, it is the second most cited database (~193) next to SilkBase among silkworm databases (Supplementary Table S2). It is integrated with SilkDB and contains a BmArray-map to display raw data. BmMDB is based on a *PHP-HTML* web interface with its back end linked to a *MySQL* database and can be searched with sequence (BLAST) and text-based (Probe ID) search options. The search output includes functional annotation of genes (name and CDS), experimental raw data, etc. Data submission and ftp download options are absent in BmMDB (Table 1).

The fifth gene expression database, ‘SilkTransDB’, was developed by *Chinese Academy of Agricultural Sciences, China* to integrate transcriptome and genome annotated data from SilkDB (34). It comprises of whole transcriptome information of different developmental stages of *B. mori* obtained through HT RNA sequencing (RNA-Seq). This data has expanded the information on silkworm genome by identifying ~5500 novel transcripts and 13 195 new exons, thus uncovering the functional complexity of *B. mori* transcriptome (Supplementary Table S1). The core data of SilkTransDB consists of 3.3 gigabase (Gb) reads covering around 7-fold of *B. mori* genome and protein-coding genes that constitute 81.3% of all the predicted genes in SilkDB (34). SilkTransDB can be browsed through *GBrowse* and searched using multiple BLAST options (blastn, tblastn and tblastx). According to the publication, it has three web-interfaces: (i) SilkDB annotation (gene, CDS and mRNA), (ii) transcriptome information (gene, structure and alternative splicing (AS) events) and (iii) Map-solexa data of reads and coverage. However, this description does not comply with the interface displayed over their respective website. SilkTransDB accepts submission of annotated data files using *GBrowse*.

The status of data update in BmMDB and SilkTransDB is unclear due to the absence of relevant information in their websites. Also, there are accessibility issues associated with 'BmMDB' website (<http://www.silkdb.org/microarray/>) which should be addressed for making it usable. 'SilkTransDB' can be addressed at <http://124.17.27.136/gbrowse2/>.

The gene expression data, viz. EST, microarray and transcriptome, from the databases discussed above, will help in the identification of unique lepidoptera-specific ESTs, unique genes and proteins as well as development of molecular markers and identification/annotation of unknown proteins. These data can also have implications on evolutionary studies on insects. Apart from the techniques used above, gene expression data can also be generated via Serial analysis of gene expression (SAGE), Fluorescence *in situ* hybridization (FISH), etc. which can be included in these databases to broaden their data spectrum, or exclusive databases can be developed for the data derived from these techniques alone (35, 36).

Microsatellite databases

Microsatellites are the repeated sequences of 1–6 bp length which are widely used in fingerprinting, linkage analysis, marker studies, etc. (37). 'SilkSatDb' was the first microsatellite database of silkworm *B. mori* created by Centre for DNA Fingerprinting and Diagnostics (CDFD), Hyderabad, India in 2005 (38). It consists of microsatellite data derived using SSRF program from whole genome shotgun (WGS) and EST sequences of *B. mori* (39). It also contains data on mutations and polymorphisms, allelic frequencies, evolutionary conservation of microsatellites, etc. (Supplementary Table S1). In addition, it catalogues extraction protocols, validated primer sequences for around 200 loci (under the tab 'PrimerBase'), informative figures and methodologies adopted for inter simple sequence repeats (ISSRs)-based genotypic analyses. A database with similar type of data but with wider range of organisms was constructed by the same laboratory in 2007 (40). Named 'InSatDb', this database comprises of microsatellite information from five completely sequenced insect genomes (*B. mori*, *D. melanogaster*, *Apis mellifera*, *Tribolium castaneum* and *Anopheles gambiae*) derived using a different tool Tandem Repeat Finder version 4 (40, 41). Various characteristics of microsatellites like nature, type, frequency, motif, genome location, repeat size, copy number, etc. of the five insects can be retrieved from this database (Supplementary Table S1).

Both are MySQL relational databases developed using PHP as a server side scripting language. The search page interface is similar in both the databases; the output interface for InSatDb is interactive, displaying an array of

information like 'Repeat Kind', 'Start', 'End', 'Copy Number', 'Length', etc. and is linked to a primer-designing tool 'Primer3' while that of SilkSatDb is linked to two analytical tools 'AutoPrimer' and 'SSR Finder'. As of 11 March 2016, the search engine was dysfunctional due to 'database connection' issues. 'InSatDb' is more advantageous to use over 'SilkSatDb' due to its wide scope of comparative genome analysis among five insects, batch download options and a 'tutorial' link; in fact, the data of SilkSatDb can be found within InSatDb with additional information. Both databases lack periodical updates and data submission features (Table 1). 'SilkSatDb' is accessible at <http://www.cdfd.org.in/SILKSAT/index.php> and 'InSatDb' at www.cdfd.org.in/insatdb. Similar to the above two sections, the microsatellite databases are available only for *B. mori* while being unavailable for other lepidopteran insects, particularly wild silk moths. Genomic information can greatly assist microsatellite studies as discussed above and thus, opens new scope for constructing more microsatellite databases.

Silkworm mutant databases

Being a model organism for insects, especially silkworms, *B. mori* mutants are generated for research as well as commercial interests and are expected to fulfill requirements like enhanced production of fibroin, different colored silks, etc. (42). Web search currently shows three databases that host information on such mutants, namely, 'Database of Bombyx mutant photographs (DBMP)', 'ABURAKO Database' and 'Bombyx Trap database'. 'DBMP' has been developed by Laboratory of Insect Genetics and Bioscience, University of Tokyo, Japan to provide photographs of the mutants having mutation on different chromosome numbers. Also, this database is integrated with full-length cDNA database and SilkDB. The other two databases, i.e. 'ABURAKO' and 'Bombyx Trap database' have been developed and maintained by NIAS, Japan. 'ABURAKO' consists of information (20 mutation loci, alleles, list of mutants, translucence level, amount of uric acid accumulation and other characteristics) on silkworm mutants with translucent larval skin due to deficit of uric acid metabolism, including low-resolution images for some of the mutants. In 'Bombyx Trap database', reporter expressions in 288 transposon insertion lines (enhancer and gene trap lines) via transposon mutagenesis are included (23, 42). Also, the information regarding the position of mutation in a genome sequence, fluorescent intensity of reporter expression at various developmental stages, reporter type, etc. are available in this database. It has text-based and image-based search options. The text-based search provides information on strain ID, reporter used, measured fluorescence site, etc. while the latter is a

browsable gallery of good-resolution images for normal and fluorescent eggs, larvae, moths, etc. Both ABURAKO and DBMP lack an in-built search engine.

A critical comparison of these three databases on silkworm mutants with the other databases discussed above makes it apparent that they require enormous improvement in various matters (Supplementary Table S1). The web interface of all the three databases is not well-designed; only one has a search engine; none of them contain data submission or download links and none of them are updated periodically (Table 1). Also, to which extent can an image help a researcher is a questionable issue. Overall, if these three databases can be unified to create a common database with the addition of missing features described above, then the resultant database will have a better application periphery. Another meaningful addition to this group can be the creation of a database of mutant generation protocols used by researchers. One can access 'DBMP', 'ABURAKO' and 'Bombyx Trap database' at <http://papilio.ab.a.u-tokyo.ac.jp/genome/>, http://cse.nias.affrc.go.jp/natuo/en/aburako_top_en.htm and <http://sgp.dna.affrc.go.jp/ETDB/>, respectively.

Transposable elements (TEs) databases

'BmTEdb' is the only database exclusively available for transposable elements of *B. mori* hosted by Chongqing University, China (43). It is a comprehensive database on 1308 TE families which have been further classified into sub-families. TEs are said to represent ~40% of the silkworm genome (44). The researchers have used a combined (*de novo*, structure-based and homology-based) approach to identify and classify the TEs within the *B. mori* genome (43). TEs play a role in the function and evolution of genes/genomes which makes the database useful for researchers trying to understand the role of these mobile elements in silkworm genetics (45, 46). BmTEdb provides users with options to search, browse and download the TE sequences in single as well as in batch. Addition of analytical tools like BLAST, HMMER and GetORF enhances the analytical scope of this database (47–49). Options like public data submission (suggestions available) and update, user account sign in, etc. are not available within the database (Table 1). BmTEdb can be accessed at <http://gene.cqu.edu.cn/BmTEdb/>.

Study of transposons including their identification, characterization and annotation, is crucial as it provides insights into genome variation and evolution. This can be greatly facilitated by genomics, genetics, transgenic technologies, HT sequencing technologies, etc. In addition to *B. mori* which has been studied well in BmTEdb, there is a great scope of developing databases for many other related silkworms.

Other web-resources

Apart from the above databases, the sequence and frequency information of the ovarian small RNAs in *B. mori* can be retrieved from a web-platform 'Silkworm sRNA' supported by National Bioresource Project (<http://www.nbrp.jp/>). Currently, it contains a total of 67 700 counts of RNA of 38 493 kinds which are available at http://papilio.ab.a.u-tokyo.ac.jp/small_RNA/all_smallRNA.txt.

Protein databases

The interest in studying silkworms is deeply rooted in the proteins (fibroin and Sericin) that it produces. Therefore, protein databases serve as an essential platform in studying gene expression, post-translational modifications and other biological processes related to silkworm proteins (50). Till now, four databases are available directly related to this area, namely, 'KAIKO2DDB', 'SilkProt', 'SilkPPI' and 'SilkTF'.

'KAIKO2DDB' (Silkworm proteome database or SPD) was the first silkworm proteome database published by NIAS, Japan in 2006 (23, 50). It houses the 2D gel-electrophoresis and mass spectrometry information of seven major tissues of silkworm (midgut, malpighian tubule, ovary, middle silk gland, posterior silk gland, fat bodies and hemolymph) (Supplementary Table S1). The data can be accessed by accession number, description ID or gene name, author, spot id/serial number, identification methods and pI/Mw range. The database was developed using Make-2DDB II software and is hosted on a web interface based on HTML. The other three databases: 'SilkProt', 'SilkPPI' and 'SilkTF' were developed by Bioinformatics Centre, CSRI&TI (Central Sericulture Research and Training Institute), Mysore, India. SilkProt database contains annotated protein data of silkworm which helps in predicting structure and pathways. SilkPPI, i.e. Silkworm Protein-Protein Interaction database provides details on protein-protein interactions of *B. mori* which facilitates the study of biological and cellular processes (51, 52). It uses protein sequences from SilkDB along with computational methods, e. g. interlog based method for data predictions (24, 51). Around 7736 protein interaction pairs including 2700 unique proteins that were predicted using Interlog method are included in the database (51). SilkDB accession number can be used to search the database for the information regarding interaction proteins, GO annotation, Pfam domains and nominal *P*-value of the microarray data. The database can be accessed through <http://210.212.197.30/SilkPPI/> but currently the link is non-functional (11 March 2016). Again, Silkworm transcription factor (SilkTF) database hosts information on transcription factors (TFs) of *B. mori* silkworm. The database can be browsed and searched either by SilkDB sequence ID or domain search. 'Sequence search'

facilitates finding of transcription factors present in the sequence, PfamID, domain name, regions and e-value information; 'Domain search' tool gives an output of list of sequence IDs having specific domain, locations of sequences with the specific domains and their corresponding e-values.

Among the four databases, KAIKO2DDB is developed slightly better than SilkProt, SilkTF and SilkPPI. The latter two have issues related to accessibility. SilkProt can have a better web interface rather than having the search engine as its homepage. Due to the lack of home page, it does not offer users any other options like data upload/download, data analysis or help page. SilkTF has similar problems to that of SilkPPI, except that it has an in-built BLAST tool (Table 1). KAIKO2DDB has numerous search options and global search options (50). It can be accessed through KAIKO Proteome Database (<http://KAIKO2DDB.dna.affrc.go.jp/>) or SWISS-2DPAGE (<http://kr.expasy.org/ch2d/make2ddb/>) under the silkworm genome database. SilkProt, SilkTF and SilkPPI are accessible at <http://www.btismysore.in/silkprot/>, <http://www.btismysore.in/SilkTF/> and <http://210.212.197.30/SilkPPI/>, respectively. Apart from these, an upcoming database 'Silkgpcr' has also been reported on the web page of *Bioinformatics Centre, CSR&TI, Mysore, India*. It will aim to provide information about the G protein coupled receptor protein and its various classifications (Rhodopsin like, Secretin like receptor, Metabotropic glutamate receptors, etc.) in *B. mori*.

Unlike nucleotide databases, there is dearth of databases related to the protein structure, sequence, protein-protein interactions, etc. The four databases discussed above have a common data type protein, however, they address three different facets. Similarly, new databases focused on silkworm protein structure or sequences can be developed by the researchers. Implementation of combinatorial approach (Proteomics and Transcriptomics) can be one of the ways to understand how variation in the proteome and transcriptome is associated with physiological changes in silk production, to characterize strains, etc. The scope of database development in this field is huge, so the issues of silk protein data scarcity should be addressed.

Silkworm genetic resource databases

Silkworm genetic resource databases which deal with data like varieties, strains, races, etc. other than genomes, transcriptomes and genes, also form an integral part of seridatabases. 'Silkworm Gene Resources database' (SGRDB) and 'SilkwormBase' are the two databases that can be classified under this group. SGRDB is a MySQL relational database developed by ERWin Data Modeler software where the data is stored in Oracle relational database management system and maintained by National Academy of

Agricultural Science (NAAS), Korea. SilkwormBase is an integrated genetic resource database of silkworms developed as a part of National BioResource Project (NBRP) between the resource centre (Graduate School of Agriculture, Kyushu University) and the information centre (National Institute for Genetics). SGRDB provides the characterization information (e.g. strain, accession number, color, shape, etc.) of 321 varieties collected from different regions including Korea, China, Japan, Europe, tropical region and non-classified group along with 1132 photo images of different life stages of these silkworm varieties. It also allows the users to access information regarding silkworm races such as univoltine, bivoltine, multivoltine and others (53). SilkwormBase hosts around 456 phenotypically classified strains and a total of 419 genes. It also facilitates the users to access information regarding genetic stock resources including strains, larval period, images of strains at different stages (egg, larva, pupa and adult), feeding habits of artificial diets, etc. and enlists the genes expressed at various life stages, their features, classification as well as linkage maps (Supplementary Table S1). SGRDB has four main functional categories, namely, variety search, characterization viewer, photo gallery and general information whereas SilkwormBase is equipped with three search options: 'Strain', 'Gene' and 'References' along with an additional option 'Distribution request' where one can online request the eggs or other developmental stages such as larva, pupa and adult of various silkworm strains. 'SGRDB' and 'SilkwormBase' are available at <http://www.naas.go.kr/silkworm/english> (not accessible on 11 March 2016) and http://www.shigen.nig.ac.jp/silkwormbase/about_kaiko.jsp, respectively.

SilkwormBase (both English and Japanese versions) is found to be more helpful in comparison to SGRDB since it is accessible and regularly updated (last updated 27 April 2015). It also links to a new website (last updated: 5 December 2014) created as a part of NBRP which deals with the collection, preservation and distribution of wild silkworms (*S. cynthia pryeri* Butler, *S. cynthia ricini* Donovan, *A. yamamai* Guérin-Méneville, *A. pernyi* Guérin-Méneville and *Rhodinia fugax* Butler). Since there are a diverse range of wild silkworms existing in the world, exploration and inclusion of those genetic resources can be a remarkable feature of this database. Alternatively, new genetic resource databases can be developed to fill up the gap in seri-related field.

Insect pathway databases

'iPathDB' is the only insect pathway database that houses pathway data on Lepidoptera (10 different orders) with a total of ~52 insects. It was developed by Li Lab: Insect

Genomic and Bioinformatics Lab, China in 2014 (54). Currently, 12 111 pathways for 52 different species associated with disease, xenobiotic metabolism signaling, insect hormone and wing development are available in this database. iPathDB has options for search (drop-down schematic and text-based). Its strongest feature is the inclusion of a pathway construction software 'iPathCons' which facilitates the users to construct pathways from transcriptome as well as official gene sets (OGSs) data of insects. Users can download the pathways constructed through iPathCons by sorting species list or species on the phylogenetic tree. Additionally, it provides batch download for the raw data files and in-built softwares (Table 1). The database was designed using *HTML*, *PHP*, *CSS* and *JavaScript* which operates under *Apache HTTP server*. These pathways will be helpful for entomological research community and are available at URL: <http://ento.njau.edu.cn/ipath/>. iPathDB has tried to cover necessary pathways and can act as a better platform for various insect pathway studies. Further, it can act as a stand-alone database if it adds more pathways in a single platform. The pathways related to insect behavior, immunity, metabolism (carbohydrate and fatty acid synthesis/metabolism), function and evolution of genes and pathways involved in sex-determination, wounding/herbivory signaling pathways, etc. can be included to make it a full-fledged database or new databases can be constructed based on essential pathway studies.

Silkworm host plant databases

Host plants are the most important resources in sericoculture as they provide food and nutrition to the silkworms. Based on the preference of feeding by the silkworms, host plants can be divided as primary (1°), secondary (2°) and tertiary (3°) host plants. The quality and yield of silk produced by the silkworms depends on the selection of these host plants. For example, the cocoon color and tensile strength of cocoon fibers varies for polyphagous silkworms. Certain host plants of silkworms have economic importance other than sericulture and have been studied with different focus. For instance, fruits of *Morus alba* (mulberry) are a great source of nutrients and anti-oxidants (55). *Jatropha* is more popular as a biofuel crop and most of the molecular and genetic research has been focused on that aspect (56). Similarly, *Ricinus communis* is known for the production of castor oil having applications as lubricant, food, medicine, etc. (57). Those host plants, which are only specific for sericulture, have been rarely reported. Overall, the host plants can be further divided into domesticated and wild silkworm host

plants. There are about 23 databases developed so far for these host plants which are discussed as below (Figure 3).

Databases of mulberry

Mulberry is the primary host plant of *B. mori* belonging to family Moraceae. About 150 species of mulberry have been identified till date, which provide shelter to several sericigenous insects in nature. *B. mori* requires specific sugars, proteins and vitamins for its normal growth and silk gland nourishment. Mulberry leaves play a very important role in providing adequate amount of nutrients for the production of good quality cocoons (58). Recent advances in HT sequencing technology have led to the generation of several mulberry specific databases that are 'Morus Genome Database' (MorusDB), 'Mulberry Microsatellite Database' (MulSatDB) and other databases.

'MorusDB' was the first mulberry genome database constructed by *Southwest University (SWU), China* and recently published in 2014. This database, available at URL <http://morus.swu.edu.cn/morusdb>, houses a wide range of genomic and biological information of *M. notabilis* C.K. Schneid (Mulberry). The core data of MorusDB constitutes 236-fold coverage of 330.79 Mb assembled mulberry genome sequence and reference-based assembled transcriptome sequence (59, 60). This information includes annotated genes, GO, ESTs, TEs, orthologs and paralogs, horizontally transferred genes, taxonomy, etc. (Supplementary Table S1). It has a user-friendly web interface designed and implemented using *MySQL* and *PHP*, embedded with helpful analytical tools like BLAST, WEGO, GO browse, genome browser, etc. One of the main advantages is availability of data download feature provided by *FTP* and *File Browser* which allows specific and batch download of the genome and transcriptome data. However, MorusDB lacks some features found in other popular databases like GenBank such as public data submission, user-registration, etc. (Table 1). Addition of the former can widen the range and amount of data in it while that of the latter will make it easier to use for the public.

'MulSatDB', on the other hand, was the first mulberry microsatellite database constructed by *CSR&TI, Mysore, India*. It comprises of mulberry genome as well as EST based microsatellite (SSR) markers (61). Currently, it hosts 217 293 WG SSRs out of which 2772 SSRs were mapped to *F. vesca* chromosomes and 361 functionally annotated SSRs among 962 present EST SSRs. The markers can be searched and browsed through two search sections: 'Whole genome' and 'EST' based on various criteria's (repeat size, repeat type, motif type, etc.). This database was based on *MySQL RDBMS* and its web interface was designed using

HTML, PHP and Javascript operating on Apache 2.2 web server. The presence of various inbuilt analysis tools (CMap, Primer3plus and BLAST), public data query and submission makes it an interactive and user-friendly database. Unlike MorusDB, this database accepts data submission on new SSR markers, marker information, research projects and publications which should be further standardized. However, it lacks data download and help/FAQ features, addition of which will further improve the database utility (Table 1). 'MulSatDB' can be accessed at <http://btismysore.in/mulsatdb/index.html>.

Apart from MorusDB and MulSatDB, few other web-resources/databases on mulberry are under constructions which are hosted at the *Bioinformatics Centre, CSRI, TI, India* (URL <http://www.btismysore.in/pgene.html>, <http://www.btismysore.in/dbase.html>). One is a relational database called 'Mulberry Genome Database' that provides data on molecular marker, DNA fingerprints, similarity and dissimilarity index matrices, phylogenetic relationship (in dendrograms) and marker segregation pattern. It is also available in the form of compact disks (CD). Other three databases include 'Database of DNA sequences for important plant genes in mulberry', 'MulDis (A Comprehensive Mulberry Disease and Pest Database)' and 'Sample Web Application for Analysis of Molecular ID'. The first database is accessible via internet while the other two are not. 'MulTF' is another database proposed in their webpage for transcription factors of mulberry. While being informative, these resources lack proper representation as well as design of a database and are tough to access, browse or even understand as no help page or publication is associated with the resources. They require further refinement in different areas of which dynamic web design is prime importance. Improvement in the existing web resources and the development of planned ones would help in the future research of mulberry.

Databases of castor

Ricinus communis (castor bean) has enormous economic and ecological importance as a popular biofuel crop (62). It also serves as the primary host plant of *S. cynthia* (Eri silkworm). Much of the genomic and molecular studies on castor are completed focusing on its economic significance. This information can be helpful in understanding the silkworm and host plant interactions. Databases have therefore been developed and published on castor, however, few of these database URLs are currently not accessible. One such database is the 'Castor Database' developed by TNAU (Tamil Nadu Agricultural University), Coimbatore, India. It hosts the phenotypic and germplasm data of the currently available castor

varieties. About 294 different germplasm including 20 FC5 plants and YRCH (Yethapur *Ricinus communis* Hybrid) plants are documented here. Users can access the information on qualitative characters (such as type of internodes, spike shape, length of primary spike, compactness of inflorescence, branching pattern of the stem, petiole length, lacination of leaf, type of inflorescence) as well as the quantitative characters (number of lobes in leaves, height of the plant, nodes in main stem, etc.) of castor plants. In addition to these traits, the yield information of various germplasm is proposed to be included in this database. This will help the farmers to select the varieties of improved traits. However, the database cannot be currently accessed through its available URL: <http://www.tnaugenomics.com/castor/index.php>. It also suffers from demerits like lack of a search engine that can do specific searches, analytical tools, data submission or download options and a help/FAQ page (Table 1). A namesake of Castor database also exists at <http://glbrc.bch.msu.edu/castor/login> hosted and maintained by Michigan State University, but it is not an open source and is accessible to registered users with no option for new registration on the homepage.

Another database on castor is 'JCVI Castor Bean Genome Database' developed at *J. Craig Venter Institute, USA*, which hosts 4X draft assembled genome sequence of *R. communis* (~400 Mbp) generated using whole genome shotgun strategy (63). The database also has ~31 221 putative proteins and ~50 000 ESTs generated from different tissues to aid in gene discovery and annotation. The data from whole genome assembly as well as auto-annotation is available to download from the ftp server. However, this database lacks a common browse or search page for the hosted data. Even GBrowse is non-functional (11 March 2016). The other analytical tool, i.e. BLAST (all categories) works fine with the data. One of its branches of data named as 'Castor Bean TAs' has been discontinued and users are referred to other related databases. This database is available at <http://castorbean.jcvi.org/index.shtml>.

'CastorDB' is a comprehensive knowledgebase DB for *R. communis* developed at *M. S. University of Baroda, India* (64). It was based on integration of the genome sequence information obtained from NCBI and the previously discussed JCVI Castor Bean Genome Database. The database facilitates the users to retrieve information on protein localization, domains, pathways, sumoylation sites, gene expression, protein-protein interactions, etc. Implemented using MySQL, Perl API (application programming interface), Java, CGI, HTML and Javascript, one of the important features of this database is the presence of three way search method- simple, advanced and

similarity search based on BLAST tool. The database is accessible at <http://castordb.msubiotech.ac.in>. The URL has not been functional since Oct' 2015 till date (11 March 2016) and hence, more information could not be provided on it.

Databases of papaya

Carica papaya (Papaya) is a highly nutritious tropical fruit plant that is popular worldwide. It is also one of the important secondary host plants of *S. cynthia*. The 3× draft WGS of *C. papaya* Linnaeus was first published in 2008 with genome of size 372 Mb (65). It has facilitated the molecular and genetic study of the plant which has industrial and agricultural significance. Also, WGS and the largely sequenced sex-determining region of papaya have provided a deep insight into its genome structure and organization. The databases describing information on papaya include *Papaya-DB* and *CPR-DB* (Papaya Repeat Database).

'Papaya-DB' is an online genomic data resource of papaya which was developed by *Center for Applied Genetic Technologies, USA* and is accessible through URL: <http://www.plantgenome.uga.edu/papaya/>. It acts as an interface to various data including WGS, EST sequences, physical and genetic maps, sex-determining region, etc. It also offers access to the Plant Genome Duplication Database (PGDD) that enables users to perform whole-genome alignments with other plant species. While the database is linked to GBrowse for WGS data browsing, the links on the web-page are non-functional (11 March 2016). Being a sole representative of papaya genome database, it lacks many important features such as non-availability of data download/deposition options, analytical tools, help page, etc. (Table 1). Developers should plan to improvise the database to make it user friendly and easily accessible by incorporating the necessary features.

The other database, 'CPR-DB', developed at the *University of Maryland, USA*, provides data on repetitive elements of papaya which constitute ~56% of its genome (66). These repetitive elements are TEs (52%), tandemly arrayed sequences (1.3%) and high copy number (HCN) genes (3%). Among transposons and tandem repeats (TRs), retrotransposons and microsatellites constitute the most abundant portion (about 43.3%), minisatellites (0.19%) and satellites representing the least portion of genome. However, the database does not have the typical web interface and is merely hosted via ftp server. The data (i.e. novel TE, TR sequences, HCN transcripts and protein sequences) can be downloaded as .fasta files but cannot be browsed or searched separately, which constitute the major demerits of CPR-DB (Table 1). The data is accessible at <ftp://ftp.cbcb.umd.edu/pub/data/CPR-DB>.

Databases of *Jatropha*

Jatropha curcas is an economically important plant which has enormous potential for biodiesel production. It is also important for sericulture, being the secondary host plant of *S. cynthia*. Much research on *Jatropha* is available but at present, only one open-source database exists for this plant. 'Jatropha Genome Database', developed at *Kazusa DNA Research Institute, Japan*, hosts genomic information and DNA markers of *Jatropha* (67). Currently, the database consists of total 297 661 187 bp sequence elements with an average G + C content of 33.7%. Also, the presence of keyword search, data download via ftp server and constant updates are the strong suites of this database (Table 1). It has undergone many revisions (current version 4.5) and is highly cited (Supplementary Table S2). Homology based search option (BLAST) is also available as an analytical tool to search full length sequence, predicted CDS sequence, predicted amino acid sequence and unigene sequences. Additional features like help/FAQ and data submission will make the database more user-friendly. This database is available at <http://www.kazusa.or.jp/jatropha/>.

Databases of cassava

Cassava is an important nutritious food, popular in the regions of Africa, Asia and South America (68). Besides *Papaya* and *Jatropha*, *Manihot esculenta* (Cassava) is also an important secondary host plant of *S. cynthia*. There are three databases available for cassava- *Cassava Genome Database* (CGDB), *Chinese Cassava Genome Database* (CCDB) and *Cassavabase*.

CGDB, developed at *Institute for Genome Sciences, USA* comprises of BAC-based fingerprint maps of an inbred cultivar of cassava and simultaneously provides data on cassava ESTs, assembled ESTs, WGS sequences and SNPs from physical maps as well as genes. It also aims to identify the traits associated with drought tolerance in cassava. On the other hand, CCDB, developed at *Fudan University, China*, provides BAC and cDNA libraries, annotated genome and transcriptome data on various pathways (gene discovery, starch metabolism, photosynthesis, drought/cold acclimatization, etc.), linkage maps, markers, etc. 'Cassavabase', another comprehensive database or rather a web portal, is different from the above two databases in matters of the information hosted by it. The database provides a combination of information ranging from genomic sequences to phenotypes, genetic maps, breeding programs, etc. This database has been developed for both researchers and breeders under *Next Generation Cassava (NEXTGEN Cassava) Breeding project* and

hosted by *Boyce Thompson Institute for Plant Research, USA*. It employs the advanced breeding machineries to improve cassava productivity and yield.

CGDB, CCDB and Cassavabase are accessible through <http://cassava.igs.umaryland.edu/cgi-bin/index.cgi>, <http://www.cassava-genome.cn> and <http://www.cassavabase.org/>, respectively. All of them have GBrowse and download options. However, link for GBrowse tool in CGDB is not functional (as on 11 March 2016). BLAST links (blastn, blastp, blastx, tblastn and tblastx) for CGDB have selected databases for checking similar sequences with several matrices to use. Addition of features like search, data submission, database update, help, user-registration, cross-referencing of the hosted data, etc. will increase the reliability of both CGDB and CCDB databases. Cassavabase is an ideal database that can be used a model by other database developers. Some of its merits are user-friendly interface, incorporation of data analysis tools for breeders (breeder home, phenotype analyze, barcode tools, genomic selection and population structure) and researchers (BLAST, ontology browser), the genomic map of Cassava with markers, etc. It also has features like data submission, query search, help topics and manuals, etc. (Table 1).

Although the three cassava databases exist as web-resources, these have not yet published. Proper publication of these databases will help in detailed understanding of the databases.

Databases of *Quercus*

Quercus or the oak is a popular timber tree and also the 1^o host plant of oak tasar silkworm (*A. proylei*). ‘Quercus Portal’ is the first web resource which deals with almost all facets of *Quercus* data (69). It is an integrative web portal developed under the EvolTree project at *Institut National de la Recherche Agronomique (INRA), France* which provides information on genome, genetic resources, biodiversity, evolution, phylogeny and taxonomy of *Quercus* (tree/shrub). Based on the information it carries, the portal is divided into eleven sub-databases. Among these, *Oak genome*, *EST* and *Candidate genes* databases are three genome/EST/gene related databases which comprise of whole genome sequence; three unigene sets for the genus *Quercus* (OCV1, OCV2 and OCV3); and putative genes related to biotic/abiotic stress, phenology and growth, respectively (70, 71). Others include marker and mapping information related databases which are *QuercusMap*, *CMap*, *SSR* and *SNP*. *QuercusMap* provides genotypic and phenotypic data on mapping pedigrees; *CMap* contains genetic and comparative maps while microsatellite and single nucleotide polymorphism data in candidate genes of oak are provided by *SSR* and *SNP* databases. Likewise, the phenotypic, genotypic,

geographic, genetic diversity and fossil data of the oak trees and their populations can be accessed through *TreePop*, *(GD)²*, *Oak provenance* and *FossilMap* databases. Besides being highly resourceful, Quercus portal also has a global search bar which allows query search across all the above-mentioned databases. However, a useful additional navigation guide with help page will make the portal easily accessible (Table 1). Quercus portal, with all the resources mentioned above, is available at URL <https://w3.pierroton.inra.fr/QuercusPortal/index.php?p=fmap>.

Other generalized plant databases

Apart from the host plant specific databases, several generalized databases exist which contain genomic, proteomic and taxonomic information of some silkworm host plants. These web-resources are non-specific and include data on the host plants which will further contribute to better understanding of their biology (Supplementary Table S1). One such database is ‘HOSTS’, a database of host plants of the lepidopteran insects (around 15%) around the world created at *Natural Museum of History, UK*. Since it consists of ~180 000 records of host plants for about 22 000 Lepidopteran species from ~1600 published and manuscript sources, HOSTS claims to be ‘the best and most comprehensive compilation of host plant data available’ (72). It has two good search modules: ‘Text Search’ or ‘Drill down search’ that allow the users to search information using two criteria’s: ‘Lepidoptera’ or ‘Host plant’. Search can be performed by family, genus, species names (only scientific names) and location to obtain the host plant data of respective insects. No record of update was found in HOSTS; regular updates, data download and addition of data submission option can broaden its knowledgebase. It is available at <http://www.nhm.ac.uk/our-science/data/hostplants/>.

Two general plant genome databases, *PlantGDB* and *Phytozome* are also available which were developed at *Indiana University* and *Department of Energy’s Joint Genome Institute, USA*, respectively. Both contain genome data of many plant species including silkworm host plants (73, 74). *PlantGDB* contains ESTs, cDNA sequences and microarray probes while *Phytozome* (current version v11) hosts 55 annotated genomes clustered into gene families. The development of *PlantGDB* was done using *MySQL-PHP-Perl-Apache server* while that of *Phytozome* by *LAMPJ stack (Linux, Apache, MySQL, PHP/Perl and Java)*. Both databases have important features like data download, browse, help, search and analysis enabled via many embedded tools. *PlantGDB* provides multiple analytical tools (BLAST, BioExtract Server, DAS, FindPrimers, GeneSequer, GenomeThreader, MuSeqBox, PatternSearch, ProbeMatch, TE_Nest, Tracemblem, yrGATE); bulk data

download facility via ftp server, and individual data files in several formats (FASTA, GenBank, GFF3 or EMBL format, bzip2 files and MySQL tables). Data download in Phytozome can be done through JGI Genome Portal only after user registration in OBO format, as HTML table and tab delimited text. Its analytical tools include JBrowse, BLAST, BLAT, PhytoMine and BioMart (Table 1). Both are highly cited comparative genomics databases (Supplementary Table S2) but PlantGDB has been discontinued in 1 July 2015 while Phytozome is regularly being updated with new versions. These two databases can be accessed at <http://www.plantgdb.org/> and <http://www.phytozome.net/>, respectively.

Few other generalized databases include 'PLAZA', 'The Chromatin Database' (ChromDB), 'Plant Transcription Factor Database' (PlantTFDB) and 'PLANTS'. PLAZA is a web-portal developed to perform comparative genomics and phylogenetic analyses (75, 76). The information of gene families and genome homology of important host plants e.g. *R. communis*, *M. esculenta*, *C. papaya* can be explored using 'Analyse' tools. The database is available through URL <http://bioinformatics.psb.ugent.be/plaza/>. ChromDB includes the sequence information of chromatin linked proteins of some silkworm host plants such as *R. communis*, *M. esculenta* and *C. papaya* (77). This database is available at <http://www.chromdb.org/index.html>. However, currently the link to access this database is not functional (11 March 2016). Another database, PlantTFDB provides the identification and classification data of TFs of few host plants of silkworms. Users can also download the list of TF families and protein sequences of TFs of the plants through the database link <http://planttfdb.cbi.pku.edu.cn/> (78–80). Besides genomic and proteomic databases, taxonomic information also plays an essential role in studying the biology of a plant. PLANTS database is the one that includes data like images, classification, ecology, etc. of a few host plants (*M. esculenta*, *Quercus* spp., *Shorea robusta*) (81). The database can be accessed through <http://plants.usda.gov/java/>. All four databases are equipped with necessary features like search, help, analytical tools and download option (download lacking in ChromDB; Table 1).

Review of the host plant databases in this section showed that the number of specialized data resources is not enough and on an average, the ones that are available are not well-equipped. Most of them lack one or the other important feature. Lack of database development expertise may be one reason behind this. Merging information sciences with biological data has been going on for quite a long time now and it is time for plant scientists to update their set of skills. Another observation was the lack of cross-references and analytical tools in the databases,

which should be made an obvious requisite for any biological database. It has also been observed that generalized host plant databases are more cited than the specific databases. For instance, Jatropha database (C-139) is cited next to Phytozome (C-637) and PlantTFDB v 2.0 (C-163) while the citation value of MorusDB, MulSatDB, etc. is small (Supplementary Table S2). Overall, host plant databases are fewer in number than silkworm or rather insect databases. There is a need to bring together the scattered data on these host plants together in one piece, as was done in the HOSTS database. Also, a need of plant-specific bioinformatics tools was seen which needs to be addressed sooner than ever, as the HT technologies are quickly generating vast arrays of data that needs to be scoured for meaningful outputs.

Pest and pathogen databases

Silkworms in association with host plants inhabit diverse niches and get affected with several viruses, bacteria, fungi and parasites ranging from mutualistic symbiosis to pathogenesis. Study on these organisms is equally essential to understand the host–pathogen interactions, studying molecular mechanisms involved in the pathogenesis, host immune response, developing new strategies against infectious pathogens, etc. (82, 83). Keeping this in mind, 'SilkPathDB' was constructed as first pathogen database by State Key Laboratory of Silkworm Genome Biology (SKLSGB) at Southwest University, China. This database deals with genomic and biological data of a variety of silkworm pathogens including fungi, bacteria, virus and microsporidia. The data includes genome sequences, gene annotation, proteomic and transcriptomic profile of silkworms under infected conditions, etc. (84). SilkPathDB is a user-friendly and full-fledged database having all necessary features like search, browse, download, help and multi-analytical tools (SilkPathDB BLAST engine, SearchGO, Browse GO, Genome Browse, EuSecPred and ProSecPred). The database is constantly upgraded (last updated 08 July 2015) and users also have freedom to upload data in this database (Table 1). These features make this database highly useful to the users interested in lepidopteran and other insect-related pathogenetic studies. One can access the database at <http://silkpathdb.swu.edu.cn/>. Despite being fully developed, the database has not yet been published. Apart from SilkPathDB, a new comprehensive silkworm disease and pest database 'SilkDis', being developed by Bioinformatics Centre at CSRTI, Mysore, India has been mentioned at URL <http://www.btismysore.in/dbase.html>. It will aim to serve as a data resource on silkworm diseases and pests; providing detailed information on

disease occurrence, infection mode, biotic/abiotic factors and effective pest/disease management approaches.

Silkworm–pathogen interaction studies suffer from a lack of understanding of silkworm pathology. It has been observed that among the seri-resources, pest and pathogen databases are the least in number. SilkPathDB can be expanded with more information on other pests and pathogens infecting various silkworms and their host plants. Moreover, new databases including pest and pathogens should be made in order to further explore the cross-talk among host, pest and pathogens.

Combined databases

In addition to the silkworm, host plant and pest/pathogen databases, we have reviewed few other databases that comprise of generalized information of seri-resources (Figure 3). These databases are not only specific to any individual silkworm, plant or pest databases but contain integrated information of these resources. Seventeen databases have been found which are classified based on the data type (Supplementary, Table S1) and briefly described as below.

Barcode databases

Barcodes are short DNA sequences serving as signatures for the identification and classification of species, process called as DNA barcoding (85). ‘BOLD’ (The Barcode of Life Data System) is one such user-friendly data resource developed by *Consortium for the Barcode of Life (CBOL)* to enable collection, storage, analysis and publication of DNA barcode sequences by amassing distributional, morphological and molecular information (86). Around 1 180 314 specimen records of lepidopteran insects including 53 476, 34 950 and 2429 members of Saturniidae, Sphingidae and Bombycidae families, respectively, are provided by BOLD (11 March 2016). Also, plants and fungi specimen records are available in this database. The data can be accessed through four main sections: (i) public data portal, (ii) a database of barcode clusters, (iii) a data collection workbench and (iv) an educational portal. Since past few years, BOLD has become a potential and central online platform for the researchers working in DNA barcoding fields. It has diverse data files which can be downloaded in different formats like Specimen data as XML or TSV, sequences as FASTA, Trace files as .ab1 or .scf and specimen details/sequences as XML or TSV formats. All essential features like data upload/download, public query, cross-referencing, user registration, help and analytical tools (Distribution Map Analysis, Taxon ID Tree, Distance Summary, sequence composition tools, etc.) make BOLD extremely perfect and user-friendly (Table 1).

It is a *PostgreSQL relational database* (www.postgresql.org) constructed using *Java*, *C++*, *PHP* and can be accessed at <http://www.boldsystems.org/>. In order to further explore huge amount of global molecular data, the ‘Barcode of Life Data Portal’ (BDP; <http://bol.uvm.edu>) was constructed by *CBOL* using *PHP* (87). It is a central resource to access the information from BOLD as well as other public databases like NCBI GenBank. Thus, it bridges the gap between the DNA barcoding scientists and the biodiversity informatics researchers. It can also assist in accessing a vast array of approaches for exploration and cataloguing of the molecular data for DNA barcoding applications.

Taxonomy/distribution related databases

Proper identification, classification, taxonomy, distribution and geographical information of species are the foremost things for biology, genetics, molecular studies, etc. Databases consisting of such information play a pivotal role in exploring, identifying and classifying species. The two such main databases are ‘Database of Insects and their Food Plants’ (DBIF) and ‘Database of Butterflies and Moth of the World’ (DBMW).

‘DBIF’ has been developed by *Biological Research Centre (BRC)*, *England* as the main part of *National Biodiversity Network (NBN)*. It is a database of invertebrates (including insects, silkworms) and their host plants with three search options- ‘Search invertebrates’, ‘Search host plants’ and ‘Search source’. Interactions for members of different families of Lepidoptera (~7 butterfly families, ~19 macro moth families and ~42 micro moth families) are provided by this database. The output is displayed in a tabular format for the selected search and can be accessed through <http://www.brc.ac.uk/dbif/homepage.aspx>.

‘DBMW’ on the other hand, was developed by *Natural History Museum (NHM)*, *UK* which catalogues around 32 000 generic names of world’s lepidopteran insects (butterflies and moths, including wild silkmths). Information of around 88, 356 and 449 members of three families: Bombycidae, Saturniidae and Sphingidae can be retrieved from the database. The data can be searched or browsed through family, genus, species, classification or images and accessible at URL: <http://www.nhm.ac.uk/our-science/data/butmoth/>.

In addition to the above data resources, few other databases describing the morphological, ecological, taxonomical data of diverse live forms are available. These include: ‘Common Names of Insects Databases’ (CNIDB; http://www.entsoc.org/pubs/common_names), ‘Moths of Borneo’ (<http://www.mothsofborneo.com/>) under Host-Parasite database (<http://www.nhm.ac.uk/research-curation/scientific-resources/taxonomy-systematics/host-para>

sites/), 'Butterflies and Moths of North America' (BAMONA, <http://www.butterfliesandmoths.org/>), 'EPPO Global Database' (<https://gd.eppo.int/>), 'Encyclopedia of Life' (EOL) Database (<http://www.eol.org/>), 'Integrated taxonomic information system' (ITIS) database (<http://www.itis.gov/>) and 'INPN' (<http://inpn.mnhn.fr/>).

Pheromone databases

Since the discovery of sex pheromones in *B. mori* (1959), huge amount of data on pheromones and other signaling compounds of insects including some silkworms has been generated. These signaling chemicals are required for communication, interaction, behavior, sexual attraction, defense, behavioral activities, etc. 'Pherobase' is the world's largest useful database of semiochemicals, i.e. pheromones and allelochemicals developed by El-Sayed AM, HortResearch, Lincoln, New Zealand in 2014 (88). Presently, the database hosts around 30 000 entries, 3500 semiochemicals and 8000 organic compounds of not only insects but also plants (floral compounds), invasive species, etc. The classification was based on various criteria's like functional groups, behavior, molecular weight, formula, etc. which can be browsed via taxa, family, genus and species. It has two inbuilt tools such as 'Kovats calculator' and 'Formula generator' to calculate kovats values and formula for specific ions respectively. It is a full-fledged database with facilities like regular updates, search, data submission, and sign in, etc. (Table 1). However, addition of download option will further enhance its applicability. This database is accessible through URL: <http://www.pherobase.com/>.

Silk-based databases

Silk has evolved as a great source of economy during past decades as discussed earlier. Few databases like 'Biomat_dBase', 'Spatio-temporal database of the Silk Road' (SRDB) and 'Silk Fabric Specification Database' (SFSDB) have been reported by different groups to cover the available information related to silk (biomaterial, Silk Road and fabric characteristics).

'Biomat_dBase' has been constructed by Indian Institute of Technology, Kharagpur, India using HTML/CSS, PHP and Javascripts. This database combines the biomaterial information with main focus on natural biomaterials including silk (89). This information includes fabrication of silk into different matrices, applications in tissue engineering, regenerative medicine, etc. Although database URL: <http://dbbiomat.iitkgp.ernet.in> is mentioned in the publication but the link is not functional currently. Development and availability of this database will

be helpful for the researchers working in the related fields in utilizing the existing resources and fabricating new biomaterials.

'SRDB' and 'SFSDB' are other similar databases mentioned in the publication but not accessible over internet. SRDB is a collaborative SQL server database developed by Chinese Academy of Sciences, Surveying and Land Information Engineering of Central South University, China, that contains the data related to Silk Road during the ancient times (90). The database mainly focuses on historical, field, geographical, remote sensing, thematic data of Han and Tang Dynasties. Accessibility of this database can act as a platform for combining both modern and archaeological technologies thus making a way towards development of a traditional archaeology. On the other hand, SFSDB was developed by National Engineering Laboratory for Modern Silk and College of Textile and Clothing Engineering, Soochow University, China (91, 92). The database constructed using SQL Server 2000 and Visual Basic.NET, deals with the fabric specification information (fabric name, number, weaving information, etc.) and its analyses (calculation of cover tightness, fabric balance coefficient, fabric shrinkage, etc.). This database if made available will help the researchers in silk fabric designing and their development.

Other web resources

During our search, we found few other databases ('miRNEST', 'miRBase' and 'MEROPS') which directly or indirectly contained information of few seri-resources.

'miRNEST' (current version: miRNEST v 2.0) is an integrated micro RNA database managed by Laboratory of Functional Genomics, Adam Mickiewicz University, Poland. Constructed using HTML, CSS, PHP 5.2.11 and MySQL 4.0.31, it includes structure and targets of miRNA candidates of the silkworms and plants (93, 94). The miRNA sequences of silkworms namely *B. mori*, *S. cynthia* and host plants such as *P. americana*, *R. communis*, *Q. robur*, *C. papaya*, *M. esculenta*, *J. curcas* are included in this database. 'miRBase' is another miRNA resource managed by the Griffiths-Jones lab at the Faculty of Life Sciences, University of Manchester, UK. It is a MySQL database that comprises of information on miRNAs, their annotation and sequences of taxa like insects (e.g. *B. mori*), host plants (e.g. *R. communis*, *M. esculenta*) (95–98). 'MEROPS' (current Release 9.13), on the other hand, is a peptidase database designed and developed at EMBL-European Bioinformatics Institute, Cambridge CB10 1SD, UK using MySQL. Peptidases (proteolytic enzymes, proteases, proteinases) are the enzymes which degrade the proteins by hydrolyzing the peptide bonds and constitute around 2% of all the proteins

Table 2. Technologies for data generation in sericulture field

Sl. No.	Sericultural research area	Technologies	References
1	Genomics	BAC-end sequencing filter, hybridization, fingerprinting, WGS, Transgenesis technology, comparative genomics, Linkage mapping, Sanger sequencing, Roche 454 Genome Sequencing, Pyrosequencing technology (454 GS-FLX), Combination of Illumina and 454, Illumina short-read sequencing, de novo seq, exome seq, targeted seq, Microarray based and genome wide association studies (GWAS)	(8, 104, 107)
2	Proteomics	SDS-PAGE, Tandem MS (Tandem Mass Spectroscopy), MALDI MS, two-dimensional gel electrophoresis (2-DE), protein microarray	(108,117, 118)
3	Transcriptomics	High-throughput RNA sequencing technology (RNA-Seq), Comparative transcriptome analysesTotal RNA and mRNA sequencing, Targeted RNA Sequencing, Small RNA and Non-coding RNA Sequencing, Serial analysis of gene expression (SAGE)	(107, 109, 110)
4	Metabolomics	MS –based system with GC (Gas chromatography) and LC (Liquid Chromatography) for initial separation, NMR analysis of crude extracts and its direct examination by MS	(113, 119)
5	Epigenomics	Methylation sequencing, Illumina high-throughput bisulfite sequencing (MethylC-Seq), ChIP Sequencing, Ribosome Profiling, Pyrosequencing technology	(107, 112)
6	Metagenomics	Metagenomics: Amplicon seq (16S rRNA), Shotgun Sequencing	(113, 115)
7	Metatranscriptomics	Metatranscriptomics: Functional study of microbial populations Illumina RNA-Seq	(113, 115)
8	Genetic Manipulation	Genetic technologies- Transposon based or genome-editing technologies	(116)

in any organism. The database offers hierarchical classification and nomenclature of the peptidases, their substrates as well as inhibitors (99, 100).

The above three databases can be accessed freely through URL <http://mirnest.amu.edu.pl>, <http://www.mirbase.org/> or as flat file from <ftp://mirbase.org/pub/mirbase/> and <http://merops.sanger.ac.uk>, respectively.

The comparative analysis of all the databases discussed in the ‘Combined Databases’ section demonstrated that most of them are better designed than databases discussed in other sections and are equipped with most of the essential features. Many of these are highly cited: miRBase being the most cited database followed by BOLD, MEROPS and so on (Supplementary Table S2). The search visibility of all the databases is good and the databases are easy to access. However, few features like data update, analytical tools, help and user registration was not uniformly observed in all combined databases. Also, some accessibility issues related to non-functional URLs of a few databases (for eg- Biomat_dBase) were observed which require troubleshooting.

Technology for data generation in sericulture field

Genomic technologies utilized in silkworm research involve structural to functional genomics. Structural genomics deals

with three-dimensional structure of gene products. Functional genomics, on the other hand, is primarily concerned with the transcriptome, i.e. gene expression analysis and with proteome i.e. protein analysis (101). It deals with the expression profiling, usage of genome by the organism under physiological or developmental conditions, etc. This review covers a time frame of ~12 years, from 2003 to 2015 (continued) (Figure 2). During this period, the technologies adapted for silkworm research have observed tremendous improvement (Table 2). The genome of *B. mori* was sequenced by BAC-end cloning and WGS sequencing (17). At present, sequencing technologies have progressed beyond old Sanger sequencing methodologies. NGS technologies are revolutionizing many areas of molecular biology such as genomics, transcriptomics, proteomics, etc. owing to their cost-effectiveness and unprecedented speed (102–106). Its main advantage is that gene discovery and expression profiling is possible through *de novo* assembly of short reads generated i.e. without any reference genome. Several NGS platforms such as Illumina/Solexa, ABI/SOLiD, 454/Roche, etc. provide broad opportunities for HT functional genomics e.g. insect chemical ecological studies such as pheromone production, reception, insect–plant interactions; genetic manipulation studies; proteomics studies, etc. (107). NGS has been applied for expression analyses of pheromone receptors, adverse effects of phoxim exposure in the *B. mori*

(108). Further, functional complexity of seri-transcriptome necessitates the exploration of diverse fields which are not yet clarified. Studies reveal that the comparative transcriptomics through RNA-Seq technology unravel the genetic basis of silk production and strength, cocoon coloration, etc. among wild and domesticated silk moths (109, 110). Apart from the above 'omics' technologies, currently metabolomics, epigenomics and metagenomics are emerging as advanced approaches for studying the metabolome, epigenome and metagenome of the insects including silkworms along with their resources (111). Changes in host metabolism, measurement of sugars, amino acids, redox agents or complex metabolite mixtures, epigenetic divergence and regulation through methylomics are few hidden areas which are yet to be applied in sericulture field (112, 113). Similarly, the metagenomic studies can reveal the microbial complexity in the gut of the insects (114, 115). Moreover, genetic studies based on transposon-mediated transgenesis and genome-editing technologies also have significant impact on genetic manipulations among silkworms (116).

Outcome of the study: SeriPort

The review of literature as well as exploration of cyberspace for all available data resources culminated into a huge amount of seri-database related information. As mentioned in the database descriptions or demerits above, some of these databases are not easily available to users due to their minimal turn-up during common searches using popular search engines. Again, some of the related web-resources are not published in literature and hence, are not known to the public. In order to address these issues, we have created an HTML based web portal which can act as a common platform for all kinds of data available on sericulture in the internet. The portal has been named 'SeriPort' and is available at <http://seriport.in/> (Supplementary Figure S1). The workflow for the construction of SeriPort is schematically represented in Supplementary Figure S2. The database for SeriPort was designed using Basic HTML 5 and CSS 3 for front-end and PhpMyAdmin and MySQL 5 for the back end (Supplementary Figure S3). The working language of website is PHP and HTML 5. The connectivity between front end and back end was done by using PHP. The central data in SeriPort, i.e. databases on silkworms, host plants, pest and pathogens, etc. were categorized in the portal in a similar manner to that of the present review. However, a separate webpage was created for each database which consisted of a short description and its web link. Apart from that, the portal also has a webpage on relevant references on these databases. The main features of SeriPort

include a user-friendly and dynamic user interface, in-built search engine and options for data download and submission. SeriPort, which is an outcome of this review, is expected to serve as a supportive portal between general users and the niche occupied by sericulture in the internet.

Conclusion

In this review, we have highlighted the databases which currently provide information on the biotic components of a silkworm's ecological niche. Silkworm thrives on plant leaves and co-inhabits its host plant with numerous other insects and micro-organisms which may act as its pest or pathogen. The efforts to understand silkworm or its interactions with other organisms have generated a plethora of information which has been converted into different types of electronic databases. The applicability as well as advantages and limitations of these databases have been previously discussed. It has been observed that problems related to data update, public data submission and incorporation, data analysis tools were common among the databases. Most of these drawbacks can be dealt with proper database architecture and programming. If we attempt to measure the usefulness of these databases, the citation count per publication can be taken into account. The database related articles have been cited ~13 553 times (on 11 March 2016) which is a significant value. Amongst the databases, the highly cited ones are SilkBase, BmMDB, SilkDB, SilkDB v2.0, Phytozome, PlantTFDB, Jatropha genome database, miRBase, BOLD and MEROPS (Supplementary Table S2).

The importance of seri-databases cannot be emphasized more. They can play crucial roles in conservation of the silkworm species, especially the wild varieties. For example, *A. assamensis* is a semi-domesticated silkworm which is endemic to North-Eastern part of India, mainly due to the climatic and environmental conditions of the place. WildSilkbase which provides the complete EST set of *A. assamensis* can assist us in understanding the functional part of its genome. This might help in engineering the organism to be able to survive in unfavorable conditions. Similarly, plant databases providing information on gene and protein sequences, diseases, etc. will facilitate the conservation of host plants, which is further important for the conservation of silkworms.

Among other scientific benefits are studies on complex interactions among silk moths, host plants and microbes as a model system to understand ecological balance; studies to understand genetics of silk materials from different silkworms; development of SNP-based molecular markers to aid in species differentiation; host plant improvement via

genetic engineering; development of effective and cheap methodologies for detection and elimination of pest and pathogen infestations in silkworms and host plants; etc. Also, proper documentation of huge information in a secure but accessible place is another reason to have more sericulture-related databases in the future. Implementation of informatics and HT technologies can aid in this regards to a great extent. From an economic standpoint, we believe that the databases can indirectly influence the economy of a region which is dependent on silkworms or their host plants for daily income. India and China are some of the sericulture-intensive countries in the world providing employment to around 9 million people, according to International Sericultural Commission. We have already witnessed the devastating consequences of Colony Collapse Disorder of honey bees in Europe. If an epidemic of pest infestation is to occur in these places, a huge number of families will be hit by it. These serio-bioresource databases will facilitate more studies on understanding the likeliness of such epidemics and development of future strategies to tackle them. Moreover, serio-bioresources will be highly benefited by the databases and it is imperative that the process of developing more databases goes on. New databases can be created for pest and pathogens, geographical locations, silkworm and host plant disease statistics, taxonomy, compounds, pathways, and so on.

SeriPort, the web-portal which is an outcome of this review can help general users in finding the sericulture-related databases over the internet in a more effective manner. Also, this portal will shed light on useful databases which are not known or seldom accessed due to invisibility in top search results.

Supplementary data

Supplementary data are available at Database Online.

Acknowledgements

DS, HC and DK express gratitude to MHRD (Government of India) for financial support in the form of fellowship. They also thank Institutional Biotech Hub (Project BT/04/NE/2009) established under Department of Biotechnology (DBT), Government of India for providing computational facility to carry out the research work.

Funding

Department of Biotechnology, Government of India, New Delhi for supporting the research through U-Excel Project (Sanction Order No. BT/411/NE/U-Excel/2013 dated 06.02.2014).

Conflict of interest. None declared.

Websites URL

EPPO. (2016). EPPO Global Database (available online). <https://gd.eppo.int>

Integrated Taxonomic Information System (ITIS) <http://www.itis.gov>

Common Names of Insects and Related Organisms http://www.entsoc.org/pubs/common_names

Butterflies and moths of North America <http://www.butterfliesandmoths.org/>

INPN <http://inpn.mnhn.fr/>

Butterflies and Moths of the World <http://www.nhm.ac.uk/our-science/data/butmoth/>

Database of Insects and their Food Plants <http://www.brc.ac.uk/dbif/homepage.aspx>

Chinese Cassava Genome Database <http://www.cassava-genome.cn/>

CassavaBase <http://www.cassavabase.org/>

Cassava Genome Database <http://cassava.igs.umaryland.edu/cgi-bin/index.cgi>

Databases Developed and Maintained at Bioinformatics Centre, CSRTI, Mysore (<http://www.btismysore.in/dbase.html>)

Castor Database <http://www.tnaugenomics.com/castor/index.php>

SilkwormBase http://www.shigen.nig.ac.jp/silkwormbase/about_kaiko.jsp

ABURAKO database-The world of silkworm larval translucent skin mutants (http://cse.nias.affrc.go.jp/natuo/en/aburako_top_en.htm)

Database of Bombyx mutant photographs, IGB Lab, Univ. Tokyo (<http://papilio.ab.a.u-tokyo.ac.jp/genome/>)

Moths of Borneo (<http://www.mothsofborneo.com/>)

References

1. Marie-Laure,R., Emerson,L. and Robertson,B.J. (2014) *The Johns Hopkins Guide to Digital Media*. Johns Hopkins University Press, Baltimore.
2. Bernstein,F.C., Koetzle,T.F., Williams,G.J. *et al.* (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112, 535–542.
3. The Flybase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, 31, 172–175.
4. Bilimoria,K.Y., Stewart,A.K., Winchester,D.P. *et al.* (2008) The National Cancer Data Base: a powerful initiative to improve cancer care in the United States. *Ann. Surg. Oncol.*, 15, 683–690.
5. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, 40, D136–D143.
6. Kumar,A., Chetia,H., Sharma,S. *et al.* (2015) Curcumin resource database. *Database (Oxford)*, 2015, bav070.
7. Debin,M.A. (1998) The great silk exchange: how the world was connected and developed. In Flynn, D., Frost, L. and Latham, A.J.H. (ed.), *Pacific Centuries: Pacific and Pacific Rim History since the 16th Century*, Routledge Press, London.

8. Goldsmith, M.R., Shimada, T., and Abe, H. (2005) The genetics and genomics of the silkworm, *Bombyx mori*. *Annu. Rev. Entomol.*, 50, 71–100.
9. Inoue, S., Kanda, T., Imamura, M. *et al.* (2005) A fibroin secretion-deficient silkworm mutant, Nd-sD, provides an efficient system for producing recombinant proteins. *Insect Biochem. Mol. Biol.*, 35, 51–59.
10. Daimon, T., Kozaki, T., Niwa, R. *et al.* (2012) Precocious metamorphosis in the juvenile hormone-deficient mutant of the silkworm, *Bombyx mori*. *PLoS Genet.*, 8, e1002486.
11. Okullo, A.A., Temu, A.K., Ogwok, P. *et al.* (2012) Physicochemical properties of biodiesel from *Jatropha* and castor oils. *Int. J. Renew. Energy Res.*, 2, 47–52.
12. Krishna, K.L., Paridhavi, M., and Patel, J.A. (2008) Review on nutritional, medicinal and pharmacological properties of Papaya (*Carica papaya* Linn.). *Nat. Prod. Rad.*, 7, 364–373.
13. Reddy, B.K. and Rao, J.V.K. (2009) Seasonal occurrence and control of silkworm diseases, grasserie, flacherie and muscardine and insect pest, uzi fly in Andhra Pradesh, India. *Int. J. Indust. Entomol.*, 18, 57–61.
14. Kuribayashi, S. (1981) Studies on the effect of pesticides on the reproduction of the silkworm *Bombyx mori* L. I. Effects of chemicals administered during the larval stage on egg-laying and hatching. *J. Toxicol. Sci.*, 6, 169–176.
15. Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24, 133–141.
16. Aparicio, O., Geisberg, J.V., Sekinger, E. *et al.* (2005) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr. Protoc. Mol. Biol.*, 69, 21.3:21.3.1–21.3.33.
17. Mita, K., Kasahara, M., Sasaki, S. *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res.*, 11, 27–35.
18. Xia, Q., Zhou, Z., Lu, C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, 306, 1937–1940.
19. Wang, J., Xia, Q., He, X. *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res.*, 33, D399–D402.
20. Adams, M.D., Celniker, S.E., Holt, R.A. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, 287, 2185–2195.
21. Holt, R.A., Subramanian, G.M., Halpern, A. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298, 129–149.
22. International Silkworm Genome Consortium (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.*, 38, 1036–1045.
23. Shimomura, M., Minami, H., Suetsugu, Y. *et al.* (2009) KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics*, 10, 486.
24. Duan, J., Li, R., Cheng, D. *et al.* (2010) SilkDB v2.0 a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.*, 38, D453–D456.
25. Jiang, M., Ryu, J., Kiraly, M. *et al.* (2001) Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. U.S.A.*, 98, 218–223.
26. Kawasaki, H., Ote, M., Okano, K. *et al.* (2004) Change in the expressed gene patterns of the wing disc during the metamorphosis of *Bombyx mori*. *Gene*, 343, 133–142.
27. Ote, M., Mita, K., Kawasaki, H. *et al.* (2004) Microarray analysis of gene expression profiles in wing discs of *Bombyx mori* during pupal ecdysis. *Insect. Biochem. Mol. Biol.*, 34, 775–784.
28. Wang, Z., Gerstein, M., and Snyder, M. (2009) RNA-Seq a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63.
29. Mita, K., Morimyo, M., Okano, K. *et al.* (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl. Acad. Sci. U.S.A.*, 100, 14121–14126.
30. Arunkumar, K.P., Tomar, A., Daimon, T. *et al.* (2008) WildSilkbase: an EST database of wild silkmths. *BMC Genomics*, 9, 338.
31. Papanicolaou, A., Gebauer-Jung, S., Blaxter, M.L. *et al.* (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res.*, 36, D582–D587.
32. Parkinson, J., Anthony, A., Wasmuth, J. *et al.* (2004) PartiGene—constructing partial genomes. *Bioinformatics*, 20, 1398–1404.
33. Xia, Q., Cheng, D., Duan, J. *et al.* (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. *Genome Biol.*, 8, R162.
34. Li, Y., Wang, G., Tian, J. *et al.* (2012) Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA sequencing. *PLoS One*, 7, e43713.
35. Funaguma, S., Hashimoto, S., Suzuki, Y. *et al.* (2007) SAGE analysis of early oogenesis in the silkworm, *Bombyx mori*. *Insect. Biochem. Mol. Biol.*, 37, 147–154.
36. Lécuyer, E. (2011) High resolution fluorescent in situ hybridization in *Drosophila*. *Methods Mol. Biol.*, 714, 31–47.
37. Toth, G., Gaspari, Z., and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, 10, 967–981.
38. Prasad, M.D., Muthulakshmi, M., Arunkumar, K.P. *et al.* (2005) SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Res.*, 33, D403–D406.
39. Sreenu, V.B., Ranjithkumar, G., Swaminathan, S. *et al.* (2003) MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. *Appl. Bioinformatics*, 2, 165–168.
40. Archak, S., Meduri, E., Kumar, P.S. *et al.* (2007) InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Res.*, 35, D36–D39.
41. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27, 573–580.
42. Uchino, K., Sezutsu, H., Imamura, M. *et al.* (2008) Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. *Insect. Biochem. Mol. Biol.*, 38, 1165–1173.
43. Xu, H.E., Zhang, H.H., Xia, T. *et al.* (2013) BmTEDb: a collective database of transposable elements in the silkworm genome. *Database (Oxford)*, 2013, bat055.
44. Osanai-Futahashi, M., Suetsugu, Y., Mita, K. *et al.* (2008) Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. *Insect. Biochem. Mol. Biol.*, 38, 1046–1057.
45. Feschotte, C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, 9, 397–405.
46. Fedoroff, N.V. (2012) Transposable elements, epigenetics, and genome evolution. *Science*, 338, 758–767.

47. Altschul,S.F., Madden,T.L., Schaffer,A.A. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.,
48. Finn,R.D., Clements,J., and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, 39, W29–W37.
49. Rice,P., Longden,I., and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, 16, 276–277.
50. Kajiwar,H., Nakane,K., Piyang,J. *et al.* (2006) Draft of silkworm proteome database. *J. Electrophoresis*, 50, 39–41.
51. Sumathy,R., Rao,A.S., Chandrakanth,N. *et al.* (2014) *In silico* identification of protein-protein interactions in Silkworm, *Bombyx mori*. *Bioinformatics*, 10, 56–62.
52. Legrain,P., Wojcik,J., and Gauthier,M. (2001) Protein-protein interaction maps: a lead towards cellular functions. *Trends Genet.*, 17, 346.
53. Kim,C., Kim,K., Park,D. *et al.* (2010) An integrated database for the enhanced identification of silkworm gene resources. *Bioinformatics*, 4, 436.
54. Zhang,Z., Yin,C., Liu,Y. *et al.* (2014) iPathCons and iPathDB: an improved insect pathway construction tool and the database. *Database (Oxford)*, 2014, bau105.
55. Legay, J.M. (1958) Recent advances in silkworm nutrition. *Annu. Rev. Entomol.*, 3, 75–86.
56. Paulillo,L.C., Mo,C., Isaacson,J. *et al.* (2012) *Jatropha curcas*: from biodiesel generation to medicinal applications. *Recent Pat. Biotechnol.*, 6, 192–199.
57. Scarpa,A. and Guerci,A. (1982) Various uses of the castor oil plant (*Ricinus communis* L.). A review. *J. Ethnopharmacol.*, 5, 117–137.
58. Seidavi,A.R., Bizhannia,A.R., Sourati,R. *et al.* (2005) The nutritional effects of different mulberry varieties on biological characters in silkworm. *Asia Pac. J. Clin. Nutr.*, 14, S122.
59. Li,T., Qi,X., Zeng,Q. *et al.* (2014) MorusDB: a resource for mulberry genomics and genome biology. *Database (Oxford)*, 2014, bau054.
60. He,N., Zhang,C., Qi,X. *et al.* (2013) Draft genome sequence of the mulberry tree *Morus notabilis*. *Nat. Commun.*, 4, 2445.
61. Krishnan,R.R., Sumathy,R., Bindroo,B.B. *et al.* (2014) MulSatDB: a first online database for mulberry microsatellites. *Trees*, 28, 1793–1799.
62. daSilva Nde,L., Maciel,M.R., Batistella,C.B. *et al.* (2006) Optimization of biodiesel production from castor oil. *Appl. Biochem. Biotechnol.*, 130, 405–414.
63. Chan,A.P., Crabtree,J., Zhao,Q. *et al.* (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nat. Biotechnol.*, 28, 951–956.
64. Thakur,S., Jha,S., and Chattoo,B.B. (2011) CastorDB: a comprehensive knowledge base for *Ricinus communis*. *BMC Res. Notes*, 4, 356.
65. Arumuganathan,K. and Earle,E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.*, 9, 208–218.
66. Nagarajan,N. and Navajas-Pérez,R. (2014) Papaya Repeat Database. In Ming,R. and Moore,P.H. (ed.), *Plant Genetics and Genomics: Crops and Models. Genetics and Genomics of Papaya*. Springer, New York, pp. 225–240.
67. Sato,S., Hirakawa,H., Isobe,S. *et al.* (2011) Sequence analysis of the genome of an oil-bearing tree, *Jatropha curcas* L. *DNA Res.*, 18, 65–76.
68. Raheem,D. and Chukwuma,C. (2001) Foods from cassava and their relevance to Nigeria and other African countries. *Agr. Hum. Values*, 18, 383–390.
69. Ehrenmann,F., Jacques-Gustave,A., Labbé,T. *et al.* (2014) Quercus portal: a web resource for genetics and genomics of oaks. In Conference IUFRO: Genetics of Fagaceae, Bordeaux.
70. Lesur,I., Le Provost,G., Bento,P. *et al.* (2015) The oak gene expression atlas: insights into Fagaceae genome evolution and the discovery of genes regulated during bud dormancy release. *BMC Genomics*, 16, 112.
71. Plomion,C., Aury,J.M., Amselem,J. *et al.* (2016) Decoding the oak genome: public release of sequence data, assembly, annotation and publication strategies. *Mol. Ecol. Res.*, 16, 254–265.
72. Robinson,G.S., Ackery,P.R., Kitching,I.J. *et al.* (2010) HOSTS—A Database of the World's Lepidopteran Hostplants (<http://www.nhm.ac.uk/hosts>). Natural History Museum, London.
73. Dong,Q., Schlueter,S.D., and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, 32, D354–D359.
74. Goodstein,D.M., Shu,S., Howson,R. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, 40, D1178–D1186.
75. Proost,S., Van Bel,M., Sterck,L. *et al.* (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell*, 21, 3718–3731.
76. Proost,S., Van Bel,M., Vaneechoutte,D. *et al.* (2015) PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.*, 43, D974–D981.
77. Gendler,K., Paulsen,T., and Napoli,C. (2008) ChromDB: the chromatin database. *Nucleic Acids Res.*, 36, D298–D302.
78. Guo,A.Y., Chen,X., Gao,G. *et al.* (2008) PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res.*, 36, D966–D969.
79. Zhang,H., Jin,J., Tang,L. *et al.* (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.*, 39, D1114–D1117.
80. Jin,J., Zhang,H., Kong,L. *et al.* (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res.*, 42, D1182–D1187.
81. USDA, NRCS. (2015) The PLANTS Database (<http://plants.usda.gov>, 22 December 2015). National Plant Data Team, Greensboro, NC.
82. Kaito,C., Kurokawa,K., Matsumoto,Y. *et al.* (2005) Silkworm pathogenic bacteria infection model for identification of novel virulence genes. *Mol. Microbiol.*, 56, 934–944.
83. Ishii,K., Hamamoto,H., and Sekimizu,K. (2015) Studies of host-pathogen interactions and immune-related drug development using the silkworm: interdisciplinary immunology, microbiology, and pharmacology studies. *Drug Discov. Ther.*, 9, 238–246.
84. Pan,G., Xu,J., Li,T. *et al.* (2013) Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. *BMC Genomics*, 14, 186.

85. Hajibabaei, M., Dewaard, J.R., Ivanova, N.V. *et al.* (2005) Critical factors for the high volume assembly of DNA barcodes. *Philos. Trans. R. Soc. London [Biol]*, 360, 1959–1967.
86. Ratnasingham, S. and Hebert, P.D. (2007) BOLD: the barcode of life data system. *Mol. Ecol. Notes*, 7, 355–364.
87. Sarkar, I.N. and Trizna, M. (2011) The Barcode of Life Data Portal: bridging the biodiversity informatics divide for DNA barcoding. *PLoS One*, 6, e14689.
88. El-Sayed, A.M. (2014) The Pherobase: Database of Pheromones and Semiochemicals. <http://www.pherobase.com> (22 January 2015, date last accessed).
89. Subia, B., Mukherjee, S., Bahadur, R.P. *et al.* (2012) Biomat_dBase: a database on biomaterials. *Open Tissue Eng. Regen. Med. J.*, 5, 9–16.
90. Bi, J., Luo, G., Wang, X. *et al.* (2014) The construction of the spatio-temporal database of the ancient Silk Road within Xinjiang province during the Han and Tang dynasties. In *IOP Conference Series: Earth and Environmental Science*, Volume 17, 012169.
91. Ling, Y., Cen, F., and Jiazhong, T. (2008) 1st International Symposium of Textile Bioengineering and Informatics, August 14 - 16, 2008 at Hong Kong Polytechnic University, pp. 1160–1164.
92. Yang, X., Liu, X.M., Sun, X.K. *et al.* (2012) Study on selvage parameter based on silk fabric specifications database. *Adv. Mater. Res.*, 502, 159–163.
93. Szcześniak, M.W., Deorowicz, S., Gapski, J. *et al.* (2012) miRNEST database: an integrative approach in microRNA search and annotation. *Nucleic Acids Res.*, 40, D198–D204.
94. Szcześniak, M.W. and Makołowska, I. (2014) miRNEST 2.0: a database of plant and animal microRNAs. *Nucleic Acids Res.*, 42, D74–D77.
95. Griffiths-Jones, S., Grocock, R.J., Van Dongen, S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, 34, D140–D144.
96. Griffiths-Jones, S., Saini, H.K., Van Dongen, S. *et al.* (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, 36, D154–D158.
97. Kozomara, A. and Griffiths-Jones, S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, 39, D152–D157.
98. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, 42, D68–D73.
99. Rawlings, N.D., Barrett, A.J., and Bateman, A. (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, 40, D343–D350.
100. Rawlings, N.D., Waller, M., Barrett, A.J. *et al.* (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.*, 42, D503–D509.
101. Tittiger, C. (2004) Functional genomics and insect chemical ecology. *J. Chem. Ecol.*, 30, 2335–2358.
102. Hawkins, R.D., Hon, G.C., and Ren, B. (2010) Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11, 476–486.
103. Zhang, J., Chiodini, R., Badr, A. *et al.* (2011) The impact of next-generation sequencing on genomics. *J. Genet. Genomics*, 38, 95–109.
104. Lee-Liu, D., Almonacid, L.I., Faunes, F. *et al.* (2012) Transcriptomics using next generation sequencing technologies. *Methods Mol. Biol.*, 917, 293–317.
105. Nguyen, Q., Nielsen, L.K., and Reid, S. (2013) Genome scale transcriptomics of baculovirus-insect interactions. *Viruses*, 5, 2721–2747.
106. Zhao, Z., Wu, G., Wang, J. *et al.* (2013) Next-generation sequencing-based transcriptome analysis of *Helicoverpa armigera* larvae immune-primed with *Photobacterium luminescens* TT01. *PLoS One*, 8, e80146.
107. Morozova, O. and Marra, M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92, 255–264.
108. Altelaar, A.F.M., Munoz, J., and Heck, A.J.R. (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat. Rev. Genet.*, 14, 35–48.
109. Fang, S.M., Hu, B.L., Zhou, Q.Z. *et al.* (2015) Comparative analysis of the silk gland transcriptomes between the domestic and wild silkworms. *BMC Genomics*, 16, 60.
110. Dong, Y., Dai, F., Ren, Y. *et al.* (2015) Comparative transcriptome analyses on silk glands of six silkmoths imply the genetic basis of silk structure and coloration. *BMC Genomics*, 16, 203.
111. Xiang, H., Zhu, J., Chen, Q. *et al.* (2010) Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. *Nat. Biotechnol.*, 28, 516–520.
112. Xiang, H., Li, X., Dai, F. *et al.* (2013) Comparative methylomics between domesticated and wild silkworms implies possible epigenetic influences on silkworm domestication. *BMC Genomics*, 14, 646.
113. Ohashi, H., Hasegawa, M., Wakimoto, K. *et al.* (2015) Next-generation technologies for multiomics approaches including interactome sequencing. *BioMed. Res. Int.*, 2015, 104209.
114. Dröge, J. and Mchardy, A.C. (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief Bioinform.*, 13, 646–655.
115. Knief, C. (2014) Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Front. Plant Sci.*, 5, 216.
116. Xu, H. and O'brochta, D.A. (2015) Advanced technologies for genetically manipulating the silkworm *Bombyx mori*, a model Lepidopteran insect. *Proc. R. Soc. B*, 282, 20150487.
117. Zhou, Z., Yang, H., and Zhong, B. (2008) From genome to proteome: great progress in the domesticated silkworm (*Bombyx mori* L.). *Acta Biochim. Biophys. Sin.*, 40, 601–611.
118. Hou, Y., Zou, Y., Wang, F. *et al.* (2010) Comparative analysis of proteome maps of silkworm hemolymph during different developmental stages. *Proteome Sci.*, 8, 45.
119. Putri, S.P., Yamamoto, S., Tsugawa, H. *et al.* (2013) Current metabolomics: technological advances. *J. Biosci. Bioeng.*, 116, 9–16.
120. Ming, R., Hou, S., Feng, Y. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 452, 991–996.
121. Fernandez-Pozo, N., Menda, N., Edwards, J.D. *et al.* (2015) The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res.*, 43, D1036–D1041.
122. Parr, C.S., Wilson, N., Leary, P. *et al.* (2014) The encyclopedia of life v2: providing global access to knowledge about life on earth. *Biodivers. Data J.*, 2, e1079.

Chapter 2

DNA Barcoding: Current Trends and Future Prospects

**Deepika Singh¹, Arghya Sett¹, Pragya Sharma²,
Manab Deka² and Utpal Bora^{1*}**

¹Bioengineering Research Laboratory,

Department of Biosciences and Bioengineering,

Indian Institute of Technology, Guwahati – 781 039, Assam

²Department of Applied Science, Institute of Science and Technology,

Gauhati University, Guwahati – 781 014, Assam

**E-mail: ubora@iitg.ernet.in, ubora@rediffmail.com*

ABSTRACT

DNA barcoding has been evolved as an effective species-identification tool as compared to classical taxonomic system in diverse areas like phylogeny, ecology, population genetics, biodiversity etc. It has immense potential in identification of unexplored species which could provide insights into the taxonomic structure, function and composition. There is a remarkable increase in the barcode sequences of some taxa whereas some taxa are still under-described. DNA barcoding can not be a standalone discipline and has to be combined with morphological data for accurate taxonomic identification. Application of high throughput sequencing platforms in this approach has now been employed to unravel the diversity with rapid and error-free results. Here we present a comprehensive view of DNA barcoding approach in various taxa such as animals, plants, fungi, bacteria, virus and protists. The methodology, pros and cons, applications, current and future prospects of DNA barcoding have been discussed as well. Furthermore, we concluded that integrative approach combining morphological, molecular and the high throughput sequencing technologies which will rapidly unveil the diversity of various life forms.

Keywords: Barcoding, Diversity, Taxonomy, Barcode, Integrative, Microbiome.

Introduction

Identification and classification of species have traditionally been carried out by the researchers and professionals of taxonomic domain. Several aspects of classical taxonomy are far from complete understanding attributed to higher chances of species misidentification (Packer *et al.*, 2009). Traditional systematic based taxonomy suffers from complications in proper identification of organisms with phenotypic plasticity, genetic variability, immaturity and crypticity. Advanced alternative methods have emerged as essential tools for correct identification and to establish the field of taxonomy (Padial and De La Riva 2007, Padial *et al.*, 2009). DNA sequence-based analyses provide plethora of information that may be effective for distinguishing various hierarchical levels *i.e.* population to species and the deepest taxonomically hierarchical level. It has enabled the ecologists and evolutionary taxonomists to address the issues associated with morphological data analyses.

A molecular and bioinformatical tool termed as DNA barcoding was proposed in 2003 by Paul Hebert to aid in diagnosis of biodiversity. However it was not a novel concept. Earlier it had been used in the discrimination of microbial species and later been tested among various taxa (Eggert *et al.*, 2002, Floyd *et al.*, 2002).

This system circumvents the lapses associated with morphology based identification systems by providing a better taxonomic resolution. Moreover, it has gained considerable validation as a suitable tool to identify species, delimit species and define species boundaries (Hebert *et al.*, 2003). With the immense progress in DNA barcoding research area, the pros and cons have been emerged. DNA barcode based Next generation sequencing (NGS) is becoming new platform to enhance the level of identification and classification owing to its cost-effective and parallel sequencing at a single run.

DNA Barcoding and its Symbology

DNA barcoding is a rapid and accurate species identification tool that uses short, standardized, conserved gene regions as internal species tags called DNA barcodes. It performs species level identification by sequencing a specific barcode region derived from unknown specimens and comparing with the reference library of sequences of known specimens. Since every species most likely have a unique DNA signature code, this approach is based on the postulate that inter-species variation exceeds intra-species variation.

An ideal DNA barcoding system should serve following criteria:

1. Variation of barcode region between species but nearly identical within species.
2. Enough phylogenetic information for easy identification and classification of unknown species.
3. Robustness in highly conserved regions and highly consistent DNA amplifications and sequencing.
4. Short target gene that can amplify the degraded DNA.

The main goal of DNA barcoding is to identify unknown, cryptic, microscopic and other species with inaccessible or complex morphology as well as discovery of new species. It has proved to be effective amongst various groups of animals such as birds, fishes, spiders, lepidopterans etc. Apart from this, it has also been applied in plants, bacteria, fungi, mollusks, macroalgae and protists taxonomic classification (Hebert *et al.*, 2003, Taberlet *et al.*, 2007).

It is a broadly accepted species identification tool for budding researchers and specialists. It also complements taxonomy, molecular phylogenetics and population genetics. The integrated information on species distribution, structure and their genetic diversity will thus enhance the speed and efficacy of local population studies (Hajibabaei *et al.*, 2007).

Presently used DNA sequences are not amenable for documentation or data storage. Despite the development of variant barcode technology in manufacturing and retailing industries, no such type of barcode has been assessed for its suitability in representing the DNA barcode sequences. In contrast to one dimensional barcodes (1D), two dimensional barcodes (2D) have better potential in DNA barcoding applications. A study reveals that the use of PDF417 digital barcodes has high longevity for digital imaging of DNA sequences (Kumar *et al.*, 2008). Moreover, various 2D barcodes (*e.g.* Data Matrix, CodaBlock-F, the Aztec Code, PDF417, PDF417 Truncated, QR2005 code, and QR code) have been evaluated as symbology for barcoding by using five broadly accepted plant and animal barcode sequences (COI, *rbcl*, *matK*, ITS2 and *psbA-trnH*). Among all, Quick Response (QR) code was reported as the most accurate symbology for DNA barcodes due to its good compression efficiency and ease of scanning and decoding (Liu *et al.*, 2012).

DNA Barcoding in Various Taxa

Selection of a suitable barcode region is an important prerequisite for DNA barcoding that varies amongst various taxa. Several mitochondrial, chloroplast and nuclear regions have been studied for their efficiency to fulfill the requirements of barcoding. Each region has its own advantages and disadvantages. The barcode markers for plants, fungi, bacteria, animals and other eukaryotes are enlisted in Table 2.1.

The schematic diagram representing DNA barcoding in animals, plants, fungi, virus, and bacteria is illustrated in Figure 2.1.

DNA Barcoding in Animals

The mitochondrial genome of animals is devoid of introns, has a limited exposure to recombination and haploid type of inheritance. Therefore, it is a better candidate for analysis than the nuclear genome (Saccone *et al.*, 1999). The universal primers enable the routine amplification of specific regions of the mitochondrial genome (Simmons and Weller 2001). The use of ribosomal sequence (12S and 16S) as universal barcode in the past phylogenetic studies is now constrained by the prevalence of insertions and deletions (indels). These indels complicate sequence alignment with the existing template sequences (Doyle and Gaut, 2000). A 650-base fragment of the 5' end of the mitochondrial gene cytochrome c oxidase I (COI, *cox1*)

Table 2.1: Choice of DNA Barcode Markers in Various Taxa

Taxa	Gene/Genomic Sequence	References
Animals and many other eukaryotes	Mitochondrial gene cytochrome c oxidase subunit 1 (<i>COI</i>) gene	Hebert <i>et al.</i> , 2003
Plants	Chloroplast gene- <i>matK</i>	Lahaye <i>et al.</i> , 2008, Yu <i>et al.</i> , 2011,
	Chloroplast gene- <i>rbcL</i>	Newmaster <i>et al.</i> , 2006
	Chloroplast gene- <i>trnH-psbA</i>	Kress <i>et al.</i> , 2005
	<i>rbcL</i> + <i>matK</i>	Hollingsworth <i>et al.</i> , 2011
	Internal transcribed spacers of nuclear ribosomal DNA (nrDNA ITS)	Chen <i>et al.</i> , 2010
Fungi	Internal transcribed spacer (<i>ITS</i>) of nuclear DNA alone or in combination with <i>SSU</i> and <i>LSU</i>	Kress <i>et al.</i> , 2005, Chen <i>et al.</i> , 2010, Dumbrell <i>et al.</i> , 2011, Arfi <i>et al.</i> , 2012, Schoch <i>et al.</i> , 2012
Bacteria	Large subunit of ribosomal polymerase II (<i>RPB1</i> and <i>RPB2</i>)	Seifert 2009
	<i>rpoB</i>	Case <i>et al.</i> , 2007, Walsh <i>et al.</i> , 2012
	16S <i>rRNA</i>	Woo <i>et al.</i> , 2008
	<i>COI</i>	Michel <i>et al.</i> (1998)
	<i>Cpn60</i>	Links <i>et al.</i> , 2012
Viruses	<i>SSU rRNA</i> in some Archaea	Benlloch <i>et al.</i> , 2002
	No specific marker (S3-peptidase gene for some virus)	Cheng <i>et al.</i> , 1995, Kumar <i>et al.</i> , 2011
Protozoans	<i>COI</i> for amoebae	Kosakyan <i>et al.</i> , 2012
	<i>ITS</i> and 28S <i>rRNA</i> (for <i>Piroplasma</i>)	Gou <i>et al.</i> , 2012
	18S <i>rRNA</i> (<i>SSU</i>)	Karnati <i>et al.</i> , 2003

Abbreviations: 16S *rRNA*: 16S ribosomal RNA, 28S *rRNA*: 28S ribosomal RNA, *COI*: cytochrome c oxidase I, *cpn60*: chaperonin-60, *rpoB*: RNA polymerase beta subunit, *matK*: maturase K, *rbcL*: ribulose-bisphosphate carboxylase, *LSU*: large subunit, *SSU*: small subunit, *ITS*: internal transcribed spacer, *RPB1*/*RPB2*: ribosomal polymerase B1/B2.

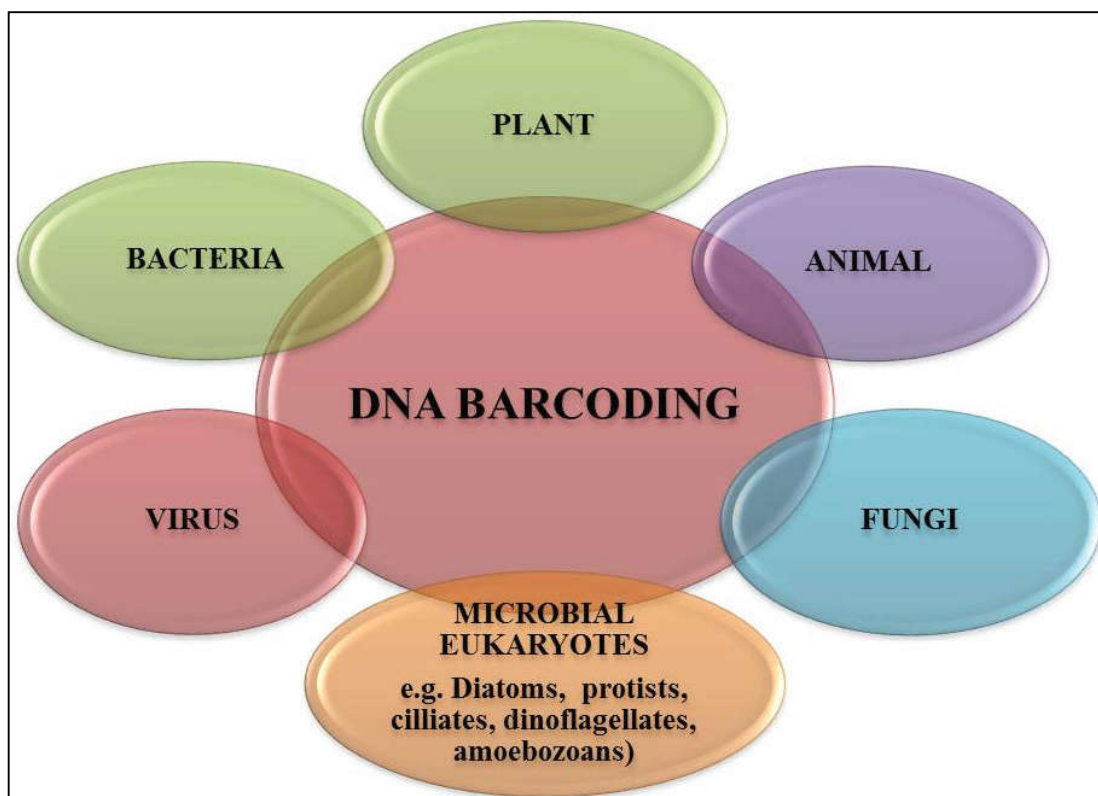


Figure 2.1: DNA Barcoding in Various Taxa.

has been considered as gold standard barcode region for animals (Hebert *et al.*, 2003, Folmer *et al.*, 1994) due to following advantages:

1. Robust universal primers for this gene.
2. Lack of insertions or deletion mutations, more conserved region of sequence.
3. A longer range of phylogenetic signal relative to ribosomal sequences (12S and 16S).

DNA barcoding resolves most of the animal species but some taxa have proved intractable. COI has been proved to be an ideal DNA barcode for most of the animal species, such as spiders, crustaceans, reptiles, fishes, birds, bryozoans, mammals *etc.* (Robinson *et al.*, 2009, Costa *et al.*, 2007, Huang *et al.*, 2015, Ivanova *et al.*, 2007, Hebert *et al.*, 2004, Kerr *et al.*, 2007, Gómez *et al.*, 2007, Bennett *et al.*, 2011). Taxonomically well-studied mammals have not been much focussed in DNA barcoding system (Wilson and Reeder, 2005). Barcoding on mammals has been epicentered on some groups such as rodents, bats and opossums (Borisenko *et al.*, 2008, Clare *et al.*, 2007).

Barcoding has also been applied for insect species, particularly Lepidopteran insects (Silva-Brandão *et al.*, 2009, Anslan and Tedersoo 2015). This is quite effective in the identification of moths, butterflies, mayflies *etc.* (Hebert *et al.*, 2003, Hebert *et al.*, 2004, Ball *et al.*, 2005). They have also been successfully employed for barcoding

of psocids of family Liposcelididae and in closely related species of silkworm pests (Jinbo *et al.*, 2011, Yang *et al.*, 2013).

DNA barcoding has emerged in field of ichthyology where fishes show morphological diversity during development (Hubert *et al.*, 2008, Steinke *et al.*, 2009). Fish is the largest group of invertebrate with significant diversity in morphology and ecology (Nelson 2006). A campaign of International Barcode of Life Project (iBOL) *i.e.* Fish Barcode of Life Initiative (FISH-BOL) provides a platform for ichthyologists to accelerate species identification (Swartz *et al.*, 2008). The high efficiency of COI in species identification has been reported in Australian marine, Canadian marine and ornamental North American fishes (Ward *et al.*, 2005). The efficiency of this marker has also been checked for Cuban freshwater fish species, billfishes and freshwater fish species from rich Balkan ichthyofauna (Lara *et al.*, 2010, Hanner *et al.*, 2011, Triantafyllidis *et al.*, 2011). A successful amplification has been achieved in the fishes of Taal Lake, the third largest lake in the Philippines, using COI as a barcode (Aquilino *et al.*, 2011). Identification of problematic fish species *e.g.* Sillaginids has also been reported where COI alone failed (Kruck *et al.*, 2013).

Furthermore, DNA barcoding has been found to be effective using COI than 16S ribosomal RNA gene (16S) for species identification of endangered amphibians (Smith *et al.*, 2008). The shortcomings of existing primers direct to inception of new primer sets. They universally amplify COI region in species like salamanders, frogs, caecilians *etc.* (Che *et al.*, 2012).

The species identification in earthworms is difficult due to rapid diversity in the evolution and lack of expertise in this field (Martin *et al.*, 2000). Use of molecular markers paves a new way for identification of new taxa, juveniles, cryptic species and earthworm biodiversity (Decaens *et al.*, 2013). Several studies revealed the efficiency of 12S rRNA (~400 bp) and 16S rRNA in species identification, due to their markedly lower genetic distances. In contrary, the efficiency of COI gene has been studied in genetic lineages and discovering cryptic species (Klarica *et al.*, 2012).

Moreover, the use of *cox1*, 16S and cytochrome b (COB) has also been explored in barcoding of primates with *cox1* showing largest amount of genetic variation (Lorenz *et al.*, 2005). Thus, COI has been proved as an ideal marker for accurate identification of primates and plays a key role in comparative primatology (Nijman *et al.*, 2010).

DNA Barcoding in Plants

The nuclear and plastid/chloroplast genomes could provide suitable barcode regions as they are variable enough to discriminate amongst species and have less intraspecific variation than interspecific variation (Chase *et al.*, 2005, Kress *et al.*, 2005). COI is not primarily used for species identification and new species discovery in plants. It is due to low substitution rate of mitochondrial genes than plastid or nuclear gene in plants (Palmer and Herbon 1988, Yang *et al.*, 1999).

Amongst all the nuclear regions, internal transcribed spacer (ITS) and regions of ITS could be used as potential barcode candidates in plant phylogenetic analysis. Although the presence of paralogous copies of ITS has been observed as a problem

in some genera (Campbell *et al.*, 2005). ITS2 region has been proved to be the most promising universal barcode candidate for identification of medicinal plants (Chen *et al.*, 2010). ITS1/ITS2 has been found to be the most potential markers for assessment of diversity and species identification of Euphorbiaceae family (Pang *et al.*, 2010).

Apart from the nuclear genes, chloroplast genes could also be good candidates in plant barcoding. It has been observed that non-coding regions are appropriate than coding regions as they exhibit more variation and reduced functional constraints (Shaw *et al.*, 2005). The plastid *trnH-psbA* spacer (119 to >1000 bp) has been proposed as a promising candidate locus for high species level identification due to following criteria:

1. Short sequence length, easy DNA extraction and amplification.
2. Highly conserved priming sites.

However, alignment and analysis of *trnH-psbA* is one of major challenge which is not a strict criterion for DNA identification but a critical prerequisite for developing bioinformatics tools (Blaxter, 2004). It shows improved discriminatory power in the identification of closely related plant taxa (Bruni *et al.*, 2012). Studies showed the successful amplification of *trnH-psbA* spacer in gymnosperms, mosses, ferns, liverwort etc. but not in all angiosperms. The length variation due to multiple indels makes the alignment difficult or impossible (Yamashiro *et al.*, 2004). In addition, *atpF-atpH*, a noncoding spacer, has been proposed as a universal DNA barcoding marker for species-level discrimination of duckweeds, an aquatic monocot belonging to the family Lemnaceae (Wang *et al.*, 2010). It has also been reported that the ITS and *trnH-psbA* could be used as a pair of potential barcode for identifying most divergent angiosperm taxa but failed in the order *Cycadales* (Sass *et al.*, 2007).

The most characterized plastid coding region in GenBank *i.e.* *rbcL* has also been evaluated for molecular identification of plants due to its universality, ease of amplification and alignment (Newmaster *et al.*, 2006). Studies reveal that it had been previously rejected as a candidate gene for species identification by Gielly and Taberlet (1994), Renner (1999), and Salazar *et al.* (2003). But it is a broad spectrum candidate for all major groups and provides a good baseline for barcoding of land plants as compared to other plastid genes. The promising results have been obtained using *rbcL* as a barcode with less variation between the species and large interspecific variation. It can be used to resolve congeneric species and different populations of mosses. Moreover, its discriminatory power has been evaluated for many arid wild plants (Bafeel *et al.*, 2012). Apart from this gene as a potential barcode, other plastid regions may also provide greater utility in barcoding of plants.

The plastid gene, *matK* has higher substitution rates than *rbcL* in angiosperms/flowering plants but not in all bryophytes and ferns (Lahaye *et al.*, 2008, Yu *et al.*, 2011). It has high resolving power in river weeds of angiosperm family Podostemaceae (Kelly *et al.*, 2010). It has also been proposed as a barcode marker for land plants. The problem in amplification of *matK* region across the plant kingdom implied the need of multiple primer set rather than a single pair (Dunning *et al.*, 2010). The potential of *matK*, *trnH-psbA* and *rbcL* regions in designing environmental

DNA (eDNA) markers has been proposed as a surveillance tool in monitoring of invasive water plants (Scriver *et al.*, 2015).

No single plant marker can substitute COI and thus combination of different markers (*trnH-psbA*, *matK*, *rpoC*, *rpoB*, *rbcL*) is now being employed in the barcoding of plants (Kress *et al.*, 2007). Consortium for the Barcode of Life-Plant Working Group (CBOL) recommended a two-locus standard barcode (*rbcL* + *matK*), with 70 per cent discriminatory power, for barcoding of plant species based on two criteria: efficient recovery of good-quality sequences and high species discriminatory potential (Vijayan *et al.*, 2010). In a study, ITS and *trnH-psbA* are used as suitable barcode loci for DNA barcoding of tropical tree species of India. *rbcL* considered as best barcode marker failed to identify tropical tree species whereas *matK* too showed low amplification rate. The combination of *rbcL* and *matK* failed to discriminate tropical tree species (Tripathi *et al.*, 2013). But recently, the combination of markers (*matK* and *rbcL*) allowed the significant discrimination among the *Annona* species (Larranaga and Hormaza, 2015). In addition to two-locus barcodes, three-locus barcodes (*rpoC1+rpoB+matK* or *rpoC1+matK+trnH-psbA*) are also being employed by various reserachers to develop plant DNA barcodes. However development of a single universal locus for discrimination of plant species is still challenging.

DNA Barcoding in Bacteria

Since 1980s, molecular characterization of bacterial species has been performed using ribosomal RNA and DNA sequences which served as the standard for species identification (Busse *et al.*, 1996). 16S rRNA was established as the “gold standard” for assessing phylogenetic analyses and microbial diversity. It has advantages like non-variable function, 1500 bp size sufficient for informatics work, presence in almost all bacterial species makes it a frequently used market. However, its utility in some bacterial species identification has been found to be problematic due to ubiquitous distribution, lack of informative characters, intragenomic heterogeneity, difficulty in sequence alignment (presence of indels) (Kjer 1995, Bennasar *et al.*, 1996).

In addition, 16S rRNA does not permit the accurate species identification among some taxa such as Enterobacteriaceae (Martinez- Murcia *et al.*, 1992, Roux and Raoult 1995). To confront the above limitations, use of alternate markers (conserved protein coding genes) including *rpoB* (the beta subunit of DNA polymerase) *gyrB*, the structural gene for the DNA gyrase β subunit has been proposed for accurate species identification (Ade 'kambi *et al.*, 2009, Holmes *et al.*, 2004). Amongst all markers, *rpoB* offers several advantages as described below:

1. *rpoB* is typically present in a single copy within a genome.
2. High level of genetic divergence and more informative characters provides an enhanced basis for phylogenetic reconstruction. Higher resolution of *rpoB* marker can reveal molecular evolution down to the population level.
3. Easy interpretation of the data generated using this marker in an evolutionary framework.
4. Easy identification and removal of the sequence errors which can disrupt the reading frame.

There are an incessant number of studies using single-copy protein-encoding genes for assessment of subspecies diversity. *rpoB* has widely been exploited as a potential barcode marker for bacterial species identification (Case *et al.*, 2007, Vos *et al.*, 2012). To date, many studies reveal the efficiency of using *rpoB* gene sequences for the identification and taxonomic classification of bacterial species including *Bacillus*, *Mycobacterium*, *Legionella*, *Staphylococcus*, *Pseudomonas* and *Streptomyces* (Ki *et al.*, 2009, Richert *et al.*, 2007). It has been demonstrated that the polymorphism exhibited by this gene can be used to identify marine *Vibrio* isolates in marine waters. The phylogenetic relationship between the gram-positive *Actinobacteria* and *Firmicutes* has also been studied using *rpoB* as a marker gene. It has been found to be an alternative marker for phylogenetic analyses and species identification in Archaea and bacteria. A remarkable level of divergence has been seen among enteric bacteria as compared to 16S *rRNA* gene (Mollet *et al.*, 1998).

For past few decades, novel, as yet to be cultured haloarchaeal isolated were isolated and identified using *SSU rRNA* (Benlloch *et al.*, 2002). *SSU rRNA* is a multi-copy gene present in a single genome. There is less divergence (<1 per cent) between these copies in most of the archaea and bacteria but not in many haloarchaea (Acinas *et al.*, 2004). Intragenomic heterogeneity is one of the problems associated with the study of taxonomy and molecular evolution in haloarchaea especially the genera *Haloterrigena* and *Natrinema* (Boucher *et al.*, 2004). Thus, *rpoB* has thus been proposed as an ideal phylogenetic marker to study the taxonomic diversity in haloarchaea due to its universal and conserved nature (Walsh *et al.*, 2012). The utility of COI has also been observed to discriminate *Wolbachia* species which are the common endosymbiotic bacteria (Smith *et al.*, 2012). Apart from the above regions, chaperonin-60 (*cpn60*) has been found to be a bacterial barcode of 555 bp length (Links *et al.*, 2012). Hence *COI*, 16S *rRNA*, *cpn60* and *rpoB* genes can be suitable for bacterial DNA barcode development process. These barcodes can be effective in the identification, characterization and exploration of microbiota of the organisms (probiotic microbes) having economic significance (Woo *et al.*, 2008, Michel *et al.*, 1998, Reva *et al.*, 2015).

DNA Barcoding in Fungi

The use of *COI* as a standard barcode has also been excluded in fungi due to its difficulty in amplification and insufficient variation (Schoch *et al.*, 2012). *ITS*, a multi-copy and easily amplified region of the ribosomal DNA is considered as the most ideal barcode for identification of most of the fungal taxa at species or lower level (Hillis 1991, Alvarez 2003). It comprises of *ITS1* and *ITS2* regions, distinguished by the 5.8S gene and is covered by the 18S (*SSU*) and 28S (*LSU*) genes in the nuclear *rDNA* repeat units (Figure 2.2).



Figure 2.2: Map of Nuclear *rRNA* Genes and their *ITS* Regions in Fungi.

A major success has been achieved in the Basidiomycota division. However, multiple sequence alignments and subsequent phylogenetic analyses using this region are often hindered by its considerable variable length. Furthermore, ITS lacks species resolution, especially in the economically important microfungal groups in Ascomycota division, which constitute 60 per cent of the presently recognized fungal taxa (Seifert, 2009). So, there is an immense requirement of some other region which will allow the robust species identification in Ascomycota division.

To address all these issues, Assembling the Fungal Barcode of Life (AFTOL) project (2008) proposed the use of several protein coding genes *e.g.* ribosomal polymerase B2 (*rpb2*), β -tubulin for *Penicillium* and translation elongation factor 1- α for *Fusarium* for species identification and phylogenetic analyses in fungi particularly in Ascomycota division (Frisvad *et al.*, 2004, Schoch *et al.*, 2009, O'Donnell *et al.*, 2010). They provide better species level resolution in most of the fungi (Robert *et al.*, 2011). In Nectriaceous fungi, which are economically important and highly diverse fungi, beta-tubulin gene has been considered as a potential barcode marker (Zhao *et al.*, 2011). Among the protein coding genes, the largest subunit of RNA polymerase II (*RPB1*) has been proposed as a candidate barcode in Basidiomycota, Microsporidia and zygomycota and some protists due to the presence of single copy and comparatively low level of sequence variance (Garnica *et al.*, 2009). However, low PCR amplification rate than ITS might be a critical issue in its usage as a barcode candidate (McLaughlin *et al.*, 2009).

Presently, more than 130000 fungal ITS sequences have been submitted to the BOLD database and this number is growing exponentially (Table 2.2) (Hibbett *et al.*, 2011). ITS region is being employed as an ideal marker for identification of fungi in environmental samples (environmental DNA barcoding) (Nilsson *et al.*, 2009). The presence of multiple copies of ITS per cell (~250) makes it an ideal target for sequencing environmental templates having low quantity of DNA (Vilgalys *et al.*, 1990). In some environmental studies, the combination of ITS and LSU has been explored as candidate for barcoding (Geml *et al.*, 2009).

Table 2.2: List of Public Databases of DNA Barcoding

Databases	Data Content	Website
BOLD (Barcode of Life Data Systems)	Specimen Records: 5,800,485, Specimens with Barcodes: 4,353,038, Species with Barcodes: 245,957 Animals 5201271, Plants 392238, Fungi 137666, Protists 23091	http://www.boldsystems.org
MMDBD (Medicinal Materials DNA Barcode Database)	1661 Medicinal species with 51375 sequences available.	http://www.cuhk.edu.hk/icm/mmdbd.htm
ITS2 Database Ribosomal RNA Database	ITS2 sequences and 25000 secondary structures for the known ITS sequences	http://its2.bioapps.biozentrum.uni-wuerzburg.de
The ITS2 Database III- sequences and structures for phylogeny	Identification of more than 160,000 correct full length and more than 50,000 partial structures	http://its2.bioapps.biozentrum.uni-wuerzburg.de

Besides ITS, the ribosomal subunits, small (SSU) and large (LSU) have also been reported for identification of novel fungal clade (Arfi *et al.*, 2012). The small subunit of rRNA (18S) is a conserved region with fewer hypervariable domains and is widely used in phylogenetic analyses in several Basidiomycota genera (Uhart *et al.*, 2007). The V6 and V9 domains of SSU displayed a high divergence level between *Tricholoma* species (Mouhamadou *et al.*, 2008). ITS1 and ITS2 gene regions have failed to infer for species identification in aquatic hyphomycetes where ITS1-5.8S-ITS2 rRNA gene has proved as a potential barcode (Seena *et al.*, 2010). Moreover, the combination of ITS along with SSU and LSU has been found to be a basis of DNA barcoding in arbuscular mycorrhizal fungi (Stockinger *et al.*, 2010). Thus for simultaneous species identification and phylogenetic studies, ITS supplemented with additional sequences such LSU or SSU should be used as barcoding markers (Toju *et al.*, 2012, Schoch *et al.*, 2012). DNA barcoding has thus proved to be useful in exploring the fungal diversity, studying pathogenicity, geographical and ecological roles.

DNA Barcoding in Viruses

The complexity of morphology-based system has led to the need of molecular methods for species identification in viruses (Ward 1993). DNA barcoding has been suggested as a new tool for species identification in viruses. Furthermore, it has been exploited as a tool for Avian influenza virus (AIV) isolation and identification of host bird species, thereby, helping to understand the AIV ecology, evolution and epidemiology in wild birds (Lee *et al.*, 2010). A 408 bps region encoding S3-peptidase located in structural gene of viral genome has been proposed as a hypothetical barcode which is yet to be confirmed (Cheng *et al.*, 1995, Kumar *et al.*, 2011). The viral DNA barcode records available on BOLD and the number of publications are negligible. No DNA or RNA marker has been developed to elucidate the complexity of the virus and studying viral diversity. Work has to be done in exploring a universal barcode region for accurate identification of viral taxa.

DNA Barcoding in Eukaryotic Microbes

DNA barcoding would be highly advantageous for the identification of eukaryotic microbes. This study is emerging as a new field for cataloguing species identification in microbial communities. COI was found to be efficient in species identification in different groups of eukaryotic microbes, such as ciliates, dinoflagellates, diatoms and amoebozoans (Lin *et al.*, 2009, Evans and Mann 2009, Heger *et al.*, 2011). Several DNA barcodes are being explored for identification of unknown and cryptic species of diatoms. Diatoms are the largest group of microalgae varying in shape, size and pattern during the life cycle with only <1 per cent species having full known cycle. *rbcL* and *LSU* rDNA have been proposed as universally acceptable barcode marker for diatoms (Mann *et al.*, 2010).

The identification of protist lineages is quite promising due to lack of species-specific morphological traits. Single gene cannot be sufficient for proper identification of protists. DNA barcoding is a novel approach for species demilitation in protists using different genes such as COI, 18S rRNA, 28S rRNA and ITS as barcode regions (Karnati *et al.*, 2003). It has been reported that COI is as an ideal barcode for amoebae and ITS2 for *Piroplasma* (Kosakyan *et al.*, 2012, Gou *et al.*,

2012). An improved barcoding system has been proposed for rapid and accurate identification of cryptic species (Hoef-Emden 2012). It has also been exploited for identification of freshwater and marine red algae of order Rhodophyta using COI sequences (Rueness 2010).

Methodology for DNA Barcoding

DNA barcoding methodology is a sequential process comprising of following steps- collection, identification and preservation of specimens, extraction of DNA from tissues, amplification of target gene using universal primers, sequencing and data analysis followed by submission to the database. The schema of methodology is depicted in Figure 2.3.

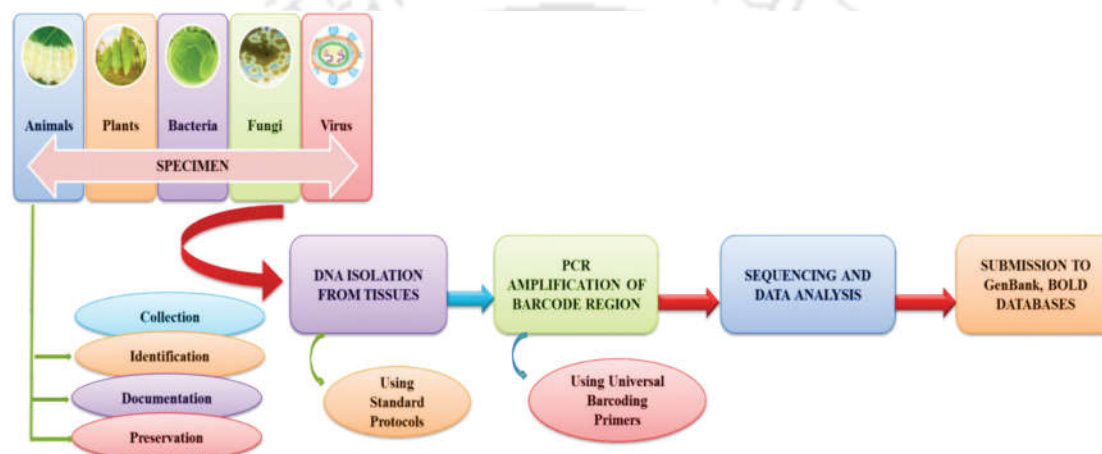


Figure 2.3: Schematic Representation of Methodology for DNA Barcoding.

Specimen Collection, Documentation and Preservation

Specimens (either fresh or preserved ones) can be collected from the repositories of biological materials such as natural history museums, herbaria, aquaria, seed banks etc and get identified through taxonomists. The identified specimens are then preserved with proper documentation (collection date, site, taxonomic description, photographic records etc.).

DNA Isolation

Total genomic/organellar DNA is isolated from the collected samples using standard DNA extraction protocols varying upon the type of specimen (Sharma *et al.*, 2010, Singh *et al.*, 2011). The quality of the extracted DNA is determined by gel electrophoresis and UV spectrophotometer in terms of A260/280 ratio. The ratio of 1.8, 2.0 and 0.6 indicates pure DNA, RNA and protein respectively. The extracted DNA can be stored at -20 °C for future use and in -80°C for longer storage.

PCR Amplification and Sequencing of Target Gene

Various universal barcoding primers have been reported for the amplification of the target gene. In case any of these primers fail to amplify, suitable primers can

be designed from the conserved sequences registered in GenBank. The target genes/ barcode regions are amplified by PCR using organelle-specific universal primers and then verified by gel electrophoresis. Concentration of the gel varies with the size of target barcode region. Generally, 1-2 per cent gel is preferred for quantitative and qualitative analysis of amplicons. The relevant PCR amplified products are then sequenced to obtain exact DNA barcode sequences from these specimens. The data are then analysed for subsequent submission in a public domain database (NCBI, BOLD *etc.*).

Submission to Database

Sequence analysis and phylogenetic tree construction is done using various appropriate bioinformatics tools (BioEdit, ORFinder, CLUSTAL W, MEGA 5.2.2 *etc.*). After proper annotation, sequence is submitted to NCBI GenBank database (<http://www.ncbi.nlm.nih.gov/WebSub/?tool=barcode>) using sequence submission platform (Sequin/BanKit). The barcode sequences can also be submitted as projects in BOLD databases (<http://www.boldsystems.org>), which are characterized by an automatic submission tool for publishing sequences to GenBank. Conservation of entire morphological reference for species is one of the conditions for data submission to BOLD.

Analysis of DNA Barcode Data

The unknown specimens are identified by comparing the closest similar reference record in the database. A Data Analysis Working Group has been created by CBOL to develop the methods for analysing and maintenance of DNA barcode data.

Applications of DNA Barcoding

DNA barcodes can be useful in many fields such as ecology, biomedicine, epidemiology, evolutionary biology, biogeography, conservation biology, bio-industry and biomonitoring. Some of the applications of DNA barcoding are as follows:

1. It facilitates the discovery of many unknown species and identification of pathogenic species with ecological, medical and agronomical significance and thus serves as a promising tool in epidemiology studies (Ball and Armstrong 2006, Nzelu *et al.*, 2015, Gumiel *et al.*, 2015).
2. Further, barcoding system aids in recognition, detection and to trace dispersal of patented organisms in agro-biotechnology, either to certify the source organism *e.g.* truffles or to secure intellectual property rights for bio- resources (Rastogi *et al.*, 2007, Taberlet *et al.*, 2007).
3. DNA barcode is an important tool to associate adults with immature specimens (fish larvae, fungal sexual stage) and in cases, when morphological traits fail to discriminate species (*e.g.* red algal species and fungal species) (Pegg *et al.*, 2006, Shenoy *et al.*, 2007).

4. It is useful in identification of the cryptic species. The existence of many “cryptic” fungal species, which are morphologically alike, emphasizes that molecular information is essential (Sato *et al.*, 2008).
5. This approach helps in identification of species interactions. For example, it has been used to study the host specificity of Australian beetles and their interaction with plants (Jurado-Rivera *et al.*, 2009).
6. It has been useful in identifying species-specific contaminants in fish consumed as food that may cause health problems (Lowenstein *et al.*, 2009).
7. Authentication of seafood species, which are commercially important having high market value, is one of the areas where DNA barcoding has been attempted. This approach aims to circumvent several issues related to food safety, species misbranding and accurate identification and conservation of wildlife and sustainable fishery fields (Pinto *et al.*, 2013).
8. It is widely used in food traceability to identify biological specimens for both raw and processed food (Galimberti *et al.*, 2013).
9. Authentication of natural health products (NHPs), derived from animals or plant origin has become possible by reliable DNA barcoding method. This reduces the chances of misrepresentation and substitution of unprotected species in the commercial market (Wallace *et al.*, 2012).
10. ‘Biomonitoring’ or ‘bioassessment’ is the method of evaluation for overall ecological health of surrounding environment (e.g. streams, reservoirs, wetlands, estuaries) on the basis of change in the composition and structure of the benthic (bottom dwelling) community. DNA barcoding using COI has also recently gained much interest in the field of environmental biomonitoring by providing a level of data standardization that lacked in the previous environmental assessments using aquatic life forms such as in-stream alga (diatoms or soft-bodied algae) or benthic invertebrates (Sweeney *et al.*, 2011, Pilgrim *et al.*, 2011).
11. Authentication and quality control of aesthetically potent materials (plants or animals) can be executed during DNA barcoding tools. An integrated MMDBD (Medicinal Materials DNA Barcode Database) have been developed for the collection and storage of DNA barcodes of these materials and act as a resource for researchers of various fields such as conservation, forensic, systematic study and herbal industry (Lou *et al.*, 2010).
12. Identification of traditional medicinal plants can also be performed using barcodes which is essential for proper medicinal production, research and application (Huang *et al.*, 2015).

Current Status of DNA Barcoding

The concept of DNA barcoding gained immense support to revolutionize taxonomy and species discovery (Blaxter 2005). Till date, a total of 4353038 DNA barcode sequences have been recorded in BOLD that satisfied DNA barcoding

criteria. A taxonomic bias has been observed in DNA barcoding due to limited research in search of universal barcode and classifying all live forms. The animal kingdom offered the maximum success rate (over 90 per cent) with Insecta representing around 46 per cent (Figure 2.4). Among insects, mostly, the progress has been made in the order Lepidoptera. Currently, more than 50000 barcode data have been collected which represent around 50 per cent of the total lepidopteran species (Lepidoptera-BOL). The proportion of barcode sequences is limited for plants, fungi and protists based on the BOLD results (BOLD search on 22-08-2015).

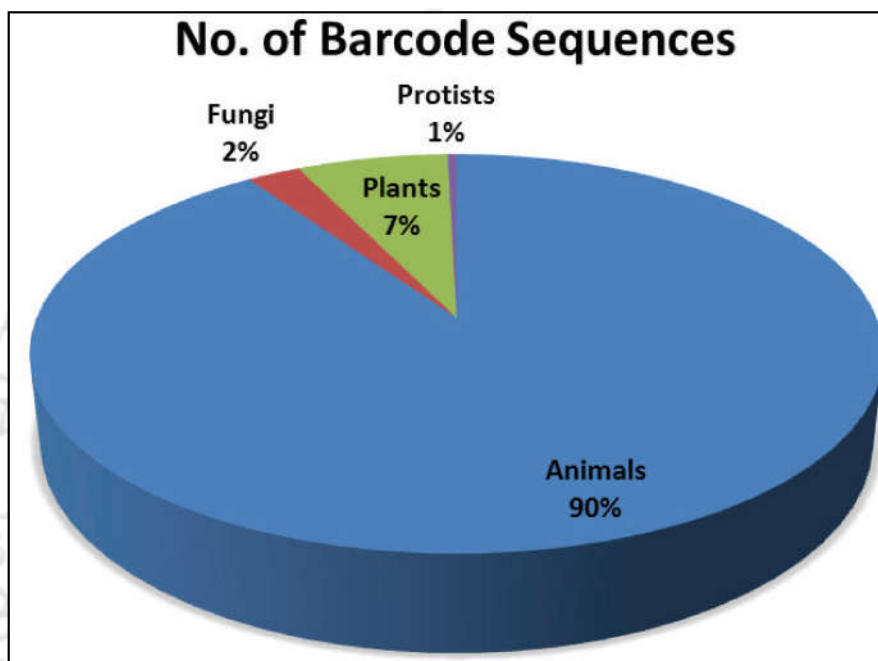


Figure 2.4: Taxonomic Bias in DNA Barcoding
(Data obtained from BOLD database as on 22 August, 2015).

Presently, DNA barcoding is a global enterprise, operating three international databases (BOLD, CBOL and iBOL) for construction of DNA barcode libraries that will enable researchers for better understanding and biodiversity conservation. Several campaigns have been initiated for constructing DNA barcode libraries for various taxa (Table 2.3). Resolving technical, analytical and fundamental issues is a prerequisite for data submission in BOLD. Unlike other well-known sequence libraries (*e.g.* NCBI), the sequences deposited in BOLD can be revised and taxonomically classified.

There is a hefty progress in scientific assessment for efficient identification of earthworm using barcoding approach. “Barcoding Earthworm” campaign (Earthworm- BOL) assembles ~5000 barcode sequences representing 800 species from more than 50 different countries and this number is projected to rise by next few years (Decaëns *et al.*, 2013). To date, there are numerous publications on the topic with 2930 papers entitled ‘DNA barcoding’ (Figure 2.5).

Table 2.3: List of DNA Barcoding Consortium and Campaigns Initiated for Barcoding of Various Taxa

Main International DNA barcoding Consortium	
International Barcode of Life (iBOL)	Established in 2010 for rapid and automated identification of organisms based on DNA barcode library.
Consortium for the Barcode of Life (CBOL)	Established in 2004 that aims to enable development of DNA barcoding and to construct DNA barcoding libraries.
The Barcode of Life Data Systems (BOLD)	Established in 2005 that acts as an interface for submission of DNA sequences to GenBank.
Campaigns for DNA barcoding of various taxa launched by CBOL and iBOL	
<i>Organism</i>	<i>Barcoding Campaigns</i>
Fishes	Fish Barcode of Life Initiative (Fish-BOL)
Sharks	Shark Barcode of Life
Moths and butterflies	Lepidoptera Barcode of Life
Saturniids	Lepidoptera barcode of life: Saturniidae
Ants	Formicidae Barcode of Life
Caddisflies	Trichoptera Barcode of Life
Birds	All Bird Barcoding Initiative
Mammals	Mammalia Barcode of Life

There are several studies which can demonstrate DNA barcoding as a tool to classify the biome distinctly. DNA barcoding separates 10 cryptic species from morphologically indistinguishable skipper butterflies (Hebert *et al.*, 2004). In another

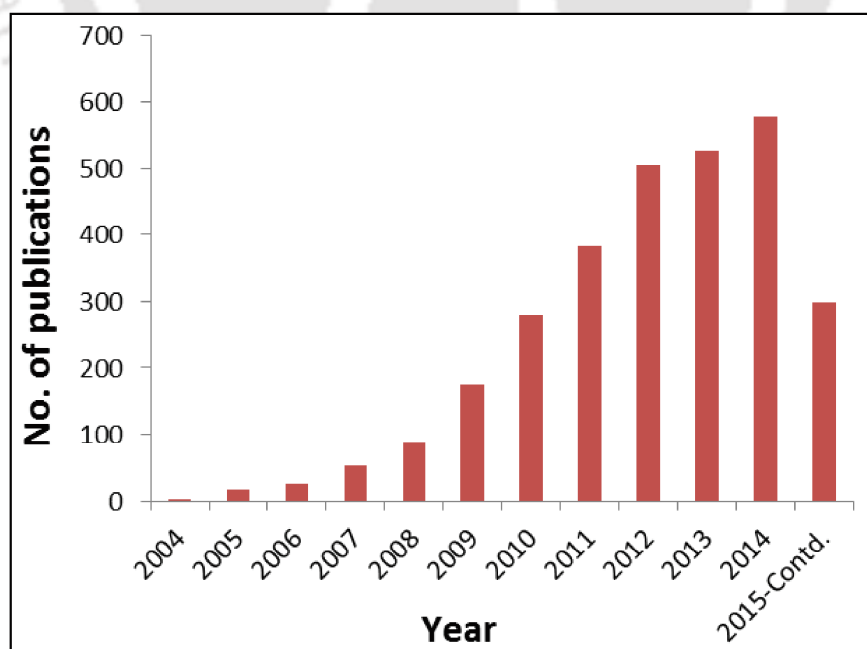


Figure 2.5: Trend of Publications Entitled 'DNA Barcoding'
(Search from Scopus database as on 22 August, 2015).

study, it has identified cryptic species from morphologically identical parasitoid flies (Tachinidae) (Smith *et al.*, 2006). It can also be used to identify adult and larval stages within species (Rasmussen *et al.*, 2009).

So far, COI gene has been proved as a universal barcode for various taxa including gastropods, springtails, butterflies, mayflies, spiders, ants, birds, crustaceans, reptiles, fish and recently, diatoms and protists (Ward *et al.*, 2005, Smith *et al.*, 2005, Hajibabaei *et al.*, 2006, Costa *et al.*, 2007, Gou *et al.*, 2012, Huang *et al.*, 2015). It has also been used to effectively discriminate among species (>95 per cent) in three families of Lepidoptera- Hesperidae (skipper butterflies), Sphingidae (sphinx moths), and Saturniidae (wild silk moths) (Hajibabaei *et al.*, 2006).

Use of a multi-locus approach, including COI- region and independent markers (*e.g.* nuclear subunit ribosomal RNA genes) is progressively emerging for species identification.

Advantages of DNA Barcoding

The advantages of DNA barcoding are described in the following section.

Studying Species Biology and Ecology

DNA barcoding has emerged as an effective tool in ecology and species biology. This approach is now widely used for understanding species interactions and detecting the presence of elusive and endangered species (Pfenninger *et al.*, 2007). For example, long range wolf colonization of France and Switzerland has been found out using extracted DNA barcodes from faeces and hair samples (Valie're *et al.*, 2003).

Moreover, the tool is widely used to elucidate the transmission pathways to study the interactions among the beetles (Lecythidaceae) and their endosymbiotic yeasts (Berkov *et al.*, 2007). It allowed the detection of North American bullfrog (*Rana catesbeiana*), one of the world's worst invasive species in southwest France using water samples from ponds (Ficetola *et al.*, 2008). The ecology and biology of fungus, invasive plant species and microbes has also been studied using this powerful tool (Scriver *et al.*, 2015). Thus, DNA barcoding has immense potential to solve important ecological problems.

Barcoding : A Multidisciplinary and Technically Advanced Approach

The prerequisite of DNA barcoding approach is to assemble the reference DNA barcode data which in turn, depends on the conventional methods of DNA isolation, amplification followed by sequencing. The reference library is extremely useful in rapid identification of low taxonomic level taxa using specific short-DNA sequences *i.e.* mini-barcode 100 bp, 300 bp (Min and Hickey, 2007).

An increasing number of techniques (*e.g.* silicon-based microarrays, nylon-membrane-based macroarrays, bio-barcode amplification assay (BCA), fluorescent DNA barcode-based immunoassay and the Luminex system of DNA-tagged polystyrene beads (patented) sorted by flow cytometry) have recently been developed as a next step of DNA barcoding. Silicon-platform microarray was first

used to study fungal systems and BCA for virus systems (Summerbell *et al.*, 2005, Cao *et al.*, 2009, Yin *et al.*, 2011).

High-throughput sequencing enables the most efficient means of rapid barcode-based species identification (in case of mixed samples such as stomach contents, food, and blood or water columns). SASI-Seq (Sample assurance Spike-Ins) is one of the simple and inexpensive approaches to check the samples quality. Sample mix up or identification of cross-contaminants could be easily achieved through high throughput sequencing methods (Quail *et al.*, 2014). The full viral genome sequencing has also been carried out using multi-faceted approach that includes nucleic acid preparation, Illumina Seq and a novel iterative sequence classification algorithm (Cotten *et al.*, 2014).

Moreover, the use of cutting edge sequencing techniques in metagenomics could be promising in DNA barcoding as a new approach called metabarcoding (Hudson 2008, Reva *et al.*, 2015). It can be used for exploration of both culturable as well as non-culturable microbes. Conservation of entire morphological reference for species is one of the conditions for submission of data to BOLD database for which new techniques have been developed for extraction of non-destructive DNA from the specimens (Pook and McEwing 2005, Hunter *et al.*, 2008, Rowley *et al.*, 2007).

Limitations of DNA Barcoding

Inspite of the immense potential offered by DNA barcoding, it has some drawbacks which cannot be overlooked and are discussed in the following section.

Exploration of Microdiversity

Barcoding the microdiversity (*i.e.* archaea, bacteria, unicellular fungi and protists) is one of the major challenges for the BOL objectives (Chantangsi *et al.*, 2007). The diversity of microscopic and sub-microscopic communities is masked by the dominant populations which lead to unexploration of microbes (Sogin *et al.*, 2006). It can be explored using COI-based identification system but for macroscopic life, multi-locus barcoding system may be needed (Blaxter *et al.*, 2005, Evans *et al.*, 2007).

The use of SSU has been reported in identification of human parasite *Blastocystis hominis* (Scicluna *et al.*, 2006). Similarly other barcode regions (18S rRNA, 28S rRNA, ITS etc.) are being developed to study the diversity of protozoans, archaea etc. Advanced methods such as metagenomics and community genomics have been emerging as new fields to study microdiversity, population studies and evolutionary relationships (DeLong 2004, Tringe *et al.*, 2005). But the expensive metagenomics approach makes DNA barcoding as a useful tool for studying microdiversity.

No Universal Barcode for Some Taxa

COI serves as a universal DNA barcode in animals but its utility in protists, fungi and plants is quite challenging. Amphibians have high variability in mtDNA and its COI priming sites and therefore large-scale effort is needed to barcode the amphibians using the same primary barcode region of COI (Smith *et al.*, 2008). COI fails to distinguish closely related animal species requiring nuclear regions *e.g.* *Cytb* and *Rhod* to identify all teleost fish species (Sevilla *et al.*, 2007).

In algae, there are shortcomings with the primers for COI and therefore, universal plastid amplicon (UPA) was reported as an alternative tool, although having a drawback of low interspecific divergence (Clarkston *et al.*, 2010). COI has failed in barcoding of some protists (some group of amoeba) and elicits the use of an alternative barcode (Gile *et al.*, 2010). It is also not suitable for DNA barcoding of plants and fungi due to low rate of molecular evolution (Chase and Fay, 2010).

To address all these issues, alternative stretches of DNA have been proposed as barcodes for the barcoding of these organisms. Combination of two plastid regions *matK* and *rbcL* is recommended as universal barcodes for terrestrial plants by CBOL Plant Working Group (2009). But it is not still widely accepted for all plant species (Hollingsworth *et al.*, 2011). Best results have been obtained using the internal transcribed spacer (ITS) region of the ribosomal DNA as a barcode for fungi. However, ITS offers less remarkable result in the DNA barcoding of *Penicillium* moulds (Seifert *et al.*, 2007). Additional markers (single or in combination) are now being explored to resolve the fungal diversity. The development of universal barcode is need of the hour towards the exploration of biodiversity.

Barcoding of Formalin Preserved Specimens

Preserved tissues and specimens are vast repositories of genetic information which form the basis for construction and development of barcode libraries. Formalin-preserved specimens provide a great resource for this approach (Zhang *et al.*, 2010). However, extraction of DNA and amplification of the target gene from these specimens is quite challenging (Zimmermann *et al.*, 2008).

Tissue lysis is the primary obstacle for formalin preserved tissues which affect the integrity of the extracted DNA. Formalin preserves morphological structure more effectively by fixing proteins through peptide linkage conjugation but degrades DNA (Kolesárová *et al.*, 2012). It also affects the downstream applications such as PCR amplification and sequencing reactions due to the crosslink formations by formaldehyde.

Unfavourable preservation condition is a major constraint on DNA quality (Strutzenberger *et al.*, 2012). Use of alternative preservatives such as ethanol is preferred than formalin as it denatures proteins (at high concentration $\sim$ 95 per cent) that may degrade DNA (Flournoy *et al.*, 1996). For DNA barcoding, ethanol at a concentration of 95 per cent is recommended for storage. Fresh sample and alcohol preservation is more preferable for DNA barcoding studies than formalin preservation to maintain maximum integrity (Nagy 2010, Baird *et al.*, 2011).

Presence of Barcoding Gaps

Constructing a database for DNA barcoding is essential where each taxon is optimally represented. Lack of reference data may lead to the mis-interpretation of barcode-based identification. Limited geographical sampling may lead to 'Barcoding gaps' which should be considered during the database construction phase. This may lead to failure of more recently diverged species (sister-species) identification (Meyer and Paulay 2005, Hebert *et al.*, 2003, Wiemers and Fiedler 2007). The unknown specimens may be problematic if belongs to under-described

part of biodiversity (Rubinoff 2006, Rubinoff *et al.*, 2006). Collaboration with expert taxonomists for identification of undocumented voucher specimens is essential for proper construction of barcode database.

Delimitation of Species

Species delimitation refers to the process which determines the species boundaries and discovery of new species. The accuracy of species identification using DNA barcoding has been criticised by several taxonomists that depends on the overlap between intraspecific and interspecific variation (Moritz and Cicero 2004, Lee 2004, Ebach and Holdrege 2005, Rubinoff *et al.*, 2006). This approach is based on the hypothesis that intraspecific divergences will be less than interspecific divergences which is absent in several taxa (Meyer and Paulay 2005, Cognato 2006, Meier *et al.*, 2006, Whitworth *et al.*, 2007). This implies a query whether DNA barcoding can attain its goal to unravel the biodiversity. It is possible only if the target region is conserved among the members of a species and differs from others (Prendini 2005).

Utility of Mitochondrial Genome

The utility of a single mitochondrial gene in DNA barcoding to delineate species boundaries has also triggered several criticisms. Classifications solely based on the one character (*i.e.* mitochondrial gene) are inadequate for species identification (Rubinoff *et al.*, 2006). Nuclear loci is only required to resolve phylogenetic relationships in several cases where one host species can bear different symbionts which may lead to intraspecific (*i.e.* interpopulation) variation in mtDNA sequences (Hurst *et al.*, 1999).

Paraphyly, polyphyly, heteroplasmy (*e.g.* Mussels) or presence of nuclear pseudogenes in mitochondrial genome *i.e.* nuclear mitochondrial DNA (NUMTs) are some of the artifacts associated with mitochondrial genes. This may cause the overestimation of sample divergence and thus obscure the real evolutionary history of organisms (Williams and Knowton 2001, Funk and Omland 2003, Meyer and Paulay 2005, Terranova *et al.*, 2007). It has been reported that heteroplasmy level differs among tissues and DNA extracted from large tissues displays less polymorphism (Magnacca and Brown, 2010).

Methodological Advances

There are two major drawbacks of DNA barcoding: (i) damage of specimen during DNA extraction method and (ii) nature of sample. A simple, rapid and reliable DNA extraction protocol is essential to obtain efficient yields of DNA for further process.

Various non-destructive DNA extraction protocols have been proposed which enable the determination of barcode with minimal damage of museum specimens or small insects (Acarina, Araneae, Coleoptera, Diptera, Hemiptera, Hymenoptera etc.) (Rowley *et al.*, 2007, Hunter *et al.*, 2008, Castalanelli *et al.*, 2010). However, these methods have been useful in some insects; DNA extraction is challenging in plants (*e.g.* Malvaceae, Asteraceae) having mucilaginous compounds, polysaccharides,

phenolic compounds and other secondary metabolites. These compounds co-precipitate with DNA and thus hinder the extraction of high quality DNA. In addition, they complicate the downstream applications such as polymerase chain reaction (PCR) or other enzymatic reactions.

To overcome these issues, several modified protocols have been reported that enable effective DNA extraction from problematic plant samples (Michiels 2003, Shepherd *et al.*, 2011, Kundu *et al.*, 2011, Souza *et al.*, 2012). Recent technical advancements such as highly efficient DNA polymerase and a DNA-repairing enzyme allow the extraction and amplification of DNA from historical museum specimens as well as fossilized samples (Hajibabaei *et al.*, 2005, Juen and Traugott, 2006, Ball and Armstrong, 2008).

Distance or Character Based Barcoding Analyses

Enumeration of molecular distance is proposed to be a standard method for barcode analysis (Hebert *et al.*, 2003). But it has lately been criticized and should not be considered as it is a phenetic measure non-indicative of common origin (Ward *et al.*, 2005, DeSalle *et al.*, 2005, Borisenko *et al.*, 2008, Kerr *et al.*, 2009).

On the contrary, character-based analysis has proven to be an effective tool for species identification and discrimination (Kelly *et al.*, 2007, Rach *et al.*, 2008, Bergmann *et al.*, 2009). It addresses all the objectives with no loss of information as multiple nucleotide differences are reduced to a single distance-based measure. This approach has been implemented through development of Characteristic Attributes Organization System (CAOS) software (Sarkar *et al.*, 2008). Some studies have been reported on character-based barcoding analysis but conflict still lies between distance and diagnostics (Locke *et al.*, 2010). However, the character-based barcoding is currently proposed to meet the requirement.

Conclusions and Future Prospects

DNA barcoding can be exploited to identify different species when it allows discriminating amongst closely related species, including sister taxa (Newmaster and Ragupathy 2009). Numerous studies focus on plethora of species with distant genetic relationships. It is still unclear whether this approach is restricted to species identification or species discovery and delimitation. Several studies are currently being performed for both intra- and inter species to clarify whether it can successfully delineate species rather than only higher taxonomic levels (*e.g.* genus or species groups).

In current era, data generation and analysis process is continuously evolving, especially from DNA and genomes which possibly transformed the biological sciences since past decades or so. But these advances have been incremental due to omnipresence of powerful computing system and market-driven activities to formulate cheaper and more advanced molecular biology procedures. Two paradigm shifts reported in the past decade have led to the advent of a new form of molecular systematics having a wide range of applications. These paradigms clarify how their convergence is influencing biodiversity studies and understanding of ecosystem (Hajibabaei, 2012).

The first paradigm shift occurred in 2003 employing 'DNA barcodes' for species identification. DNA barcoding has emerged as an initiative for the discovery of new species and retraining present era of taxonomy (DNA taxonomy). The paradigm shift lies in the transformation of the taxonomic method for unexplored microbiome, standardization of barcode although seen a significant advancement. This can be addressed by "Standard Signatures for DNA barcoding". Standards may continuously evolve over time, to address additional DNA information for resolving difficult taxonomic puzzles (Blaxter *et al.*, 2005, Ragupathy *et al.*, 2009, Pauls *et al.*, 2010).

Indeed, the use of only a single locus in barcoding research elicits contradictions and can result in the description of groups that are not monophyletic. It has been noted that DNA barcoding, in absence of other evidences can only suggest new species rather than outright discovering them (Goldstein and DeSalle 2011). Even a character-based methodology would struggle to identify the cryptic species. The advent of novel species using DNA barcoding is merely an iterative concept (Forister *et al.*, 2008). Thus, efforts should be made towards the improvements in this approach. Moreover, gold standard is yet to be developed with huge volume of data in order to create a genetic reference library (Lucking 2008, Packer *et al.*, 2009).

Barcoding- Towards Integrative Taxonomy

The purpose of barcoding coincides with that of traditional taxonomy and therefore should supplement the morphological data rather than replacing it. Incorporation of DNA and morphological data as an integrated taxonomy would certainly be a solution to the current crisis generated by single data resources. One of the main aims of this initiative is the discovery and description of new taxa which is difficult to accomplish using only sequence based data (Ebach and Holdrege 2005). This initiative would be helpful in various fields like recognition of cryptic species, validation of taxonomy thus facilitating the better understanding of biodiversity (Smith *et al.*, 2008, Hamada *et al.*, 2010).

An integrative taxonomy has been proposed to delimit the diversity of live forms in terms of morphology, ecology, development, behaviour, phylogeography etc. Species that are under-described would be most benefitted from this proposal. Thus, identification and characterization of species in a manner which best reflects the biodiversity is one of the challenges for the future of taxonomy (Dayrat 2005).

Integrative taxonomy is not a new concept. Wilson's "encyclopedia of life", a database was created in 2003 that includes information on morphology, DNA (EOL, available at <http://www.eol.org/>) and natural history of every species. Later, the significance of integrative taxonomy was recognized by DNA barcoding community (Smith *et al.*, 2008, Fisher and Smith 2008). This approach facilitated comprehensive and detailed understanding of parasitoid diversity and host specificity on the combined basis of morphology, DNA barcoding and ecology. Also it enabled DNA barcoding for the collection and storage of voucher specimens and Dor tissue samples for biodiversity research (Peterson *et al.*, 2007).

Thus the integration of information is an appropriate way to promote the scientific enterprise. The progress of an integrated approach is just one of the issues encountered by DNA barcoding that could be accelerated by the advent of NGS.

Application of NGS in DNA Barcoding

The second paradigm shift occurred by ground breaking advancements in DNA sequencing technologies. Traditional Sanger sequencing, invented in 1977, paves the way of breakthrough research in genomics, providing the provision of whole genome sequencing including the human genome. These technological advances minimized the labor-intensiveness and time for DNA sequencing, without altering the core Sanger-based method. But the emergence of NGS technologies revolutionized the DNA sequencing concept in the mid-2000s.

The advancement of novel technologies can perform parallel sequencing, producing thousands or millions of sequences at a single run. Several platforms like Roche 454 pyrosequencing, Illumina Hi Seq®, SOLiD sequencing, Helicos' HeliScope of modern era NGS are much relevant for DNA barcoding due to longer average read length. There is a huge difference in the amount of sequence information as compared to the classic Sanger method (Hall, 2007).

The technology has achieved some success in the investigation of single nucleotide polymorphisms, meiofaunal assemblages and microbial eukaryotes (Creer *et al.*, 2010, Stoeck *et al.*, 2010). The field of 'ecological genomics' has also been proposed to be an emerging discipline accessing the full genome data (Tautz *et al.*, 2010).

Furthermore, it has become the first and foremost choice for many areas of microbial research due to its ability to unveil the microbial communities (Schuster 2008). There is a rapid increment in the number of publications using high-throughput sequencing in fungal studies (Lim *et al.*, 2010, Dumbrell *et al.*, 2011). Now a steady flow of advancements allow DNA to be rapidly sequenced at ever-increasing volumes with simultaneous reduction in the cost per sequenced base.

DNA barcoding alone fails in the field of 'environmental barcoding' which aims at exploring large amount of environmental samples such as soil, water etc. With the increasing data generation (single or whole genome sequence) and lack of proper tools to organize, verify and analyse the data, the DNA barcoding cannot become a standalone discipline.

The utility of NGS for 'mini-barcodes' (even smaller fragments of DNA) in species identification have been suggested by DNA barcoding exponents, but this seems illogical particularly for species delimitation (Hajibabaei *et al.*, 2011). The Canadian Centre for DNA Barcoding *i.e.* the hub of the barcoding effort continues to support Sanger sequencing as one of the studies on techniques and logistics that focusses entirely on the 'high-throughput' 96-well microplates (Borisenko *et al.*, 2009).

Application of NGS in DNA barcoding will accelerate the process of exploration of diversity of various life forms. It has revolutionized the viral discovery using multiple barcodes for viral genome detection and discovery (Cotten *et al.*, 2014).

Also an emerging approach termed as “DNA metabarcoding” is being implied in the exploration of microbiome (both cultured and uncultured). It is a robust method which implies barcode based taxonomic structure, functional and composition exploration through high throughput sequencing technologies (Reva *et al.*, 2015).

Since the advent of NGS and ecological genomics technology might influence DNA barcoding with the risk of becoming irrelevant. It is thus essential to find a way to evolve their methods to encompass these emerging developmental tools. To that end, a combinatorial approach *i.e.* integration of taxonomic, molecular and application of NGS platforms can be a better solution to unravel the huge diversity among various taxa. This will facilitate easy and error-free analyses in areas like taxonomic profiling, phylogenetics, population genetics, computational biology etc.

Further Suggested DNA Barcoding Resources

Other DNA barcode resources are described in the following section.

DNA Barcode Databases

With the advancement of information technology, various databases and projects have been established for proper documentation of biodiversity, some of the public databases are enlisted.

Barcode Related Websites

1. Barcodes of Life (<http://www.barcodeoflife.org>)
2. Canadian Centre for DNA Barcoding (<http://www.dnabarcoding.ca/>),
3. ECBOL - European Consortium for the Barcode of Life (<http://www.ecbol.org/>)
4. BOLNET.ca- Canadian Barcode of Life Network (<http://www.bolnet.ca/>)
5. iBOL-International Barcode of Life Project (<http://www.dnabarcoding.org/>)
6. All Birds Barcoding Initiative (<http://www.barcodingbirds.org/>)
7. FISH-BOL: Fish Barcode of Life Initiative (<http://www.fishbol.org/>)
8. All Leps - Barcode of Life (<http://www.lepbarcoding.org/>)
9. BOLD- Barcode of Life Database (<http://www.barcodinglife.org/>)
10. MarBOL - Marine Barcode of Life (<http://www.marinebarcoding.org/>)
11. Sponge Barcoding Project (<http://www.spongebarcoding.org/>)
12. Polar Barcode of Life (<http://www.polarbarcoding.ca/>)
13. Belgian Network for DNA Barcoding (<http://www.bebol.myspecies.info>)
14. CBOL- Consortium for the Barcode of Life (<http://www.barcoding.si.edu>)
15. NorBOL (<http://dnabarcoding.no/en/>)
16. MexBOL (<http://www.mexbol.org/>)
17. JBOLI (<http://www.jboli.org/>)

Books on DNA Barcoding

1. Miller FP, Vandome AF and McBrewster J. DNA barcoding: taxonomy, species, molecular phylogenetics, consortium for the barcode of life, identification (biology), eukaryote, mitochondrion. VDM Publishing. 2009.
2. Ortman BD. DNA Barcoding the Medusozoa and Ctenophora, Doctoral Dissertations. Paper AAI3345202. 2008.
3. Yu CR. Data Analysis for microarray experiment and DNA barcode of life. UMI Dissertation Publishing. 2008.
4. Ghosh SK. A textbook on DNA barcoding. Book Space, Kolkata. 2012.

References

- Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rrn* operons. *J. Bacteriol.* 186: 2629–2635. 2004.
- Adekambi T, Drancourt M, Raoult D. The *rpoB* gene as a tool for clinical microbiologists. *Trends Microbiol.* 17: 37–45. 2009.
- Alvarez I, Wendel JF. Ribosomal ITS sequences and plant phylogenetic inference. *Mol. Phylogenet Evol.* 29: 417–434. 2003.
- Anslan S, Tedersoo L. Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. *Euro. J. Soil Biol.* 69: 1–7. 2015.
- Aquilino SVL, Tango JM, Fontanilla IKC, Pagulayan RC *et al.*, DNA barcoding of the ichthyofauna of Taal Lake, Philippines. *Mol. Ecol. Resour.*, 11(4): 612–619. 2011.
- Arfi Y, Bue'e M, Marchand C, Levasseur A *et al.*, Multiple markers pyrosequencing reveals highly diverse and host specific fungal communities on the mangrove trees *Avicennia marina* and *Rhizophora stylosa*. *FEMS Microbiol. Ecol.* 79: 433–444. 2012.
- Bafeel SO, Arif IA, Bakir MA, Al Homaidan AA *et al.*, DNA barcoding of arid wild plants using *rbcL* gene sequences. *Genet. Mol. Res.* 11(3): 1934–1941. 2012.
- Baird DJ, Pascoe TJ, Zhou X, Hajibabaei M. Building freshwater macroinvertebrate DNA-barcode libraries from reference collection material: formalin preservation vs specimen age. *J. North Amer. Benthol. Soc.* 30(1): 125–130. 2011.
- Ball SL, Armstrong KF. DNA barcodes for insect pest identification: a test case with tussock moths (Lepidoptera: Lymantriidae). *Can. J. For. Res.* 36: 337–350. 2006.
- Ball SL, Armstrong KF. Rapid, one-step DNA extraction for insect pest identification by using DNA barcodes. *J. Econ. Entomol.* 101: 523–532. 2008.
- Ball SL, Hebert PDN, Burian SK, Webb JM. Biological identifications of mayflies (Ephemeroptera) using DNA barcodes. *J. North Amer. Benthol. Soc.* 24: 508–524. 2005.

- Benlloch S, Lopez-Lopez A, Casamayor EO, Øvreås L *et al.*, Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. *Environ. Microbiol.* 4: 349–360. 2002.
- Bennasar A, Mora RR, Lalucat J, Moore ERB. 16S rRNA gene sequence analysis relative to genomovars of *Pseudomonas stutzeri* and proposal of *Pseudomonas balearica* sp nov. *Intl. J. Syst. Bacteriol.* 46: 200–205. 1996.
- Bennett CE, Wilson BS, Desalle R. DNA barcoding of an invasive mammal species, the small Indian mongoose (*Herpestes javanicus*; E. Geoffroy Saint-Hillaire 1818) in the Caribbean and Hawaiian Islands. *Mitochondrial DNA.* 22(1–2): 12–18. 2011.
- Bergmann T, Hadrys H, Breves G, Schierwater B. Character-based DNA barcoding: a superior tool for species classification. *Berliner Und Munchener Tierarztliche Wochenschrift.* 122: 446–450. 2009.
- Berkov A, Feinstein J, Small J, Nkamany M. Yeasts isolated from neotropical wood-boring beetles in SE Peru. *Biotropica.* 39: 530–538. 2007.
- Blaxter M, Mann J, Chapman T, Thomas F *et al.*, Defining operational taxonomic units using DNA barcode data. *Philos. Trans. R. Soc. Lond. B.* 360: 1935–1943. 2005.
- Blaxter M. The Promise of DNA taxonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 359: 669–679. 2004.
- Borisenko AV, Lim BK, Ivanova NV, Hanner RH *et al.*, DNA barcoding in surveys of small mammal communities: a field study in Suriname. *Mol. Ecol. Resour.* 8: 471–479. 2008.
- Borisenko AV, Sones JE and Hebert PDN. The front-end logistics of DNA barcoding: challenges and prospects. *Mol. Ecol. Resour.* 9: 27–34. 2009.
- Boucher Y, Douady CJ, Sharma AK, Kamekura M *et al.*, Intragenomic heterogeneity and intergenomic recombination among haloarchaeal ribosomal RNA genes. *J. Bacteriol.* 186(12): 3980–90. 2004.
- Bruni I, De Mattia F, Martellos S, Galimberti A *et al.*, DNA Barcoding as an effective tool in improving a digital plant identification system: A Case Study for the Area of Mt. Valerio, Trieste (NE Italy). *Plos One.* 7(9): e43256. 2012.
- Busse HJ, Denner EBM, Lubitz W. Classification and identification of bacteria: Current approaches to an old problem. Overview of methods used in bacterial systematics. *J. Biotech.* 47: 3–38. 1996.
- Campbell CS, Wright WA, Cox M, Vining TF *et al.*, Nuclear ribosomal DNA internal transcribed spacer 1 (*ITS1*) in *Picea* (Pinaceae): sequence divergence and structure. *Mol. Phylogenet. Evol.* 35: 165–185. 2005.
- Cao C, Dhumpa R, Bang DD, Ghavifekr Z *et al.*, Detection of avian influenza virus by fluorescent DNA barcode based immunoassay with sensitivity comparable to PCR. *Analyst.* 135: 337–342. 2009.
- Case RJ, Boucher Y, Dahllöf I, Holmstrom C *et al.*, Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* 173: 278–288. 2007.

- Castalanelli MA, Severtson DL, Brumley CJ, Szito A *et al.*, A rapid non-destructive DNA extraction method for insects and other arthropods. *J. Asia Pacific Entomol.* 13: 243–248. 2010.
- Chantangsi C, Lynn DH, Brandl MT, Cole JC *et al.*, Barcoding ciliates: a comprehensive study of 75 isolates of the genus *Tetrahymena*. *Intl. J. Syst. Evol. Microbiol.* 57: 2412–2425. 2007.
- Chase MW, Fay MF. Barcoding of plants and fungi. *Science.* 325: 682–683. 2010.
- Chase MW, Salamin N, Wilkinson M, Dunwell JM *et al.*, Land plants and DNA barcodes: short-term and long-term goals. *Philos. Trans. Royal Soc. Ser. B.* 360: 1889–1895. 2005.
- Che J, Chen HM, Yang JX, Jin JQ *et al.*, Universal COI primers for DNA barcoding amphibians. *Mol. Ecol. Resour.* 12(2): 247–258. 2012.
- Chen S, Yao H, Han J, Liu C *et al.*, Validation of the ITS2 Region as a Novel DNA Barcode for Identifying Medicinal Plant Species. *Plos One.* 5(1): e8613. 2010.
- Cheng RH, Kuhn RJ, Olson NH, Rossman MG *et al.*, Nucleocapsid and glycoprotein organization in an enveloped virus. *Cell.* 80: 621–630. 1995.
- Clare EL, Lim BK, Engstrom MD, Eger JL *et al.*, DNA bar-coding of Neotropical bats: species identification and discovery within Guyana. *Mol. Ecol. Notes.* 7: 184–190. 2007.
- Clarkston BE, Saunders GW. A comparison of two DNA barcode markers for species discrimination in the red algal family Kallymeniaceae (Gigartinales, Florideophyceae), with a description of *Euthoratimburtonii* sp nov. *Botany-Botanique.* 88: 119–131. 2010.
- Cognato AI. Standard percent DNA sequence difference for insects does not predict species boundaries. *J. Econ. Entomol.* 99: 1037–1045. 2006.
- Costa FO, deWaard JR, Boutillier J, Ratnasingham S *et al.*, Biological identifications through DNA barcodes: the case of the Crustacea. *Can. J. Fish Aquat. Sci.* 64: 272–295. 2007.
- Cotten M, Oude Munnink B, Canuti M, Deijs M *et al.*, Full genome virus detection in fecal samples using sensitive nucleic acid preparation, deep sequencing, and a novel iterative sequence classification algorithm. *PLoS ONE.* 9 (4): e93269. 2014.
- Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM *et al.*, Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises. *Mol. Ecol.* 19: 4–20. 2010.
- Dayrat B. Towards integrative taxonomy. *Biol. J. Linn. Soc.* 85: 407–415. 2005.
- Decaëns T, Porco D, Rougerie R, Brown GG *et al.*, Potential of DNA barcoding for earthworm research in taxonomy and ecology. *Appl. Soil Ecol.* 65: 35– 42. 2013.
- DeLong EF. Microbial population genomics and ecology: the road ahead. *Environ. Microbiol.* 6 (9): 875–878. 2004.
- DeSalle R, Egan MG, Siddall M. The unholy trinity: taxonomy, species delimitation and DNA barcoding. *Philosophil. Trans. Royal Soc. B-Biol. Sci.* 360: 1905–1916. 2005.

- Doyle JJ, Gaut BS. Evolution of genes and taxa: a primer. *Plant Mol. Biol.* 42: 1–6. 2000.
- Dumbrell AJ, Ashton PD, Aziz N, Feng G *et al.*, Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytol.* 190: 794–804. 2011.
- Dunning LT, Savolainen V. Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Bot. J. Linnean Soc.* 164(1): 1-9. 2010.
- Ebach, MC, Holdrege C. DNA barcoding is no substitute for taxonomy. *Nature.* 434: 697. 2005.
- Eggert LS, Rasner CA, Woodruff DS. The evolution and phylogeography of the African elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. *Proc. Royl. Soc. Lond. Ser. B-Biol. Sci.* 269: 1993–2006. 2002.
- Evans KM, Mann DG. A proposed protocol for nomenclatural effective DNA barcoding of microalgae. *Phycologia*, 48: 70–74. 2009.
- Evans KM, Wortley AH, Mann DG. An assessment of potential diatom “barcode” genes (*cox1*, *rbcL*, *18S* and *ITS rDNA*) and their effectiveness in determining relationships in *Sellaphora* (Bacillariophyta). *Protist.* 158: 349–364. 2007.
- Ficetola GF, Miaud C, Pompanon F, Taberlet P. Species detection using environmental DNA from water samples. *Biol. Lett.* 4: 423–425. 2008.
- Fisher BL, Smith MA. A Revision of Malagasy Species of *Anochetus* Mayr and *Odontomachus* Latreille (Hymenoptera: Formicidae). *Plos One.* 3(5): e1787. 2008.
- Flournoy LE, Adams RP, Pandey RN. Interim and archival preservation of plant specimens in alcohols for DNA studies. *Biotechniques.* 20: 657-660. 1996.
- Floyd R, Abebe E, Papert A, Blaxter M. Molecular barcodes for soil nematode identification. *Mol. Ecol.* 11: 839–850. 2002.
- Folmer O, Black M, Hoeh W, Lutz R *et al.*, DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3: 294–299. 1994.
- Forister ML, Nice CC, Fordyce JA, Gompert Z *et al.*, Considering evolutionary processes in the use of single-locus genetic data for conservation, with examples from the Lepidoptera. *J. Insect Conserv.* 12: 37–51. 2008.
- Frisvad JC, Samson RA. Polyphasic taxonomy of *Penicillium* subgenus *Penicillium*—a guide to identification of food and air-borne terverticillate penicillia and their mycotoxins. *Stud. Mycol.* 49: 1–173. 2004.
- Funk DJ, Omland KE. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Ann. Rev. Ecol. Evol. S.* 34: 397-423. 2003.
- Galimberti A, De Mattia F, Losa A, Bruni I *et al.*, DNA barcoding as a new tool for food traceability. *Food Res. Intl.* 50(1): 55-63. 2013.
- Garnica S, Weiß M, Oertel B, Ammirati J *et al.*, Phylogenetic relationships in *Cortinarius*, section *Calochroi*, inferred from nuclear DNA sequences. *BMC Evol. Biol.* 9: 1. 2009.

- Geml J, Laursen GA, Timling I, McFarland JM *et al.*, Molecular phylogenetic biodiversity assessment of arctic and boreal ectomycorrhizal *Lactarius Pers.* (Russulales; Basidiomycota) in Alaska, based on soil and sporocarp DNA. *Mol. Ecol.* 18: 2213–2227. 2009.
- Gielly L, Taberlet P. The use of chloroplast DNA to resolve plant phylogeny: noncoding versus *rbcL* sequences. *Mol. Biol. Evol.* 11: 769–777. 1994.
- Gile GH, Stern RF, James ER, Keeling PJ. DNA barcoding of chlorarachniophytes using nucleomorph ITS sequences. *J. Phycol.* 46: 743–750. 2010.
- Goldstein PZ, DeSalle R. Integrating DNA barcode data and taxonomic practice: determination, discovery, and description. *Bioessays*, 33: 135–147. 2011.
- Gómez A, Wright PJ, Lunt DH, Cancino JM *et al.*, Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. *Philos. Trans. Royl. Soc. B.* 274: 199–207. 2007.
- Gou H, Guan G, Liu A, Ma M *et al.*, A DNA barcode for *Piroplasma*. *Acta Trop.* 124: 92–97. 2012.
- Gumiel M, Da Mota FF, Rizzo VDS, Sarquis O *et al.*, Characterization of the microbiota in the guts of *Triatoma brasiliensis* and *Triatoma pseudomaculata* infected by *Trypanosoma cruzi* in natural conditions using culture independent methods. *Parasites Vectors.* 8 (1): 245. 2015.
- Hajibabaei M, deWaard J, Ivanova NV, Ratnasingham S *et al.*, Critical factors for assembling a high volume of DNA barcodes. *Philos. Trans. Royl. Soc. Lond. Series B.* 360: 1959–1967. 2005.
- Hajibabaei M, Janzen DH, Burns JM, Hallwachs W *et al.*, DNA barcodes distinguish species of tropical Lepidoptera. *Proc. Natl. Acad. Sci. USA.* 103: 968–971. 2006.
- Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23: 167–172. 2007.
- Hajibabaei M. The golden age of DNA metasystematics. *Trends Genet.* 28(11): 535–537. 2012.
- Hajibabaei MHM, Shokralla S, Zhou X, Singer GAC *et al.*, Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *Plos One.* 6: e17497. 2011.
- Hall N. Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.* 210: 1518–1525. 2007.
- Hamada N, Pepinelli M, Mattos-Glória A, Luz SLB. A new black fly species from Brazil, closely related to *Simulium guianense* Wise (Diptera, Simuliidae), revealed by morphology and DNA barcoding. *Zootaxa*, 2428: 22–36. 2010.
- Hanner R, Floyd R, Bernard A, Collette BB *et al.*, DNA barcoding of billfishes. *Mitochondl. DNA.* 22(1): 27–36. 2011.

- Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. *Proc. Royl. Soc. Lond. B. Bio.* 270: 313–322. 2003.
- Hebert PDN, Penton EH, Burns JM, Janzen DH *et al.*, Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci. USA.* 101: 14812–14817. 2004.
- Heger TJ, Pawlowski J, Larac E, Leandere BS *et al.*, Comparing potential COI and SSU rDNA barcodes for assessing the diversity and phylogenetic relationships of *Cyphoderiid testate* Amoebae (Rhizaria: Euglyphida). *Protist.* 162, 131–141. 2011.
- Hibbett DS, Ohman A, Glotzer D, Nuhn M *et al.*, Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biol. Rev.* 25: 38–47. 2011.
- Hillis DM, Dixon MT. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.* 66: 411–453. 1991.
- Hoef-Emden K. Pitfalls of establishing DNA barcoding systems in protists: the Cryptophyceae as a test case. *Plos One.* 7(8): e43652. 2012.
- Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. *Plos One.* 6: e19254. 2011.
- Holmes DE, Nevin KP, Lovley DR. Comparison of 16S rRNA, *nifD*, *recA*, *gyrB*, *rpoB* and *fusA* genes within the family *Geobacteraceae* fam. nov. *Intl. J. Syst. Evolul. Microbiol.* 54: 1591–1599. 2004.
- Huang Y, Zhang Y, Zhao C, Xu Y *et al.*, DNA barcoding and its utility in commonly-used medicinal snakes. *China J. Chinese materia medica.* 40(5): 868–874. 2015.
- Hubert N, Hanner R, Holm E, Mandrak NE *et al.*, Identifying Canadian freshwater fishes through DNA barcodes. *Plos One.* 3(6): e2490. 2008.
- Hudson ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. *Mol. Ecol. Resour.* 8: 3–17. 2008.
- Hunter SJ, Goodall TI, Walsh KA, Owen R *et al.*, Nondestructive DNA extraction from blackflies (Diptera: Simuliidae): retaining voucher specimens for DNA barcoding projects. *Mol. Ecol. Resour.* 8(1): 56–61. 2008.
- Hurst GDD, Jiggins FM, von der Schulenburg JHG, Bertrand D *et al.*, Male-killing *Wolbachia* in two species of insect. *Proc. Roy. Soc. B.* 266: 735–740. 1999.
- Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN. Universal primer cocktails for fish DNA barcoding. *Mol. Ecol. Notes.* 7: 544–548. 2007.
- Jinbo U, Kato T, ITO M. Current progress in DNA barcoding and future implications for entomology. *Entomol. Sci.* 14(2): 107–124. 2011.
- Juen A, Traugott M. Amplification facilitators and multiplex PCR: tools to overcome PCR-inhibition in DNA-gut-content analysis of soil-living invertebrates. *Soil Biol. Biochem.* 38: 1872–1879. 2006.
- Jurado-Rivera JA, Vogler AP, Reid CA, Petitpierre E *et al.*, DNA barcoding insect-host plant associations. *Proc. Biol. Sci.* 276(1657): 639–648. 2009.

- Karnati SK, Yu Z, Sylvester JT, Dehority BA *et al.*, Technical note: specific PCR amplification of protozoal 18S rDNA sequences from DNA extracted from ruminal samples of cows. *J. Animl. Sci.* 81: 812–815. 2003.
- Kelly LJ, Ameka GK, Chase MW. DNA barcoding of African Podostemaceae (river-weeds): a test of proposed barcode regions. *Taxon.* 59(1): 251-260. 2010.
- Kelly RP, Sarkar IN, Eernisse DJ, Desalle R. DNA barcoding using chitons (genus *Mopalia*). *Mol. Ecol. Notes.* 7: 177–183. 2007.
- Kerr KCR, Lijtmaer DA, Barreira AS, Hebert PDN *et al.*, Probing evolutionary patterns in neotropical birds through DNA barcodes. *Plos One.* 4: e4379. 2009.
- Kerr KCR, Stoeckle MY, Dove CJ, Weigt LA *et al.*, Comprehensive DNA barcode coverage of North American birds. *Mol. Ecol. Notes*, 7: 535–543. 2007.
- Ki J-S, Zhang W, Qian P-Y. Discovery of marine *Bacillus* species by 16S rRNA and *rpoB* comparisons and their usefulness for species identification. *J. Microbiol. Methd.* 77: 48–57. 2009.
- Kjer KM. Use of rRNA secondary structure in phylo-genetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. *Mol. Phylogenet. Evol.* 4: 314–330. 1995.
- Klarica J, Kloss-Brandstätter A, Traugott M, Juen A. Comparing four mitochondrial genes in earthworms – Implications for identification, phylogenetics, and discovery of cryptic species. *Soil Biol. Biochem.* 45: 23e30. 2012.
- Koetschan C, Förster F, Keller A, Schleicher T *et al.*, The ITS2 Database III - sequences and structures for phylogeny. *Nucleic Acids Res.* 38: D275-9. 2010.
- Kolesárová M, Herich R, Levkut M, Èurlík J *et al.*, Suitability of different tissue fixatives for subsequent PCR analysis of *Cysticercus ovis*. *Helminthologia.* 49: 67-70. 2012.
- Kosakyan A, Heger TJ, Leander BS, Todorov M *et al.*, COI barcoding of Nebelid testate amoebae (Amoebozoa: Arcellinida): extensive cryptic diversity and redefinition of the Hyalospheniidae Schultze. *Protist.* 163: 415–434. 2012.
- Kress WJ, Erickson DL. A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. *Plos One.* 2: e508. 2007.
- Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA *et al.*, Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. USA.* 102: 8369–8374. 2005.
- Kruck NC, Tibbetts IR, Ward RD, Johnson JW *et al.*, Multi-gene barcoding to discriminate sibling species within a morphologically difficult fish genus species (*Sillago*). *Fisheries Res.* 143: 39-46. 2013.
- Kumar CP, John BA, Khan SA, Lyla PS *et al.*, Mutational patterns and DNA barcode for diagnosing chikungunya virus. *Intl. J. Virol.* 7: 42-52. 2011.
- Kumar NP, Rajavel AR and Jambulingam P. Application of PDF417 symbology for DNA barcoding. *Comp. Methd. Prog. Biomed.* 90: 187–189. 2008.

- Kundu A, Sarkar D, Bhattacharjee A, Topdar N *et al.*, A simple ethanol wash of the tissue homogenates recovers high-quality genomic DNA from *Corchorus* species characterized by highly acidic and proteinaceous mucilages. *Electro. J. Biotech.* 14(1). 2011.
- Lahaye R, van der Bank M, Bogarin D, Warner J *et al.*, DNA barcoding the floras of biodiversity hotspots. *Proc. Natl. Acad. Sci. USA.* 105: 2923–2928. 2008.
- Lara A, de Leon JLP, Rodriguez R, Casane D *et al.*, DNA barcoding of Cuban freshwater fishes: evidence for cryptic species and taxonomic conflicts. *Mol. Ecol. Resour.* 10(3): 421–430. 2010.
- Larranaga N, Hormaza JI. DNA barcoding of perennial fruit tree species of agronomic interest in the genus *Annona* (Annonaceae). *Front. Plant Sci.* 6: 589. 2015.
- Lee DH, Lee HJ, Lee YN, Lee YJ *et al.*, Application of DNA barcoding technique in avian influenza virus surveillance of wild bird habitats in Korea and Mongolia. *Avian Diseases.* 54(1): 677–681. 2010.
- Lee MSY. The molecularisation of taxonomy. *Invertebr. Syst.*, 18: 1–6. 2004.
- Lim YW, Kim BK, Kim C, Jung HS *et al.*, Assessment of soil fungal communities using pyrosequencing. *J. Microbiol.* 48: 284–289. 2010.
- Lin S, Zhang H, Hou YB, Zhuang YY *et al.*, High-level diversity of dinoflagellates in the natural environment, revealed by assessment of mitochondrial *cox1* and *cob* genes for dinoflagellate DNA barcoding. *Appl. Environ. Microbiol.* 75: 1279–1290. 2009.
- Links MG, Dumonceaux TJ, Hemmingsen SM, Hill JE. The chaperonin-60 universal target is a barcode for bacteria that enables de novo assembly of metagenomic sequence data. *PLoS One.* 7: e49755. 2012.
- Liu C, Shi L, Xu X, Li H *et al.*, DNA Barcode goes two-dimensions: DNA QR Code web server. *Plos One.* 7(5): e35146. 2012.
- Locke SA, McLaughlin JD, Marcogliese DJ. DNA barcodes show cryptic diversity and a potential physiological basis for host specificity among Diplostomoidea (Platyhelminthes: Digenea) parasitizing freshwater fishes in the St. Lawrence river, Canada. *Mol. Ecol.* 19: 2813–2827. 2010.
- Lorenz JG, Jackson WE, Beck JC, Hanner R. The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philos. Trans. Royl. Soc. Lond. B Biol. Sci.* 360: 1869–1877. 2005.
- Lou SK, Wong KL, Li M, But PP *et al.*, An integrated web medicinal materials DNA database: MMDBD (Medicinal materials DNA barcode database). *BMC Genomics.* 11: 402. 2010.
- Lowenstein JH, Amato G, Kolokotronis SO. The real maccoyii: identifying tuna sushi with DNA barcodes – contrasting characteristic attributes and genetic distances. *Plos One.* 4: e7866. 2009.

- Lucking R. Taxonomy: a discipline on the brink of extinction are DNA barcode scanners the future of biodiversity research? *Archives Des. Sci.* 61: 75–88. 2008.
- Magnacca KN, Brown MJF. Mitochondrial heteroplasmy and DNA barcoding in *Hawaiian hylaeus* (Nesoprosopis) bees (Hymenoptera: Colletidae). *BMC Evol. Biol.* 10: 174. 2010.
- Mann DG, Sato S, Trobajo R, Vanormelingen P *et al.*, DNA barcoding for species identification and discovery in diatoms. *Cryptogamie Algologie.* 31(4): 557–577. 2010.
- Martin P, Kaygorodova I, Sherbakov DY, Verheyen E. Rapid evolving lineages impede the resolution of phylogenetic relationships among Clitellata (Annelida). *Mol. Phylogenet. Evol.* 15: 355–368. 2000.
- Martinez-Murcia AJ, Benlloch S, Collins MD. Phylogenetic interrelationships of members of the genera *Aeromonas* and *Plesiomonas* as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations. *Intl. J. Syst. Bacteriol.* 42: 412–421. 1992.
- McLaughlin DJ, Hibbett DS, Lutzoni F, Spatafora JW *et al.*, The search for the fungal tree of life. *Trends Microbiol.* 17: 488–497. 2009.
- Meier R, Shiyang K, Vaidya G, NG PKL. DNA Barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. *Syst. Biol.* 55(5): 715–728. 2006.
- Meyer CP, Paulay G. DNA barcoding: error rates based on comprehensive sampling. *Plos Biol.* 3: 2229–2238. 2005.
- Michel H, Behr J, Harrenga A, Kannt A. Cytochrome c oxidase: structure and spectroscopy. *Ann. Rev. Biophys. Biomol. Struct.* 27: 329–356. 1998.
- Michiels A, Ende WV, Tucker M, Riet LV *et al.*, Extraction of high-quality genomic DNA from latex-containing plants. *Analytical Biochem.* 315: 85–89. 2003.
- Min XJ, Hickey DA. Assessing the effect of varying sequence length on DNA barcoding of fungi. *Mol. Ecol. Notes.* 7: 365–373. 2007.
- Mollet C, Drancourt M, Raoult D. Determination of *Coxiella burnetii* *rpoB* sequence and its use for phylogenetic analysis. *Gene.* 207: 97–103. 1998.
- Moritz C, Cicero C. DNA barcoding: promise and pitfalls. *Plos Biol.* 2: 1529–1531. 2004.
- Mouhamadou B, Carriconde F, Gryta H, Jargeat P *et al.*, Molecular evolution of mitochondrial ribosomal DNA in the fungal genus *Tricholoma*: Barcoding implications. *Fungal Genet. Biol.* 45: 1219–1226. 2008.
- Nagy ZT. A hands-on overview of tissue preservation methods for molecular genetic analyses. *Organism Diversity Evol.* 10: 91–105. 2010.
- Nelson JS. Fishes of the World, 4th ed., John Wiley and Sons, Hoboken, NJ. 2006.
- Newmaster SG, Ragupathy S. Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae). *Mol. Ecol. Resour.* 9 (1): 172–180. 2009.

- Newmaster SG, Fazekas AJ, Ragupathy S. DNA barcoding in land plants: evaluation of *rbcL* in a multigene tiered approach. *Can. J. Bot.* 84: 335–341. 2006.
- Nijman V, Aliabadian M. Performance of distance-based DNA barcoding in the molecular identification of Primates. *Comptes Rendus Biologies.* 333(1): 11-16. 2010.
- Nilsson R, Ryberg M, Abarenkov K, Sjökvist E *et al.*, The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiol. Lett.* 296: 97-101. 2009.
- Nzelu CO, Cáceres AG, Arrunátegui-Jiménez MJ, Lañas-Rosas MF *et al.*, DNA barcoding for identification of sand fly species (Diptera: Psychodidae) from leishmaniasis-endemic areas of Peru. *Acta Tropica*, 145: 45-51. 2015.
- O'Donnell K, Sutton DA, Rinaldi MG, Sarver BA *et al.*, Internet-accessible DNA sequence database for identifying fusaria from human and animal infections. *J. Clinl. Microbiol.* 48: 3708–3718. 2010.
- Packer L, Grixti JC, Roughley RE, Hanner R. The status of taxonomy in Canada and the impact of DNA barcoding. *Can. J. Zoology.* 87: 1097–1110. 2009.
- Padial JM, De La Riva I. Integrative taxonomists should use and produce DNA barcodes. *Zootaxa*, 1586: 67-68. 2007.
- Padial JM, Castroviejo-Fisher S, Köhler J, Vilà C *et al.*, Deciphering the products of evolution at the species level: the need for an integrative taxonomy. *Zool. Scripta*, 38: 431-447. 2009.
- Palmer JD, Herbon LA. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *J. Mol. Evol.* 28: 87–97. 1988.
- Pang XH, Song JY, Zhu YJ, Xie CX *et al.*, Using DNA Barcoding to Identify Species within Euphorbiaceae. *Planta Medica.* 76(15): 1784-1786. 2010.
- Pauls SU, Blahnik RJ, Zhou X, Wardwell CT *et al.*, DNA barcode data confirm new species and reveal cryptic diversity in *Chilean smicridea* (Smicridea) (Trichoptera: Hydropsychidae). *J. North Amer. Benthological Soc.* 29: 1058–1074. 2010.
- Pegg GG, Sinclair B, Briskey L and Aspden WJ. MtDNA barcode identification of fish larvae in the southern great barrier reef. *Australian Sci.* 70: 7–12. 2006.
- Peterson AT, Moyle RG, Nyari AS, Robbins MB *et al.*, The need for proper voucherizing in phylogenetic studies of birds. *Mol. Phylogenet. Evol.* 45: 1042–1044. 2007.
- Pfenninger M, Nowak C, Kley C, Steinke D *et al.*, Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. *Mol. Ecol.* 16: 1957–1968. 2007.
- Pilgrim EM, Jackson S, Swenson S, Turcsanyi I *et al.*, Incorporation of DNA barcoding into a large-scale biomonitoring program: opportunities and pitfalls. *J. North Amer. Benthological Soc.* 30: 217–231. 2011.
- Pinto AD, Pinto PD, Terio V, Bozzo G *et al.*, DNA barcoding for detecting market substitution in salted cod fillets and battered cod chunks. *Food Chem.* 141: 1757–1762. 2013.

- Pook CE, McEwing R. Mitochondrial DNA sequences from dried snake venom: a DNA barcoding approach to the identification of venom samples. *Toxicon*. 46: 711–715. 2005.
- Prendini L. Comment on “Identifying spiders through DNA barcodes”. *Can. J. Zool.* 83: 481–491. 2005.
- Quail MA, Smith M, Jackson D, Leonard S *et al.*, SASI-Seq: Sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing. *BMC Genomics*. 15 (1): 110. 2014.
- Rach J, DeSalle R, Sarkar IN, Schierwater B *et al.*, Character based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc. Royl. Soc. Lond. Ser. B-Biol. Sci.* 275: 237–247. 2008.
- Ragupathy S, Newmaster SG, Murugesan M, Balasubramaniam V. DNA barcoding discriminates a new cryptic grass species revealed in an ethnobotany study by the hill tribes of the western ghats in southern India. *Mol. Ecol. Resour.* 9 (1): 164–171. 2009.
- Rasmussen RS, Morrissey MT, Hebert PDN. DNA barcoding of commercially important salmon and trout species (*Oncorhynchus* and *Salmo*) from north America. *J. Agril. Food Chem.* 57: 8379–8385. 2009.
- Rastogi G, Dharne MS, Walujkar S, Kumar A *et al.*, Species identification and authentication of tissues of animal origin using mitochondrial and nuclear markers. *Meat Sci.*, 76: 666–674. 2007.
- Ratnasingham S and Hebert PDN. Barcoding bold: The barcode of life data system (www.barcodinglife.org). *Mol. Ecol. Notes*. 1–10. 2007.
- Renner SS. Circumscription and phylogeny of the Laurales: Evidence from molecular and morphological data. *Amer. J. Bot.* 86: 1301–1315. 1999.
- Reva ON, Zaets IE, Ovcharenko LP, Kukharensko OE *et al.*, Metabarcoding of the kombucha microbial community grown in different microenvironments. *AMB Express*. 5 (1): 124. 2015.
- Richert K, Brambilla E, Stackebrandt E. The phylogenetic significance of peptidoglycan types: Molecular analysis of the genera *Microbacterium* and *Aureobacterium* based upon sequence comparison of *gyrB*, *rpoB*, *recA* and *ppk* and 16S rRNA genes. *Syst. Appl. Microbiol.*, 30: 102–108. 2007.
- Robert V, Szöke S, Eberhardt U, Cardinali G *et al.*, The quest for a general and reliable fungal DNA barcode. *Open Appl. Inform. J.* 5: 55–71. 2011.
- Robinson EA, Blagoev GA, Hebert PDN, Adamowicz SJ. Prospects for using DNA barcoding to identify spiders in species-rich genera. *Zookeys*. 16: 27–46. 2009.
- Roux V, Raoult D. Phylogenetic analysis of the genus *Rickettsia* by 16S rDNA sequencing. *Res. Microbiol.* 146: 85–396. 1995.
- Rowley DL, Coddington JA, Gates MW, Norrbom AL *et al.*, Vouchering DNA-barcoded specimens: test of a non-destructive extraction protocol for terrestrial arthropods. *Mol. Ecol. Notes*. 7: 915–924. 2007.

- Rubioff D, Cameron S, Will K. Are plant DNA barcodes a search for the Holy Grail? *Trends Ecol. Evol.* 21: 1–2. 2006.
- Rubioff D. Utility of mitochondrial DNA barcodes in species conservation. *Conserv. Biol.* 20: 1026–1033. 2006.
- Rueness J. DNA barcoding of select freshwater and marine red algae (Rhodophyta). *Cryptogamie Algologie.* 31(4): 377–386. 2010.
- Saccone C, DeCarla G, Gissi C, Pesole G *et al.*, Evolutionary genomics in the Metazoa: the mitochondrial DNA as a model system. *Gene.* 238: 195–210. 1999.
- Salazar GA, Chase MW, Arenas MAS, Ingrouille M. Phylogenetics of *Cranichideae* with emphasis on *Spiranthinae* (Orchidaceae, Orchidoideae): evidence from plastid and nuclear DNA sequences. *Amer. J. Bot.* 90: 777–795. 2003.
- Sarkar IN, Planet PJ, Desalle R. CAOS software for use in character-based DNA barcoding. *Mol. Ecol. Resour.* 8: 1256–1259. 2008.
- Sass C, Little DP, Stevenson DW, Specht CD. DNA barcoding in the Cycadales: testing the potential of proposed barcoding markers for species identification of Cycads. *Plos One.* 2: e1154. 2007.
- Sato H, Murakami N. Reproductive isolation among cryptic species in the ectomycorrhizal genus *Strobilomyces*: Population-level caps marker-based genetic analysis. *Mol. Phyl. Evol.* 48: 326–334. 2008.
- Schoch CL, Seifert KA, Huhndorf S, Robert V *et al.*, Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proc. Natl. Acad. Sci. USA.* 109: 6241–6246. 2012.
- Schoch CL, Sung GH, López-Giráldez F, Townsend JP *et al.*, The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst. Biol.* 58: 224–239. 2009.
- Schultz J, Müller T, Achtziger M, Seibel PN *et al.*, The internal transcribed spacer 2 database – a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Res.* 34: W704–707. 2006.
- Schuster SC. Next-generation sequencing transforms today's biology. *Nature.* 5: 16–18. 2008.
- Scicluna SM, Tawari B, Clark CG. DNA barcoding of *Blastocystis*. *Protist.* 157: 77–85. 2006.
- Scriver M, Marinich A, Wilson C, Freeland J. Development of species-specific environmental DNA (eDNA) markers for invasive aquatic plants. *Aquatic Bot.* 122: 27–31. 2015.
- Seena S, Pascoal C, Marvanova L, Cassio F. DNA barcoding of fungi: a case study using ITS sequences for identifying aquatic hyphomycete species. *Fungal Diversity.* 44(1), 77–87. 2010.
- Seifert KA, Samson RA, deWaard JR, Houbraken J *et al.*, Prospects for fungus identification using CO1 DNA barcodes, with *Penicillium* as a test case. *Proc. Natl. Acad. Sci. USA.* 104(10): 3901–6. 2007.

- Seifert KA. Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour.* 9: 3-89. 2009.
- Sevilla RG, Diez A, Noren M, Mouchel O *et al.*, Primers and polymerase chain reaction conditions for DNA barcoding teleost fish based on the mitochondrial cytochrome b and nuclear rhodopsin genes. *Mol. Ecol. Notes.* 7: 730-734. 2007.
- Sharma P, Joshi N, Sharma A. Isolation of genomic DNA from medicinal plants without liquid nitrogen. *Ind. J. Exptl. Biol.* 48: 610-614. 2010.
- Shaw J, Lickey EB, Beck JT, Farmer SS *et al.*, The tortoise and the hare: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *Amer. J. Bot.* 92: 142-166. 2005.
- Shenoy BD, Jeewon R, Hyde KA. Impact of DNA sequence-data on the taxonomy of anamorphic fungi. *Fungal Diversity* 26: 1-54. 2007.
- Shepherd LD, McLay TGB. Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. *J. Plant Res.*, 124: 311-314. 2011.
- Silva-Brandão KL, Lyra ML, Freitas AVL. Barcoding Lepidoptera: current situation and perspectives on the usefulness of a contentious technique. *Neotropical Entomol.* 38: 441-451. 2009.
- Simmons RB, Weller SJ. Utility and evolution of cytochrome *b* in insects. *Mol. Phylogenet. Evol.* 20: 196-210. 2001.
- Singh HR, Unii BG, Neog K, Wann SB. Isolating silkworm genomic DNA without liquid nitrogen suitable for marker studies. *African J. Biotech.* 10(55): 11365-11372. 2011.
- Smith MA, Bertrand C, Crosby K, Eveleigh ES *et al.*, *Wolbachia* and DNA barcoding insects: patterns, potential, and problems. *PLoS One.* 7: e36514. 2012.
- Smith MA, Fisher BL, Hebert PDN. DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philos. Trans Royl. Soc. Ser. B.* 360: 825-1834. 2005.
- Smith MA, Poyarkov NA, Hebert PDN. CO1 DNA barcoding amphibians: take the chance, meet the challenge. *Mol. Ecol. Resour.* 8(2): 235-246. 2008.
- Smith MA, Woodley NE, Janzen DH, Hallwachs W *et al.*, DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). *Proc. Natl. Acad. Sci. USA.* 103: 3657-3662. 2006.
- Sogin ML, Morrison HG, Huber JA, Welch DM *et al.*, Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc. Natl. Acad. Sci. USA.* 103: 12115-12120. 2006.
- Souza HAV, Muller LAC, Brandão RL, Lovato MB. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. *Genet. Mol. Res.* 11 (1): 756-764. 2012.
- Steinke D, Zemlak TS, Hebert PDN. Barcoding Nemo: DNA-Based identifications for the ornamental fish Trade. *Plos One.* 4: e6300. 2009.

- Stockinger H, Kruger M, Schussler A. DNA barcoding of arbuscular mycorrhizal fungi. *New Phytologist*. 187(2): 461-474. 2010.
- Stoeck T, Bass D, Nebel M, Christen R *et al.*, Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* 19: 21–31. 2010.
- Strutzenberger P, Brehm G, Fiedler K. DNA barcode sequencing from old type specimens as a tool in taxonomy: a case study in the diverse genus *Eois* (Lepidoptera: Geometridae). *Plos One*. 7(11): e49710. 2012.
- Summerbell RC, Lévesque CA, Seifert KA, Bovers M *et al.*, Microcoding: the second step in DNA barcoding. *Philos. Trans. Royl. Soc. Lond. B. Biol. Sci.* 360: 1897–1903. 2005.
- Swartz ER, Mwale M, Hanner R. A role for barcoding in the study of African fish diversity and conservation. *South African J. Sci.*, 104: 293–298. 2008.
- Sweeney BW, Battle JM, Jackson JK, Dapkey T. Can DNA barcodes of stream macroinvertebrates improve descriptions of community structure and water quality? *J. North Amer. Bentholological Soc.* 30: 195–216. 2011.
- Taberlet P, Coissac E, Pompanon F, Gielly L *et al.*, Power and limitations of the chloroplast *trnL* (UAA) intron for plant DNA barcoding. *Nucleic Acids Res.* 35: e14. 2007.
- Tautz D, Ellegren H, Weigel D. Next generation molecular ecology. *Mol. Ecol.* 19: 1–3. 2010.
- Terranova MS, Brutto SL, Arculeo M, Mitton JB. A mitochondrial phylogeography of *Brachidontes variabilis* (Bivalvia: Mytilidae) reveals three cryptic species. *J. Zool. Syst. Evol. Res.* 45 (4): 289–298. 2007.
- Toju H, Tanabe AS, Yamamoto S, Sato H. High-Coverage ITS Primers for the DNA-Based Identification of *Ascomycetes* and *Basidiomycetes* in Environmental Samples. *Plos One*. 7(7): e40863. 2012.
- Triantafyllidis A, Bobori D, Koliamitra C, Gbandi E *et al.*, DNA barcoding analysis of fish species diversity in four north Greek lakes. *Mitochondl. DNA*. 22(1): 37-42. 2011.
- Tringe SG, von Mering C, Kobayashi A, Salamov AA *et al.*, Comparative metagenomics of microbial communities. *Science*. 308: 554–557. 2005.
- Tripathi AM, Tyagi A, Kumar A, Singh A *et al.*, The internal transcribed spacer (ITS) region and *trnH-psbA* are suitable candidate loci for DNA barcoding of tropical tree species of India. *Plos One*. 8(2): e57934. 2013.
- Uhart M, Sirand-Pugnet P, Labarere J. Evolution of mitochondrial *SSUrDNA* variable domain sequences and rRNA secondary structures, and phylogeny of the *Agrocybe aegerita* multispecies complex. *Res. Microbiol.* 158: 203–212. 2007.
- Valière N, Fumagalli L, Gielly L, Miquel C *et al.*, Long-distance wolf recolonization of France and Switzerland inferred from non-invasive genetic sampling over a period of 10 years. *Animl. Conseron.* 6: 83-92. 2003.

- Vijayan K, Tsou CH. DNA barcoding in plants: taxonomy in a new perspective. *Current Sci.* 99(1): 1530-1541. 2010.
- Vilgalys R, Gonzalez D. Organisation of ribosomal DNA in the basidiomycete *Thanatephorus praticola*. *Current Genet.* 18: 277-280. 1990.
- Vos M, Quince C, Pijl AS, de Hollander M *et al.*, A comparison of *rpoB* and 16S *rRNA* as markers in pyrosequencing studies of bacterial diversity. *Plos One.* 7(2): e30600. 2012.
- Wallace LJ, Boilard SMAL, Eagle SHC, Spall JL *et al.*, DNA barcodes for everyday life: Routine authentication of Natural Health Products. *Food Res. Intl* 49: 446–452. 2012.
- Walsh DA, Baptiste E, Kamekura M, Doolittle WF. Evolution of the RNA polymerase B9 subunit Gene (*rpoB9*) in Halobacteriales: a complementary molecular marker to the SSU *rRNA* Gene. *Mol. Biol. Evol.* 21(12): 2340–2351. 2012.
- Wang WQ, Wu YR, Yan YH, Ermakova M *et al.*, DNA barcoding of the Lemnaceae, a family of aquatic monocots. *BMC Plant Biol.* 10: 205. 2010.
- Ward CW. Progress towards a higher taxonomy of viruses. *Res. Virol.* 144: 419-453. 1993.
- Ward RD, Zemlak TS, Innes BH, Last PR *et al.*, DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. Lond. B.* 360: 1847–1857. 2005.
- Whitworth TL, Dawson RD, Magalon H, Baudry E. DNA barcoding cannot reliably identify species of the blowfly genus *Protophormia* (Diptera: Calliphoridae). *Proc. Royl. Soc. B.* 274: 1731-1739. 2007.
- Wiemers M, Fiedler K. Does the DNA barcoding gap exist? – a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers Zool.* 4. 2007.
- Williams ST, Knowlton N. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. *Mol. Biol. Evol.* 18: 1484-1493. 2001.
- Wilson DE, Reeder DAM. Mammal species of the world: A taxonomic and geographic reference. 3rd ed. Johns Hopkins University Press, Baltimore, MA. 2005.
- Woo PC, Lau SK, Teng JL, Tse H *et al.*, Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinl. Microbiol Infect.* 14: 908–934. 2008.
- Yamashiro T, Fukuda T, Yokoyama J, Maki M. Molecular phylogeny of *Vincetoxicum* (Apocynaceae-Asclepiadoideae) based on the nucleotide sequences of cpDNA and nrDNA. *Mol. Phylogenet. Evol.* 31: 689-700. 2004.
- Yang QQ, Zhao S, Kucerova Z, Stejskal V *et al.*, Validation of the 16S rDNA and COI DNA barcoding technique for rapid molecular identification of stored product psocids (Insecta: Psocodea: Liposcelididae). *J. Econ. Entomol.* 106(1): 419-425. 2013.

- Yang YW, Lai KN, Li WH. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* 48: 597–604. 1999.
- Yin HQ, Jia MX, Shi LJ, Yang S *et al.*, Nanoparticle-based bio-barcode assay for the detection of bluetongue virus. *J. Virol. Methd.* 178: 225–228. 2011.
- Yu J, Xue JH, Zhou SL. New universal *matK* primers for DNA barcoding angiosperms. *J. Syst. Evol.* 49(3): 176-181. 2011.
- Zhang J. Exploiting formalin-preserved fish specimens for resources of DNA barcoding. *Mol. Ecol. Resour.* 10(6): 935-41. 2010.
- Zhao P, Luo J, Zhuang WY. Practice towards DNA barcoding of the nectriaceous fungi. *Fungal Diversity.* 46(1): 183-191. 2011.
- Zimmermann J, Hajibabaei M, Blackburn DC, Hanken J *et al.*, DNA damage in preserved specimens and tissue samples: a molecular assessment. *Frontiers Zool.* 5: 18. 2008.



Nanotechnology in Crop Protection

*Sunita Ojha, Deepika Singh, Arghya Sett, Hasnahana Chetia,
Debajyoti Kabiraj, Utpal Bora*

Indian Institute of Technology Guwahati, Guwahati, India

16.1 INTRODUCTION

Nanoparticles (NPs) are particles in a nanorange structure with unique optical, magnetic, electrical, and thermal properties (Luo et al., 2015). Compared to their bulk materials they have more strength with superior conductivity and reactivity (Annadhasan et al., 2014). The large number of surface atoms alters the surface-related properties of the particle in their nanosize. These surface atoms make these NPs very reactive for use as catalysts. In addition, high surface energy and spatial confinement properties of the NPs aid in quantum effects. Quantum confinement alters the energy band structure and charge carrier density, which in turn alters the optical and electronic properties (Buzea et al., 2007; Roduner, 2006).

Nanotechnology utilizing these nanomaterials is a multifaceted technology that has been applied in material science, electronics, energy sectors, biotechnology, medicine, and many other sectors. In agriculture, it has gained momentum over time with an abundance of public funding and government policies to combat food crises. Numerous scientific reports and patents filed in this field suggest the advancement of nanotechnology in crop protection and disease management (Parisi et al., 2015). In farming, nanointervention improves yield by increasing the efficiency of nutrient uptake and protects crops from pests through nanoformulations of pesticides. Development of new-generation pesticides with better carrier systems and decontamination of water and soil using nanotechnology also helps in crop protection. Moreover, bionanotechnology helps in better understanding host–parasite interaction at the molecular level. Preservation and packaging of food are some of the other applications of nanotechnology. It is a boon to the concept of precision farming, which aims for targeted delivery, firm attachment, and controlled release of the active material. Encapsulation or entrapment of agrochemicals using polymers and dendrimers by surface ionic attachment helps in their controlled release and improves solubility and stability. Because of their antimicrobial activity, nanoparticles of metals such as silver, zinc oxide, and titanium oxide, and nonmetals such as silica and sulfur can be used to eradicate bactericide- and fungicide-resistant pathogens (Chowdappa and Gowda, 2013; Aziz et al., 2015, 2016).



Research paper

De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer)



Hasnahana Chetia^a, Debajyoti Kabiraj^a, Deepika Singh^a, Ponnala Vimal Mosahari^b, Suradip Das^a, Pragma Sharma^c, Kartik Neog^d, Swagata Sharma^a, P. Jayaprakash^e, Utpal Bora^{a,b,*}

^a Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam 781039, India

^b Centre for the Environment, Indian Institute of Technology Guwahati, Assam 781039, India

^c Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati 781014, Assam, India

^d Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, 785700 Jorhat, Assam, India

^e Central Silk Board (CSB), Bangalore 560068, Karnataka, India

ARTICLE INFO

Article history:

Received 20 October 2016

Received in revised form 29 January 2017

Accepted 15 February 2017

Available online 17 February 2017

Keywords:

Alimentary canal

Antimicrobial peptide

Lepidoptera

Machilus bombycina

Next generation sequencing

Residual body

Saturniidae

Sericin

Silk gland

Silk gland factor

ABSTRACT

Antheraea assamensis (Lepidoptera: Saturniidae), is a semi-domesticated silkworm known to be endemic to Assam and the adjoining hilly areas of Northeast India. It is the only producer of a unique, commercially important variety of golden silk called “muga silk”. Herein, we report the *de novo* transcriptome of *A. assamensis* reared on *Machilus bombycina* leaves for the first time. Short reads generated by high throughput sequencing of cDNA libraries from multiple tissues, viz. alimentary canal, silk gland and residual body of the 5th instar of muga silkworm were assembled into transcripts via a *de novo* assembly pipeline followed by functional annotation and classification. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were either allocated a functional annotation or categorized under Pfam/COG/KEGG categories. Identification of differentially expressed transcripts and their comparative sequence analysis revealed candidate genes related to silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. Several candidate anti-microbial peptides which may have potential anti-bacterial, anti-fungal or anti-parasitic activity in *A. assamensis* were also identified. T/A and AT/TA were predicted to be the most abundant mono- and di-nucleotide simple sequence repeat markers in the transcriptome. Transcriptome validation was carried out by quantitative real-time PCR (qPCR) amplification of eight transcripts. The resources generated by this study will expand the periphery of existing genomic data on *A. assamensis* facilitating future in-depth studies on its unknown aspects.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Silk is an important cultural and commercial fibre largely obtained from silkworms. Some silkworms have been domesticated by humans over a period of time to exploit their potential in textiles. Still, majority of them remain semi-domesticated or wild. The silk proteins- fibroin and sericin of domesticated mulberry silkworm, *Bombyx mori* (Family:

Bombycidae) has been extensively used for tissue engineering applications (Das et al., 2014). Research on these has been further accelerated by discovery of its complete genome and transcriptome (Li et al., 2012; The International Silkworm Genome Consortium, 2008). Similar as well as novel research applications can also be expected from biomaterials of other less-studied semi- or undomesticated silkworms. The dearth of information and hindrances in domestication of these silkworms currently restricts their usage in such applications. One such semi-domesticated silkworm is *Antheraea assamensis* (Helfer), also known as the “muga silkworm”. *A. assamensis* ($n = 15$) is a multivoltine, polyphagous silkworm classified under the order - Lepidoptera and family - Saturniidae. It is mostly endemic to the Brahmaputra valley of Assam and adjoining hilly areas of Northeast India (Tikader et al., 2013). Being the sole producer of globally acclaimed “muga silk”, a lustrous golden yellow fabric, makes *A. assamensis* one of the most important components of the Assamese silk industry and it hugely contributes towards employment generation in North-Eastern India. The unique quality of this silk-based textile earned it a geographical indication tag

Abbreviations: AaCbP, *Antheraea assamensis* Carotenoid binding protein; AaFhc, *Antheraea assamensis* Fibroin heavy chain; AC, Alimentary Canal; AMP, Anti-microbial peptides; CEG, Core eukaryotic genes; COG, Cluster of Orthologous genes; EST, Expressed Sequence Tag; GO, Gene Ontology; KEGG, Kyoto Encyclopaedia of Genes and Genomes; MISA, MicroSatellite Identification Tool; PNTAa, Putative novel transcripts of *Antheraea assamensis*; POU, Pit-Oct-Unc; qPCR, Quantitative real time PCR; RB, Residual body; SG, Silk gland; SRA, Short Read Archive; SSR, Simple sequence repeats.

* Corresponding author at: Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India.

E-mail addresses: ubora@iitg.ernet.in, drutpalbora@gmail.com (U. Bora).



EXPANDING THE FRONTIERS OF RICE RESEARCH THROUGH OMICS

Debajyoti Kabiraj¹, Jonjyoti Kalita¹, Hasnahana Chetia¹, Deepika Singh¹, *Utpal Bora^{1,2}

¹Bioengineering Research Laboratory, Dept. of Biosciences and Bioengineering, Indian Institute of Technology, Guwahati (IITG), Assam 781039, India

²Institutional Biotech Hub, Centre for the Environment, Indian Institute of Technology Guwahati, (IITG), Assam 781039, India

*Corresponding author : Tel: +91 361 2582215; Fax: +91 361 2582249;
Email: ubora@iitg.ernet.in

ABSTRACT

Rice is an important food grain and a staple food for nearly 60% of the world's population. In post-genomic era, comprehensive data from "omics" technologies have been extensively applied to rice to elucidate its cellular intricacies and regulatory mechanisms. The present review deals with current scenario of "omics" applications in understanding this important food crop and highlights the essentiality of genomics and transcriptomics in rice research. It includes a brief coverage of improvement in genetic mapping, rice trait analysis, developmental understanding of the rice embryogenesis etc. aided by high-throughput sequencing technologies. The review also discusses application of basic and applied proteomics and metabolomics research in providing new directions to improvement of qualitative and quantitative traits. These include another proteome and seed development, tackling of biotic and abiotic stress, pathogen infections etc. Towards the end, the review addresses the application of data science into this field, shortlisting relevant database (for the omics) discussed above.

Key words : Rice, *Oryza sativa*, Genomics, Transcriptomics, Proteomics, Metabolomics, Database

Nucleic Acid Based Biosensors for Clinical Applications

Utpal Bora^{1,2*}, Arghya Sett¹ and Deepika Singh¹

¹Bioengineering Research Laboratory, Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati, Assam-781039, India

²Mugagen Laboratories Pvt. Ltd., Technology Incubation Centre, Indian Institute of Technology Guwahati, Guwahati, Assam-781039, India

Abstract

Biosensors are analytical tools which have applications in diverse fields like diagnostics, disease monitoring, food industry and environmental monitoring etc. Since last few decades, nucleic acid based analytics have gained much interest in detection and monitoring of various clinical manifestations. Nucleic acid based biosensors (NABs) are preferred choice due to their high sensitivity and selectivity. Novel synthetic probes (PNA, Aptamer) have been exploited to fabricate point-of-care NABs. Due to flexible and low-cost fabrication, they are evolving as rapid and reliable tools in clinical diagnostics market. In this review, we focus on different transducer platform (Optical, Electrochemical, Piezoelectrical) based NABs which have been implemented to detect broad spectrum of metabolic and infectious diseases. A few commercialized NAB products are also summarized.

Keywords: Biosensors; Nucleic Acid Biosensors (NABs); Clinical applications

Introduction

The timely monitoring and regulation of various parameters is important to many applications in food industry, clinical diagnoses, hygiene, environmental protection, drug development, forensics etc. [1]. Therefore, there is a requirement to have reliable analytical devices, which can perform rapid and accurate analyses. Biosensors have been developed using the specificity of a biological molecule integrated to a physicochemical transducer to convert a perceived signal (usually very weak) into a comprehensible electrical signal [2]. The use of nucleic acid as a recognition layer in biosensor design is relatively a new and exciting area in analytics [3]. Nucleic acid based biosensors (NABs) have applications in

(i) clinical diagnostics of inherited diseases and pathogenic infections

(ii) laboratory analysis like probe development for different hybridization techniques and microarray systems.

NABs primarily use deoxyribonucleic acid (DNA), ribonucleic acid (RNA), peptide nucleic acid (PNA), and aptamers (both DNA and RNA) as oligonucleotide probes. The fundamental principle behind NABs depend on sequence complementarity as per Chargaff's rules of base pairing (for DNA, A=T, G ≡ C) except in the case of aptamers. Principle of aptamer based detection is more akin to antigen-antibody or receptor-ligand interactions.

Some of the DNA-based sensors that can detect the presence of pathogenic micro-organisms, other infectious agents like virus, genetic polymorphisms and detection of point mutations (SNP) have reached the market recently [4].

The major advantage of DNA based probes are the ability to amplify a desired target DNA from the host pathogen using PCR and consequently augment the signal generated by the biosensor. Unlike other sensing molecules like enzymes or antibodies, DNA forms easily synthesizable biological recognition layers which are highly stable and reusable after simple thermal melting of DNA duplex [5]. DNA as a sensing tool is preferred among other nucleic acid based probes (Figure 1).

Still much advancement is awaited in development of highly precise and sensitive NABs which can detect multi-analytes reducing the time and cost.

Design and Fabrication of NABs

Hybridization based biosensors rely on the duplex formation between nucleic acids. Hybridization normally takes place between a known DNA sequence i.e., probe and an unknown counterpart i.e. target DNA, but DNA-RNA and RNA-RNA hybridizations can also occur [6]. The basic schema of NAB is depicted in the Figure 2. DNA probes can be synthesized by available chemical methods or PCR. But for RNA, a probe may be obtained by reverse-transcription (RT) of a previously isolated or specific messenger RNA (mRNA), or predicting its nucleotide sequence based on the amino acid sequence of the protein expressed by that DNA. Despite validated proof of the later strategy is limited due to codon degeneracy. Traditional nucleic

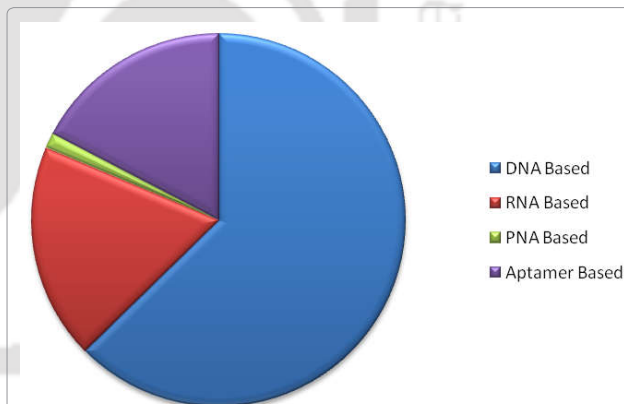


Figure 1: Publications related to different NABs for clinical applications (Last 5 year Data obtained from Scopus database).

***Corresponding author:** Utpal Bora, Bioengineering Research Laboratory, Department of Biotechnology, Indian Institute of Technology Guwahati, Guwahati, Assam-781039, India, Tel: 913612582215; Fax: +913612582249; E-mail: ubora@iitg.ernet.in, ubora@rediffmail.com

Received September 12, 2013; **Accepted** October 25, 2013; **Published** November 04, 2013

Citation: Bora U, Sett A, Singh D (2013) Nucleic Acid Based Biosensors for Clinical Applications. Biosens J 2: 104. doi: [10.4172/2090-4967.1000104](https://doi.org/10.4172/2090-4967.1000104)

Copyright: © 2013 Bora U, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.