INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Aparajita Dutta
Roll Number	:	156101002
Programme of Study	:	Ph.D.

Thesis Title: Neural Network Models for Analyzing the Splicing Cell Variable from Genome Sequences

Name of Thesis Supervisor(s)	:	Prof. Ashish Anand
Thesis Submitted to the Department/ Center	:	CSE
Date of completion of Thesis Viva-Voce Exam	:	14/09/2021
Key words for description of Thesis Work	:	Splice sites, BLSTM network, alternative splicing, neural network, MLP, doc2vec, word2vec

SHORT ABSTRACT

The escalating rate of deaths caused by complex human diseases has led to the need for unravelling the underlying genetic causation of these diseases. The study of the functional and structural information encoded in the genome facilitates genome annotation and helps in deciphering the relationship between the genome (genotype) and the disease traits (phenotype). The relationship between genotype and phenotype is critically complex, passing through several layers of complex biophysical processes. Variations in biophysical processes are manifested through the change in rate and quantity of production of several cell variables like splicing, transcription rate, polyadenylation, and DNA methylation. It is easier to associate the genotype with such more closely related measurable intermediate cell variables.

This thesis focuses on studying one such cell variable called splicing. Splicing occurs co-transcriptionally during RNA processing of genes. A gene comprises alternating regions called exons and introns. During splicing, the introns of a gene are removed, and the exons are ligated. Splicing is responsible for the transcript and protein diversity in eukaryotes. Several computational models are employed for gaining a deeper understanding of the splicing phenomenon. One way of studying the splicing mechanism is to identify splice sites from genome sequences by employing computational models.

However, the existing studies on the identification of splice sites have one or more of the following limitations:

1. The traditional computational models that identify splice sites are mainly based on functional genomic features. However, such feature sets are neither exhaustive nor optimal.

2. Several existing studies do not focus on extraction and interpretation of the biological features learnt by the model.

3. The existing studies primarily focus on identifying canonical splice sites: sites that contain the consensus GT and AG at donor and acceptor sites.

4. Most of the existing studies focus on studying splice sites from a single species.

This thesis works on the limitations mentioned above. We aim at identifying splice sites based on sequence-based features only. We employ various neural network models that learn the sequence-based features by themselves such that hand-crafted feature engineering can be eliminated to a great extent. This reduces the dependency on the existing knowledge bias. Neural network models have obtained state-of-the-art performances in identifying splice sites from the genome sequences de novo. Often such models take only nucleotide sequences as input and learn relevant features on their own. However, extracting the interpretable motifs from the model remains a challenge. We explore several existing visualization techniques in their ability to infer relevant features learnt by a neural network on the task of splice junction identification. We study a particular class of neural networks, called recurrent neural networks (RNN), in this thesis.

The existing prediction models primarily focus on identifying canonical splice sites. However, identification of non-canonical splice sites (splice sites lacking the GT - AG consensus) is also equally important for a comprehensive understanding of the splicing phenomenon. This thesis works towards this objective by studying non-canonical splice sites in greater detail to obtain features specific to the non-canonical splicing.

Furthermore, most of the existing studies focus on identifying and analysing splice sites for a single species. However, models capable of identifying splice sites from multiple species with comparable accuracy are preferable due to the robustness and generalizability. We analyse the performance of an RNN model in identifying splice sites from human, mouse, and drosophila melanogaster. We also test the model's performance on species that were not used during training. Furthermore, we extract the non-canonical splicing features learnt by the model from the three species and validate them with knowledge from the literature.