INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student : HIMADRI NAYAK

Roll Number : 09612315

Programme of Study : Ph.D.

Thesis title: On the Multiset of Factors of a String

Name of Thesis Supervisor(s) : Dr. Kalpesh Kapoor

Thesis Submitted to the Department/ Center : Department of Mathematics

Date of completion of Thesis viva-Voce Exam : 18-01-2017

Key words for description of Thesis Work : k-abelian equivance, factor multiset, l-modular equivalence, k-precedence data, partial coloring, discrepancy, separating word problem

SHORT ABSTRACT

A string is a finite or infinite sequence of letters drawn from an alphabet. A string v is said to be factor of another string s if there exists strings u and w such that s = uvw. of This thesis investigates properties of strings that have the same multiset of factors of certain length. Two strings u and v are called k-abelian equivalent if the frequency of every factor of length $\leqslant$ k are the same in u and v. Alternatively, the set of strings with identical multiset of factors of length up to k are said to k-abelian equivalent. The existing characterization of the permutations that are required to transform a string to a different string having the same multiset of factors of length k are used for showing that it is sufficient to check k and (k − 1)-length factor multiset of two strings to identify if they are k-abelian equivalent. A necessary and sufficient condition has been given for two strings to be k-abelian equivalent. The number of strings having identical multiset of k-length factors can be computed using the existing technique for finding the number of Eulerian cycles in a multidigraph. We use a variation of this theory to find a lower bound for the maximum number k-abelian equivalent strings for k = 2. Also the maximum number of strings of length n that are floor(n/2)-abelian equivalent is given. A generalization of k-abelian equivalence of words, k-precedence data equivalence, that takes into account the order of appearance of two same length factors is investigated. We study the related permutation that are required to transform one string to the other having the same k-precedence data. It is shown that the strings with length n can be transformed under some specific rules to a different string with the same k-precedence data if and only if k $\geqslant$ floor((n−1)/3)+ 1. We refine the definition of l-modular factor composition of words, which is a generalization of composition of a word given by the multiset of factors of certain length. A relation with k-abelian equivalence of words in this context is established. We introduce l-modular factor precedence data as a generalization of both l-modular composition and l-precedence data of a word. A lower bound of $\Omega((n \log n)^{(1/4)})$ on l is given to have different l-modular factor precedence data for every n-length string. We reduce the problem of finding non zero discrepancy with respect to partial coloring on a particular kind of l-regular hypergraph to the problem of finding different l-modular compositions, a restricted version of l-modular factor composition, in a pair of strings. Then the existing bounds on l to generate different l-modular composition for every n length string is used for obtaining the bounds on l to have non zero discrepancy in l-regular hypergraph with respect to a partial coloring.