Chapter 52
Novel PSSM-Based Approaches for Gene Identification Using Support Vector Machine

Heena Farooq Bhat
Department of Computer Science, University of Kashmir, India

M. Arif Wani
Department of Computer Science, University of Kashmir, India

ABSTRACT

By understanding the function of each protein encoded in genome, the molecular mechanism of the cell can be recognized. In genome annotation field, several methods or techniques have been developed to locate or predict the patterns of genes in genome sequence. However, recognizing corresponding gene of a given protein sequence using conventional tools is inherently complicated and error prone. This paper first focuses on the issue of gene prediction and its challenges. The authors then present a novel method for identifying genes that involves a two-step process. First the research presents new features extracted from protein sequences using a position specific scoring matrix (PSSM). The PSSM profiles are converted into uniform numeric representation. Then, a new structured approach has been applied on PSSM vector which uses a decision tree-based technique for obtaining rules. Finally, the rules of single class are joined together to form a matrix which is then given as an input to SVM for classification purpose. The rules derived from algorithm correspond to genes. The authors also introduce another approach for predicting genes based on PSSM using SVM. Both the methods have been implemented on genome DNAset dataset. Empirical evaluation shows that PSSM based SAFARI approach produces better results.

DOI: 10.4018/979-8-3693-3026-5.ch052

1. INTRODUCTION

As the genome sequence data grows at a very large pace, the various number of gene predicting programs have come into existence. Big data associated with the problem of identification of genes can be projected into sub-spaces or clusters so that the given problem can be divided into sub-problems. Each sub-problem can then be optimized with an appropriate model. One of the approaches of dividing a given problem into sub-problems involves projecting the big data to various sub-space grids (Wani, 2012; Wani & Yesilbudak, 2013), where each sub-space grid represents a sub-problem. Each sub-problem can then be represented independently with various models which can be combined by using a rule based system (Wani, 2001). The gene identification from large genome sequence is found to be one of the significant issues to solve in the field of bioinformatics. There is an essential requirement of developing gene finding methods and their corresponding functions. The primary issue in the process of gene forecasting is to locate the protein coding genes in genomic DNA sequence. In spite of large amount of amino acid sequences of proteins produced, only a small part of protein function has been interpreted. DNA binding proteins plays an essential role in all cell functions such as DNA replication, DNA repair, DNA modification and all the other activities allied with DNA. Most of the genes include statistics for generating proteins at definite level and these proteins then used to carry out a broad diversity of procedures in the unit. Other type of genes, known as non-coding genes, determines efficient genetic material (RNA) which is occupied in the guidelines of appearance of genes and production of proteins. This sequence of DNA is not allowed to transform into amino acids and hence be deficient in the distinctive sequence restriction of coding sequences.

The specific recognition of genes is one of the elementary steps in all meta-genomic sequencing projects (Goel et al, 2013). Gene forecasting also involves the use of Support Vector Machines. The Support Vector Machines (SVM) is a supervised learning algorithm used to categorize reserved blueprint of data (Bhat & Wani, 2014). SVM has been applied to various other domains which incorporate wind speed prediction (Wani & Bhat, 2017), fingerprint recognition (Khan & Wani, 2015), face recognition (Bhat & Wani, 2014), global solar radiation forecasting (Mujtaba & Wani, 2017) and also evaluates various information retrieval algorithms with the use of linear algebra (Bhat & Wani, 2017). The functional proteins included in organisms at the upper level are not adjacent. These are frequently divided into coding and non-coding regions. Such types of coding fragments are recognized as exons. The exonomic sequences are then mixed together by non-coding section of exceedingly changeable length known as introns. The extensively used move toward the genome annotation consists of two methods namely extrinsic and intrinsic methods. The extrinsic methods are used for homology detection (Bhat & Wani, 2017) and intrinsic methods are used for gene prediction (Mathe et al, 2002). The homology methods when executed can forecast only half portion of the genes and the rest of the genes remain unknown. Therefore, more extrapolative, fast and reliable methods are needed which can detect all the protein coding genes accurately. The method of assimilating nucleic acid similarity search has been shown practically in a long line of accomplishment, including GRAIL (Xu et al, 1996), HMMgene (Krogh, 2000), and Genscan (Burge & Karlin, 1997) and GenomeScan (Yeh et al, 2001). The most important challenge that follows the sequencing of either a small segment of DNA sequence or a long genome sequence is to establish the location of functional units like protein coding genes (exons), splice sites, terminators etc. This provides a procedure of identifying the regions that encode proteins. The protein