AN EMPIRICAL STUDY ON GENE PREDICTION TECHNIQUES

Manaswini Pradhan

Lecturer, P.G. Department of Information and Communication Technology,
Fakir Mohan University, Orissa,(India)

ABSTRACT

In recent times, bioinformatics plays an increasingly important role in the study of advanced biology. Bioinformatics deals with the management and analysis of biological information stored in databases. The field of genomics is dependent on bioinformatics which is a significant novel tool emerging in biology for finding facts about gene sequences, interaction of genomes, and unified working of genes in the formation of final syndrome or phenotype. The rising popularity of genome sequencing has resulted in the utilization of computational methods for gene finding in DNA sequences. Recently, computer assisted gene prediction has gained impetus and tremendous amount of work has been carried out on this area. An ample range of noteworthy techniques have been proposed by the researchers for the prediction of genes. An extensive review of the prevailing literature related to gene prediction is presented along with classification by utilizing an assortment of techniques. In addition, a succinct introduction about the prediction of genes is presented to get acquainted with the vital information on gene prediction.

Keywords: Genomic Signal Processing (GSP), Gene, Exon, Intron, Gene Prediction, DNA Sequence, RNA, Protein, Sensitivity, Specificity, Mrna.

I. INTRODUCTION

This paper gives a review of the literature in the context of the research undertaken for the wide range of research methodologies employed for the analysis and classification of gene prediction based upon different methods. The review of literature has been classified under the different mechanisms that has been adopted and experimented. The paper presents the research work already conducted in the area of gene prediction that supplements and supports the present research. It also gives a direction for the relevance of the present research. As far as practicable, the review of literature analyses the earlier research conducted in the context of the present research, and thus provides a scope for the present research giving a vivid description to present the conceptual terms and processes involved. Care has been taken to incorporate the available research papers that provide a scope for further research in this topic.

1.1 Reviews on Gene Prediction Methodologies

A wide range of research methodologies employed for the analysis and the prediction is presented in this section. Some of the innovative techniques have been adopted by various researchers from time to time. The reviewed gene predictions based on some mechanisms are classified and the success and limitations are the points of discussion detailed in the following subsections.

1.1.1 Support Vector Machine

Jiang Qianet *et al*. [1] presented an approach which depends upon the SVMs for predicting the targets of a transcription factor by recognizing subtle relationships between their expression profiles. Particularly, they used SVMs for predicting the regulatory targets for 36 transcription factors in the Saccharomyces cerevisiae genome which depends on the microarray expression data from lots of different physiological conditions.

MicroRNAs (miRNAs) which play an important role as post transcriptional regulators are small non-coding RNAs. For the 5' components, the purpose of animal miRNAs normally depends upon complementarities. Even though there is a lot of suggested numerous computational miRNA target-gene prediction techniques, it still have drawbacks in revealing actual target genes. MiTarget which is a SVM classifier for miRNA target gene prediction have been introduced by Sung-Kyu Kim *et al* [2].

A Bayesian framework depends upon the functional taxonomy constraints for merging the multiple classifiers have been introduced by Zafer Barutcuoglu *et al*. [3]. A hierarchy of SVMclassifiers has been trained on multiple data types. For attaining the most probable consistent set of predictions, they have merged predictions in the suggested Bayesian framework.

Hany Alashwal *et al.* [4] represented Bayesian kernel for the Support Vector Machine (SVM) in order to predict protein-protein interactions. By integrating the probability characteristic of the existing experimental protein-protein interactions data, the classifier performances which were compiled from different sources could be enhanced.

1.1.2 Gene ontology

A method for approximating the protein function from the Gene Ontology classification scheme for a subset of classes have been introduced by Jensen *et a.l*[5] This subset which incorporated numerous pharmaceutically appealing categories such as transcription factors, receptors, ion channels, stress and immune response proteins, hormones and growth factors can be calculated.

Hongwei Wu *et al.* [6] introduced a computational method for predicting the functional modules which are encoded in microbial genomes. They have also acquired a formal measure for measuring the degree of consistency among the predicted and the known modules and carried out statistical analysis of consistency measures.

Yingyao Zhou *et al* introduced an ontology-based pattern identification (OPI) is a data mining algorithm that methodically recognizes expression patterns that best symbolizes on hand information of gene function. Rather than depending on a widespread threshold of expression resemblance to describe functionally connected sets of genes, OPI obtained the optimal analysis background that produce gene expression patterns and gene listings that best predict gene function utilizing the criterion of GBA [7] have utilized OPI to a publicly obtainable gene expression data collection on the different stages of life of the malarial parasite

Remarkable advancement in sequencing technology and sophisticated experimental assays that interrogate the cell, along with the public availability of the resulting data, indicate the era of systems biology. The development of techniques that can automatically make use of these datasets to make quantified and robust predictions of gene function that are experimentally verified require comprehensive and wide variety of available data. The VIRtual Gene Ontology (VIRGO) was introduced by Naveed Massjouni *et al.* [8].

Important approach into the cellular function and machinery of a proteome has been provided using a map of protein–protein interactions. With a relative specificity semantic relation, the similarity between two Gene Ontology (GO) terms is measured. Here, a method for restructuring a yeast protein–protein interaction map that exclusively depends upon the GO observations has been presented by Xiaomei Wu *et al.* [9].

The functions of each protein are performed inside some specialized locations in a cell. For recognizing the protein function and approving its purification, this subcellular location is important. For predicting the location which depends upon the sequence analysis and database information from the homologs, there are numerous computational techniques. Few latest methods utilze text obtained from biological abstracts. The main goal of Alona Fyshe *et al.* [10] is to enhance the prediction accuracy of such text-based techniques. For improving text-based prediction, they recognized three techniques such as (1) a rule for ambiguous abstract removal, (2) a mechanism for using synonyms from the Gene Ontology (GO) and (3) a mechanism for using the GO hierarchy to generalize terms. They proved that these three methods can enhance the accuracy of protein sub-cellular location predictors considerably which utilized the texts that are removed from PubMed abstracts whose references were preserved in Swiss-Prot.

1.1.3 Homology

Jong-won Chang *et al.* [11] introduced a scheme for improving the accuracy of gene prediction that has merged the ab-initio method based on homology. Taking the advantage of the known information, the latter recognizes each gene for previously recognized genes whereas, the former rely on predefined gene features. In spite of the crucial negative aspect of the homology-based method, the proposed scheme has also adopted parallel processing for assuring the optimal system performance i.e. the bottleneck happened predictably due to the large amount of unprocessed ordered information.

Automatic gene prediction is one of the predominant confrontations in computational sequence analysis. Conventional methods to gene detection depend on statistical models derived from already known genes. Contrary to this, a set of comparative methods depend on likening genomic sequences from evolutionary associated organisms to one another. These methods were founded on the hypothesis of phylogenetic foot printing: they capitalize on the feature that functionally significant areas in genomic sequences are generally more conserved than non-functional areas. Leila Taher *et al*. [12] have constructed a web-based computer program for gene prediction on the basis of homology at BiBiServ (Bielefeld Bioinformatics Server).

Perfect accuracy is yet to be attained in computational gene prediction techniques, even for comparatively simple prokaryotic genomes. Problems in gene prediction revolve around the fact that several protein families continue to be uncharacterized. Consequently, it appears that only about half of an organism's genes can be assuredly ascertained on the basis of similarity with other known genes. Mohammed Zahir Hossain Sarker *et al.* [13] have attempted to discern the intricacies of certain gene prediction algorithms in Genomics.

1.1.4 Hidden Markov Model (HMM)

Vladimir Pavlovic *et al.* [14] have presented a well-organized framework in order to learn the combination of gene prediction systems. Their approach can model the statistical dependencies of the experts which is the main advantage. The application of a family of combiners has been represented by them in the increasing order of statistical complexity starting from a simple Naive Bayes to Input HMMs. A system has been introduced by them for combining the predictions of individual experts in a frame-consistent manner.

The computational method which was introduced for the problem of finding the genes in eukaryotic DNA sequences is not yet solved acceptably. Gene finding programs have accomplished comparatively high accuracy on short genomic sequences but do not execute well if there is a presence of long sequences of indefinite number of genes. Here, programs which exist tend to calculate many false exons. For the ab initio prediction of protein coding genes in eukaryotic genomes a program named AUGUSTUS has been introduced by Mario Stanke *et al.* [15]. Based on the Hidden Markov Model, the program was constructed and it incorporated a number of well-known methods and submodels.

The presence of processed pseudogenes: nonfunctional, intronless copies of real genes found elsewhere in the genome damaged the correct gene prediction. The processed pseudogenes are usually mistaken for real genes or exons by gene prediction programs which lead to biologically irrelevant gene predictions. Despite the fact that the methods exists for identifying the processed pseudogenes in genomes, there has not been made any attempt for incorporating pseudogene removal with gene prediction or even for providing a freestanding tool which identifies such incorrect gene predictions. PPFINDER (for Processed Pseudogene finder), a program that has been incorporated with numerous methods of processed pseudogene for finding the mammalian gene annotations have been introduced by Marijke J, Van Baren *et al.* [16].

DeCaprio *et al.* [17] demonstrated the first proportional gene predictor, Conrad which depends upon semi-Markov conditional random fields (SMCRFs). In contradictory to the best standalone gene predictors that depends upon generalized hidden Markov models (GHMMs) and accustomed by maximum probability Conrad was favourably trained for maximizing annotation accuracy.

The majority of computational tools which exists depend on sequence homology and/or structural similarity for discovering microRNA (miRNA) genes. Of late, with regards to sequence, structure and comparative genomics information, the supervised algorithms were applied for addressing this problem. Almost in these studies, experimental evidence rarely supported miRNA gene predictions. In addition to, prediction accuracy remains uncertain. In order to predict the miRNA precursors, a computational tool (SSCprofiler) which utilized a probabilistic method based on Profile Hidden Markov Models was introduced by Anastasis Oulas *et al.* [18].

1.1.5 Different Software programs for gene prediction

A computational technique to create gene models by utilizing evidence produced from a varied set of sources, inclusive of those representatives of a genome annotation pipeline has been detailed by Jonathan E. Allen *et al*. [19]. The program, known as Combiner, took into account genomic sequence as input and the positions of gene predictions from ab initio gene locators, protein sequence arrangements, expressed sequence tag and cDNA arrangements, splice site predictions, and other proofs.

Biju Issac *et al*. [20] have detailed that EGPred is an internet-based server that united ab initio techniques and similarity searches to predict genes, specifically exon areas, with high precision. The EGPred program consists of the following steps: (1) a preliminary BLASTX search of genomic sequence across the RefSeq database has been utilized to find protein hits with an $E - value < 1$; (2) a second BLASTX search of genomic sequence across the hits from the preceding run with relaxed parameters (E-values <10) assists to get back all possible coding exon regions; (3) a BLASTN search of genomic sequence across the intron database was then utilized to identify possible intron regions; (4) the possible intron and exon regions were likened to filter/remove incorrect exons; (5) the NNSPLICE program was then utilized to relocate splicing signal site locations in the outstanding possible coding exons; and (6) ultimately ab initio predictions were united with exons obtained from the fifth step on the basis of the relative strength of start/stop and splice signal regions as got from ab initio and similarity search.

Yanhong Zhou *et al.* [21] introduced a gene prediction program named GeneKey. GeneKey can attain the high prediction accuracy for genes with moderate and high C+G contents when the widely used dataset which are collected by Reese and Kulp are trained [109]. On the other hand, the prediction accuracy was lesser for CG-poor genes. They constructed a LCG316 dataset which composes of gene sequences with low C+G contents to solve this problem.

Mario Stanke *et al.* [22] have presented an internet server for the computer program AUGUSTUS, which is utilized to predict genes in eukaryotic genomic sequences. AUGUSTUS is founded on a comprehensive hidden Markov model representation of the probabilistic model of a sequence and its gene structure. The web server has permitted the user to enforce constraints on the calculated gene structure.

Overall of 143 prokaryotic genomes were achieved with an efficient version of the prokaryotic genefinder EasyGene. By Comparing the GenBank and RefSeq annotations with the EasyGene predictions, they unveiled that in some genomes up to 60% of the genes might be represented with an incorrect initial codon particularly in the GC-rich genomes. The fractional differentiation between annotated and predicted affirmed that numerous short genes are annotated in numerous organisms. Additionally, there is a chance that genes might be left behind during the annotation of some of the genomes. Out of 143, 41 genomes to be over-annotated by .5% which means that too many ORFs were represented as genes have been calculated by Pernille Nielsen *et al.* [23].

Antonio Starcevic *et al.* [24] has accomplished the program package ‘ClustScan’ (Cluster Scanner) for rapid, semi-automatic, annotation of DNA sequences encoding modular biosynthetic enzymes that consists of polyketide synthases (PKS), non-ribosomal peptide syntheses (NRPS) and hybrid (PKS / NRPS) enzymes. In addition of displaying the predicted chemical structures of products the program also allows the export of the structures in a standard format for analyses with other programs.

Kai Wang *et al.* [25] have built up a committed, publicly obtainable, splice site prediction program known as NetAspGene, for the genus Aspergillus. Gene sequences from Aspergillus fumigatus, the most general mould pathogen, were utilized to construct and experiment their model. Compared to several animals and plants, Aspergillus possesses finer introns; consequently they have utilized a bigger window dimension on single local networks for instruction, to encompass both donor and acceptor site data.

The ease of use of a huge part of the maize B73 genome sequence and originating sequencing technologies recommend economical and simple ways to sequence areas of interest from many other maize genotypes. Gene content prediction is one of the steps required to convert these sequences into valuable data. Gene predictor specifically trained for maize sequences is so far not available in public. The EuGene software merged numerous sources of data into a condensed gene model prediction and this EuGene is preferred for training by Pierre Montalent *et al* [26]. The results were compacted together into a library file and e-mailed to the user.

1.1.6 Other Training methodologies

Huiqing Liu *et al.* [27] introduced a computational method for patient outcome prediction. In the training phase of this method, they utilized two types of extreme patient samples: (1) short-term survivors who got an inconvenient result in a small period and (2) long-term survivors who were preserving a positive outcome after a long follow-up time.

According to the parent of origin, Imprinted genes are epigenetically modified genes whose expression can be determined. They are concerned in embryonic development and imprinting dysregulation is linked to diabetes, obesity, cancer and behavioral disorders such as autism and bipolar disease. A statistical model which depends on DNA sequence characteristics have been trained by Philippe P., Luedi *et al*. [28]. It not only identified potentially imprinted genes but also predicted the parental allele from which they were expressed.

1.1.7 Other Machine Learning Techniques

Stephanie Seneff *et al.* [29] described an approach incorporating constraints from orthologous human genes in order to predict the exon-intron structures of mouse genes using the techniques which are utilized in speech and natural language processing applications in the past. A context-free grammar is used in their approach for parsing a training corpus of annotated human genes. For capturing the common features of a mammalian gene, a statistical training process has generated a weighted Recursive Transition Network (RTN).

An approach to the problem of splice site prediction, by applying stochastic grammar inference was presented by Kashiwabara *et al*. [30]. Four grammar inference algorithms to infer 1465 grammars were used, and a 10-fold cross-validation to choose the best grammar for every algorithm was also used. The matching grammars were entrenched into a classifier and the splice site prediction was made to run and the results were compared with those of NNSPLICE, the predictor used by Genie gene finder.

Katharina J Hoff *et al.* [31] introduced a gene prediction algorithm for metagenomic fragments based on a two-stage machine learning approach. In the first step, for extracting the features from DNA sequences, they have used linear discriminants for monocodon usage, dicodon usage and translation initiation sites. In the second step, for computing the probability in such a way that the open reading frame encodes a protein, an artificial neural network combined these features with open reading frame length and fragment GC-content.

Single nucleotide polymorphisms (SNPs) give much assurance as a source for disease-gene association. However, the cost of genotyping the tremendous number of SNPst restricted the research. Therefore, for identifying a small subset of informative SNPs, the supposed tag SNPs is of much importance. This subset comprises of chosen SNPs of the genotypes, and represents the rest of the SNPs accurately. Additionally, in order to estimate prediction accuracy of a set of tag SNPs, an efficient estimation method is required. A genetic algorithm (GA to tag SNP problems, and the K-nearest neighbor (K-NN) which act as a prediction method of tag SNP selection have been applied by Li-Yeh Chuang *et al.* [32].

1.1.8 Digital Signal Processing

The protein-coding areas of DNA sequences have been noticed to display the period-three behaviour, which can be capitalized on to predict the position of coding areas inside genes. Earlier, discrete Fourier transform (DFT) and digital filter-based techniques have been utilized for the detection of coding areas. But, these techniques do not considerably subdue the noncoding areas in the DNA spectrum at $2\pi/3$. As a result, a non-coding area may unintentionally be recognized as a coding area. Trevor W. Fox *et al*. [33] have set up a method (a quadratic window operation subsequent to a single digital filter operation) that has restrained almost each of the non-coding areas.

The basic problem to interpret genes is to predict the coding regions in large DNA sequences. For solving that problem, Digital Signal Processing techniques have been used successfully. Furthermore, the existing tools are not able to calculate all the coding regions which are present in a DNA sequence. A predictor introduced by Anibal Rodriguez Fuentes *et al.* [34] based on the linear combination of two other methods proved good quality efficacy separately. And also for reducing the computational load, a fast algorithm was developed Fuentes *et al*. [35] earlier. Some thoughts have been reviewed concerning the combination of the predictor with other methods.

Several digital signal processing, methods have been utilized to mechanically differentiate protein coding areas (exons) from non-coding areas (introns) in DNA sequences. Mai S. Mabrouk *et al.* [36] have differentiated these sequences in relation to their nonlinear dynamical characteristics.

Genomic sequence, structure and function analysis of various organisms has been a testing problem in bioinformatics. In this context protein coding region (exon) identification in the DNA sequence has been accomplishing immense attention over a few decades. By exploiting the period-3 property present in it these coding regions can be recognized. The discrete Fourier transform has been normally used as a spectral estimation technique to extract the period-3 patterns available in DNA sequence. The conventional DFT approach loses its efficiency in case of small DNA sequences for which the autoregressive (AR) modeling is used as an optional tool. An optional but promising adaptive AR method for the similar function has been proposed by Sitanshu Sekhar Sahu *et al.* [37].

1.1.9 Neural Network

Alistair M. Chalket *et al.* [38] have presented a neural network based computational model that uses a broad range of input parameters for AO (Antisense Oligonucleotides) prediction. From AO scanning experiments in the literature sequence and efficacy data were gathered and a database of 490 AO molecules was generated. A neural network model was trained utilizing a set of parameters derived on the basis of AO sequence properties.

Takatsugu Kan *et al.* [39] have aimed to detect the candidate genes involved in lymph node metastasis of esophageal cancers, and investigate the possibility of using these gene subsets in artificial neural networks (ANNs) analysis for estimating and predicting occurrence of lymph node metastasis. With 60 clones their ANN model was capable of most accurately predicting lymph node metastasis.

In bioinformatics identification of short DNA sequence motifs which act as binding targets for transcription factors is an important and challenging task. Though unsupervised learning techniques are often applied from the literature of statistical theory, for the discovery of motif in large genomic datasets an effective solution is not yet found. For motif-finding problem, Shaun Mahony *et al.* [40] have offered three self-organizing neural networks.

Neural networks are long time popular approaches for intelligent machines development and knowledge discovery. Nevertheless, problems such as fixed architecture and excessive training time still exist in neural networks. This problem can be solved by utilizing the neuro-genetic approach. Neuro-genetic approach is based on a theory of neuroscience which states that the genome structure of the human brain considerably affects the evolution of its structure. Therefore the structure and performance of a neural network is decided by a gene created. Assisted by the new theory of neuroscience, Zainal A. Hasibuan *et al.* [41] have proposed a biologically more reasonable neural network model to overcome the existing neural network problems by utilizing a simple Gene Regulatory Network (GRN) in a neuro-genetic approach.

Liu Qicai *et al.* [42] have employed Artificial Neural Networks (ANN) for analyzing the fundamental data obtained from 78 pancreatitis patients and 60 normal controls consisting of three structural of HBsAg, ligand of HBsAg and clinical immunological characterizations, laboratory data and genetypes of cationic trypsinogen gene PRSS1. They have verified the outcome of ANN prediction using T-cell culture with HBV and flow cytometry.

1.1.10 Other techniques

Gautam Aggarwal *et al.* [43] analyzed the interpretation of three complete genomes by means of the ab initio methods of gene identification GeneScan and GLIMMER. The interpretation made by means of GeneMark is endowed in GenBank which is the standard against which these are compared.

Freudenberg *et al*. [44] introduced a technique for predicting disease related human genes from the phenotypic emergence of a query disease. Corresponding to their phenotypic similarity diseases of known genetic origin are to be clustered.

Thomas Schiex *et al*. [45] have detailed the FrameD, a program that predicts the coding areas in prokaryotic and matured eukaryotic sequences.

Rice xa5 gene produces recessive, race-specific impediment to bacterial blight disease attributable to the pathogen Xanthomonas oryzae pv. Oryzae and has immense importance for research and propagation. In an attempt to clone xa5, an F2 population of 4892 individuals was produced by Zhong Yiming *et al.* [46], from the xa5 close to isogenic lines, IR24 and IRBB5.

Bayesian variable choosing for prediction utilizing a multinomial probit regression model with data amplification to change the multinomial problem into a series of smoothing problems has been dealt with by Xiaobo Zhou *et al.* [47].

A reaction pattern library which consists of bond-formation patterns of GT reactions have been introduced by Shin Kawano *et al*. [48] and the co-occurrence frequencies of all reaction patterns in the glycan database is researched.

A comparative-based method to the gene prediction issue has been offered by Said S. Adi *et al*. [49]. It was founded on a systemic arrangement of more than two genomic sequences. In other words, on an arrangement that took into account the truth that these sequences contain several conserved regions, the exons, interconnected by unrelated ones, the introns and intergenic regions.

Linkage analysis is a successful process for combining the diseases with particular genomic regions. These regions are usually big, incorporating hundreds of genes that make the experimental methods engaged to recognize the disease gene arduous and cost. In order to prioritize candidates for more experimental study, Richard A. George *et al*. [50] have introduced two techniques: Common Pathway Scanning (CPS) and Common Module Profiling (CMP).

For deciphering the digital information that is stored in the human genome, the most important goal is to identify and characterize the complete ensemble of genes. Successful hybrid methods combining these two concepts have also been developed. A third orthogonal approach for gene prediction which depends on the detection of the genomic signatures of transcription have been introduced by Gustavo Glusman *et al.* [51] and are accumulated over evolutionary time.

Differing from most organisms, the c-proteobacterium Acidithiobacillus ferrooxidans withstand an abundant supply of soluble iron and they live in dreadfully acidic conditions (pH 2). It is also odd that it oxidizes iron as an energy source. Therefore, it faces the demanding twin problems of managing intracellular iron homeostasis when accumulated with enormously elevated environmental masses of iron and modifying the utilization of iron both as an energy source and as a metabolic micronutrient. The first model for a Fur-binding site consensus sequence in an acidophilic iron-oxidizing microorganism was given by Raquel Quatrini *et al.* [52] and he laid the foundation for forthcoming studies aimed at expanding their understanding of the regulatory networks that control iron uptake, homeostasis and oxidation in extreme acidophiles.

A generic DNA microarray design which suits to any species would significantly benefit comparative genomics. The viability of such a design by ranking the great feature densities and comparatively balanced nature of genomic tiling microarrays was proposed by Thomas Royce *et*E *et al.* [53].

For analyzing the functional gene links, the phylogenetic approaches have been compared by Daniel Barker *et al.* [54]. From species' genomes, the independent instances of the correlated gain and loss of pairs of genes have been encountered by using these approaches. They interpreted the effect from the significant results of correlations on two phylogenetic approaches such as Dollo parsminony and maximum likelihood (ML).

The complex and restrained problem in eukaryotes is accurate gene prediction. A constructive feature of predictable distributions of spliceosomal intron lengths were presented by Scott William Roy *et al.* [55].

Poonam Singhal *et al.* [56] have introduced an ab initio model for gene prediction in prokaryotic genomes on the basis of physicochemical features of codons computed from molecular dynamics (MD) simulations.

Manpreet Singh *et al.* [57] have detailed that the drug invention process has been commenced with protein identification since proteins were accountable for several functions needed for continuance of life. Protein recognition further requires the identification of protein function. The proposed technique has composed a categorizer for human protein function prediction.

The efficiency of their suggested approach in type 1 diabetes (T1D) was examined by Shouguo Gao *et al.* [58]. While organizing the T1D base, 266 recognized disease genes and 983 positional candidate genes were obtained from the 18 authorized linkage loci of T1D. A de novo prediction algorithm for ncRNA genes with factors resulting from sequences and structures of recognized ncRNA genes in association to allure was illustrated by Thao T. Tran *et al.* [59]. Bestowing these factors, genome-wide prediction of ncRNAs was performed in Escherichia coli and Sulfolobus solfataricus by administering a trained neural network-based classifier.

A comparative-based method to the gene prediction issue has been offered by Said S. Adi *et al.* [60]. It was founded on a syntenic arrangement of more than two genomic sequences.

MicroRNAs (miRNAs) that control gene expression by inducing RNA cleavage or translational inhibition are small non-coding RNAs. Most human miRNAs are intragenic and they are interpreted as a part of their hosting transcription units. The gene expression profiles of miRNA host genes and their targets which are correlated inversely have been assumed by Vincenzo Alessandro Gennarino *et al.* [61]. They have developed a procedure named HOCTAR (host gene oppositely correlated targets), which ranks the predicted miRNA target genes depending upon their anti-correlated expression behavior comparing to their respective miRNA host genes.

1.2 Supplementary Gene Classification and Prediction Techniques

Some of the recent related research works are reviewed here.

Zhenqiu Liu *et al.* [62] have offered an analytical method for categorizing the gene expression data. In the proposed method, dimension reduction has been achieved by utilizing the kernel principal component analysis (KPCA) and categorization has been achieved by utilizing the logistic regression (discrimination). KPCA is a generic nonlinear form of principal component analysis. Five varied gene expression datasets related to human tumor samples has been categorized by utilizing the proposed algorithm. The high potential of the proposed algorithm in categorizing gene expression data has been confirmed by comparing with other well-known classification methods like support vector machines and neural networks. Roberto Ruiz *et al.* [63] have proposed a novel heuristic method for selecting appropriate gene subsets which can be utilized in the classification task. Statistical significance of the inclusion of a gene to the final subset from an ordered list is the criteria on which their method is based.

Yanxiong Peng *et al*. [64] have performed a comparative analysis on different biomarker discovery methods that includes six filter methods and three wrapper methods. After this, they have presented a hybrid approach known as FR-Wrapper for biomarker discovery.

Minca Mramor *et al*. [65] have proposed a method for the analysis of gene expression data that gives an unfailing classification model and gives useful insight of the data in the form of informative perception.

Hau-San Wong *et al.* [66] have proposed regulation-level method for symbolizing the microarray data of cancer classification that can be optimized utilizing genetic algorithms (GAs). The proposed symbolization decreases the dimensionality of microarray data to a greater extent compared with the traditional expression-level features.

Ahmad M. Sarhan [67] has developed an ANN and the Discrete Cosine Transform (DCT) based stomach cancer detection system. Classification features are extracted by the proposed system from stomach microarrays utilizing DCT. ANN does the Classification (tumor or no-tumor) upon application of the features extracted from the DCT coefficients. In his study he has used the microarray images that were obtained from the Stanford Medical Database (SMD). The ability of the proposed system to produce very high success rate has been confirmed by simulation results. Georgios Papachristoudis *et al.* [68] have offered SoFoCles, an interactive tool that has made semantic feature filtering a possibility in microarray classification problems by the utilization of external, unambiguous knowledge acquired from the Gene Ontology.

Rameswar Debnath *et al.* [69] have proposed an evolutionary method that is capable of selecting a subset of potentially informative genes that can be used in support vector machine (SVM) classifiers. The proposed evolutionary method estimates the fitness function utilizing SVM and a specified subset of gene features, and new subsets of features were chosen founded on the frequency of occurrence of the features in the evolutionary approach and amount of generalization error in SVMs.

MiTarget which is a SVM classifier for miRNA target gene prediction was introduced by Sung-Kyu Kim *et al.* [70]. It employed a radial basis function kernel and was then categorized by structural, thermodynamic, and position-based features as a similarity measure for SVM features. For the first time, the features were presented and the mechanism of miRNA binding was reproduced.

Based on the information that a majority of exon sequences have a 3-base periodicity, and intron sequences do not have the sole characteristic, a technique to predict protein coding regions was developed by Changchuan Yin *et al.* [71].

On the basis of a two-stage machine learning approach a gene prediction algorithm for metagenomic fragments was proposed by Katharina Hoff *et al.* [72]. Initially, for extracting the features from DNA sequences, linear discriminants were employed for monocodon usage, dicodon usage and translation initiation sites. Secondly, for calculating the chance in such a way that the open reading frame encodes a protein and an artificial neural network combines these characteristics with open reading frame length and fragment GC-content.

Based on the physicochemical features of codons computed from molecular dynamics (MD) simulations an ab initio model for gene prediction in prokaryotic genomes was introduced by Poonam Singhal *et al.* [73].

For the genus Aspergillus a program called NetAspGene which is a dedicated, publicly available, splice site prediction was developed by Kai Wang *et al.* [74]. The most widespread mould pathogen that is the gene sequences from Aspergillus fumigatus, were employed to build and test their model. Aspergillus encloses smaller introns when compared with several animals and plants; and hence to cover both the donor and acceptor site information they have applied a larger window size on single local networks for training.

Bayesian kernel was represented for the Support Vector Machine (SVM) by Hany Alashwal *et al.* [75] so as to predict protein-protein interactions. By putting together the probability characteristic of the existing experimental protein-protein interactions data, the classifier performances that were amassed from diverse sources could be improved.

II. DIRECTIONS FOR THE FUTURE RESEARCH

In this review paper, various techniques utilized for the gene prediction has been analyzed thoroughly. Also, the performance claimed by the technique has also been analyzed. From the analysis, it can be understood that the prediction of genes using the hybrid techniques shown the better accuracy. Due to this reason, the hybridization of more techniques will attain the acute accuracy in prediction of genes. This paper will be a healthier foundation for the budding researchers in the gene prediction to be acquainted with the techniques available in it. In future lot of innovative brainwave will be rise using our review work.

III. CONCLUSION

The instant reviews is a humble attempt to explain, compare and assess the performance of different soft computing that are being applied to cancer prediction and prognosis. Specifically a number of trends have been identified with respect to the types of computational intelligent learning methods being used, the types of training data being integrated, the kinds of endpoint predictions being made, the types of cancers being studied and the overall performance of these methods in predicting cancer susceptibility or outcomes. While ANNs still predominate, it is evident that a growing variety of alternate machine learning strategies are being used and it is being applied to many types of cancers to predict at least three different kinds of outcomes. It is also clear that soft computing methods generally improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems. While most studies are generally well constructed and reasonably well validated, certainly greater attention to experimental design and implementation appears to be warranted, especially with respect to the quantity and quality of biological data. Improvements in experimental design along with improved biological validation would no doubt enhance the overall quality, generality and reproducibility of many computational intelligence-based classifiers.

Gene prediction is a very rising research in the field of bio-informatics that has received growing attention in the research community over the past decade. This paper is a comprehensive survey of the significant researches and techniques existing for gene prediction. An introduction to gene prediction has also been presented and the existing works are classified according to the techniques implemented. These researches are basically about the numerous techniques available for gene prediction analysis. From the analysis, it can be understood that the prediction of genes using the *hybrid techniques* shows a better accuracy. Due to this reason, the hybridization techniques will attain the acute accuracy in prediction of genes.

It may be concluded that most of the recent works have performed the classification study using gene expression data. The selected gene expression dataset has been optimized and classified using traditional classifier. Although the optimization is effective, the ultimate objective has not yet been achieved due to the inadequacy in classification effectiveness. Therefore, the enhancement of classifier becomes an essential pre-requisite for effective classification of microarray gene expression data.

REFERENCES

[1]. Jiang Qian, Jimmy Lin, Nicholas M. Luscombe, Haiyuan Yu and Mark Gerstein, "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data", Bioinformatics, Vol.19, No.15, pp.1917-1926, 2003

[2]. Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee and Byoung-Tak Zhang, "miTarget: microRNA target gene prediction using a support vector machine", BMC Bioinformatics, Vol.7, No.411, pp.1-14, 2006

[3]. Zafer Barutcuoglu, Robert E. Schapire and Olga G. Troyanskaya,"Hierarchical multi-label prediction of gene functions", Bioinformatics, Vol.22, No.7, pp.830-836, 2006

[4]. Hany Alashwal, Safaai Deris and Razib M. Othman, "A Bayesian Kernel for the Prediction of Protein-Protein Interactions", International Journal of Computational Intelligence, Vol. 5, No.2, pp.119-124, 2009

[5]. Jensen, Gupta, Stærfeldt and Brunak, "Prediction of human protein function according to Gene Ontology categories", Bioinformatics, Vol.19, No.5, pp.635-642, 2003

[6]. Hongwei Wu, Zhengchang Su, Fenglou Mao, Victor Olman and Ying Xu, "Prediction of functional modules based on comparative genome analysis and Gene Ontology application", Nucleic Acids Research, Vol.33, No.9, pp.2822-2837, 2005

[7]. Yingyao Zhou, Jason A. Young, Andrey Santrosyan, Kaisheng Chen, S. Frank Yan and Elizabeth A. Winzeler, "In silico gene function prediction using ontology-based pattern identification", Bioinformatics, Vol.21, No.7, pp.1237-1245, 2005

[8]. Naveed Massjouni, Corban G. Rivera and Murali, “VIRGO: computational prediction of gene functions", Nucleic Acids Research, Vol. 34, No.2, pp. 340-344, 2006

[9]. Xiaomei Wu, Lei Zhu, Jie Guo, Da-Yong Zhang and Kui Lin, "Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations", Nucleic Acids Research, Vol.34, No.7, pp.2137-2150, April 2006

[10]. Alona Fyshe, Yifeng Liu, Duane Szafron, Russ Greiner and Paul Lu, "Improving subcellular localization prediction using text classification and the gene ontology", Bioinformatics, Vol.24, No.21, pp.2512-2517, 2008

[11]. Jong-won Chang, Chungoo Park, Dong Soo Jung, Mi-hwa Kim, Jae-woo Kim, Seung-sik Yoo and Hong Gil Nam, "Space-Gene: Microbial Gene Prediction System Based on Linux Clustering", Genome Informatics, Vol.14, pp.571-572, 2003.

[12]. Leila Taher, Oliver Rinner, Saurabh Garg, Alexander Sczyrba and Burkhard Morgenstern, "AGenDA: gene prediction by cross-species sequence comparison", Nucleic Acids Research, Vol. 32, pp.305–308, 2004

[13]. Mohammed Zahir Hossain Sarker, Jubair Al Ansary and Mid Shajjad Hossain Khan, "A new approach to spliced Gene Prediction Algorithm", Asian Journal of Information Technology, Vol.5, No.5, pp.512-517, 2006

[14]. Vladimir Pavlovic, Ashutosh Garg and Simon Kasif, "A Bayesian framework for combining gene predictions", Bioinformatics, Vol.18, No.1, pp.19-27, 2002

[15]. Mario Stanke and Stephan Waack, "Gene prediction with a hidden Markov model and a new intron submodel ", Bioinformatics Vol. 19, No. 2, pp.215-225, 2003

[16]. Marijke J. van Baren and Michael R. Brent, "Iterative gene prediction and pseudogene removal improves genome annotation", Genome Research, Vol.16, pp.678-685, 2006

[17]. David DeCaprio, Jade P. Vinson, Matthew D. Pearson, Philip Montgomery, Matthew Doherty and James E. Galagan, "Conrad: Gene prediction using conditional random fields", Genome Research, Vol.17, No.9, pp.1389-1398, August 2007

[18]. Anastasis Oulas, Alexandra Boutla, Katerina Gkirtzou, Martin Reczko, Kriton Kalantidis and Panayiota Poirazi, "Prediction of novel microRNA genes in cancer-associated genomic regions-a combined computational and experimental approach", Nucleic Acids Research, Vol.37, No.10, pp.3276-3287, 2009

[19]. Jonathan E. Allen, Mihaela Pertea and Steven L. Salzberg, "Computational Gene Prediction Using Multiple Sources of Evidence", Genome Research, Vol.14, pp.142-148, 2004

[20]. Biju Issac and Gajendra Pal Singh Raghava, "EGPred: Prediction of Eukaryotic Genes Using Ab Initio Methods after combining with sequence similarity approaches", Genome Research, Vol.14, pp.1756-1766, 2004

[21]. Yanhong Zhou, Huili Zhang, Lei Yang and Honghui Wan, "Improving the Prediction Accuracy of Gene structures in Eukaryotic DNA with Low C+G Contents", International Journal of Information Technology , Vol.11, No.8, pp.17-25,2005

[22]. Mario Stanke and Stephan Waack, "Gene prediction with a hidden Markov model and a new intron submodel ", Bioinformatics Vol. 19, No. 2, pp.215-225, 2003

[23]. Pernille Nielsen and Anders Krogh, "Large-scale prokaryotic gene prediction and comparison to genome annotation ", Bioinformatics, Vol.21, No.24, pp.4322-4329, 2005

[24]. Antonio Starcevic, Jurica Zucko, Jurica Simunkovic, Paul F. Long, John Cullum and Daslav Hranueli, "ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures", Nucleic Acids Research, Vol.36, No.21, pp.6882-6892, October 2008

[25]. Kai Wang, David Wayne Ussery and Søren Brunak, "Analysis and prediction of gene splice sites in four Aspergillus genomes", Fungal Genetics and Biology, Vol. 46, pp.14-18, 2009

[26]. Pierre Montalent and Johann Joets, "EuGene-maize: a web site for maize gene prediction", Bioinformatics, Vol.26, No.9, pp.1254-1255, 2010

[27]. Huiqing Liu, Jinyan Li and Limsoon Wong, "Use of extreme patient samples for outcome prediction from gene expression data", Bioinformatics, Vol.21, No.16, pp.3377-3384, 2005

[28]. Philippe P. Luedi, Alexander J. Hartemink and Randy L. Jirtle, “Genome-wide prediction of imprinted murine genes”, Genome Research, Vol.15, pp. 875-884, 2005

[29]. Stephanie Seneff, Chao Wang and Christopher B.Burge, "Gene structure prediction using an orthologous gene of known exon-intron structure", Applied Bioinformatics, Vol.3, No.2-3, pp.81-90, 2004

[30]. Kashiwabara, Vieira, Machado-Lima and Durham, "Splice site prediction using stochastic regular grammars", Genet. Mol. Res, Vol. 6, No.1, pp.105-115, 2007

[31]. Katharina J Hoff, Maike Tech, Thomas Lingner, Rolf Daniel, Burkhard Morgenstern and Peter Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach", BMC Bioinformatics, Vol. 9, No.217, pp.1-14, April 2008.

[32]. Li-Yeh Chuang, Yu-Jen Hou and Cheng-Hong Yang, "A Novel Prediction Method for Tag SNP Selection using Genetic Algorithm based on KNN", World Academy of Science, Engineering and Technology, Vol.53, No.213, pp.1325-1330, 2009

[33]. Trevor W. Fox and Alex Carreira, "A Digital Signal Processing Method for Gene Prediction with Improved Noise Suppression", EURASIP Journal on Applied Signal Processing, Vol.1, pp.108-114, 2004

[34]. Anibal Rodriguez Fuentes, Juan V. Lorenzo Ginori and Ricardo Grau Abalo, “A New Predictor of Coding Regions in Genomic Sequences using a Combination of Different Approaches”, International Journal of Biological and Life Sciences, Vol. 3, No.2, pp.106-110, 2007

[35]. Fuentes, Ginori and Abalo, "Detection of Coding Regions in Large DNA Sequences Using the Short Time Fourier Transform with Reduced Computational Load," LNCS, vol.4225, pp. 902-909, 2006.

[36]. Mai S. Mabrouk, Nahed H. Solouma, Abou-Bakr M. Youssef and Yasser M. Kadah, "Eukaryotic Gene Prediction by an Investigation of Nonlinear Dynamical Modeling Techniques on EIIP Coded Sequences", International Journal of Biological and Life Sciences, Vol. 3, No.4, pp. 225-230, 2007

[37]. Sitanshu Sekhar Sahu and Ganapati Panda, "A DSP Approach for Protein Coding Region Identification in DNA Sequence", International Journal of Signal and Image Processing, Vol.1, No.2, pp.75-79, 2010

[38]. Alistair M. Chalk and Erik L.L. Sonnhammer, "Computational antisense oligo prediction with a neural network model", Bioinformatics, Vol.18, No.12, pp.1567-1575, 2002

[39]. Takatsugu Kan, Yutaka Shimada, Funiaki Sato, Tetsuo Ito, Kan Kondo, Go Watanabe, Masato Maeda,eiji Yamasaki, Stephen J.Meltzer and Masayuki Imamura, "Prediction of Lymph Node Metastasis with Use of Artificial Neural Networks Based on Gene Expression Profiles in Esophageal Squamous Cell Carcinoma", Annals of surgical oncology, Vol.11, No.12, pp.1070-1078,2004

[40]. Shaun Mahony, Panayiotis V. Benos, Terry J.Smith and Aaron Golden, Self-organizing neural networks to support the discovery of DNA-binding motifs", Neural Networks, Vol.19, pp.950-962, 2006

[41]. Zainal A. Hasibuan, Romi Fadhilah Rahmat, Muhammad Fermi Pasha and Rahmat Budiarto, "Adaptive Nested Neural Network based on human Gene Regulatory Network for gene knowledge discovery engine", International Journal of Computer Science and Network Security, Vol.9, No.6, ppp.43-54, June 2009

[42]. Liu Qicai, Zeng Kai,Zhuang Zehao, Fu Lengxi, Ou Qishui and Luo Xiu, "The Use of Artificial Neural Networks in Analysis Cationic Trypsinogen Gene and Hepatitis B Surface Antigen", American Journal of Immunology, Vol.5, No.2, pp.50-55, 2009

[43]. Gautam Aggarwal and Ramakrishna Ramaswamy, "Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER", J.Biosci, Vol.27, No.1, pp.7-14, February 2002

[44]. Freudenberg and Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes", Bioinformatics, Vol. 18, No.2, pp.110-115, April 2002

[45]. Thomas Schiex, Jerome Gouzy, Annick Moisan and Yannick de Oliveira, "FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences", Nucleic Acids Research, Vol.31, No.13, pp.3738-3741, 2003

[46]. ZHONG Yiming, JIANG Guanghuai, CHEN Xuewei, XIA Zhihui, LI Xiaobing, ZHU Lihuang and ZHAI Wenxue, "Identification and gene prediction of a 24 kb region containing xa5, a recessive bacterial blight resistance gene in rice (Oryza sativa L.)", Chinese Science Bulletin, Vol. 48, No. 24, pp.2725-2729,2003

[47]. Xiaobo Zhou, Xiaodong Wang and Edward R.Dougherty, "Gene Prediction Using Multinomial Probit Regression with Bayesian Gene Selection", EURASIP Journal on Applied Signal Processing, Vol.1, pp.115-124, 2004

[48]. Shin Kawano, Kosuke Hashimoto, Takashi Miyama, Susumu Goto and Minoru Kanehisa, "Prediction of glycan structures from gene expression data based on glycosyltransferase reactions", Bioinformatics, Vol.21, No.21, pp.3976-3982, 2005

[49]. Said S. Adi and Carlos E. Ferreira, "Gene prediction by multiple syntenic alignment", Journal of Integrative Bioinformatics, Vol.2, No.1, 2005

[50]. Richard A. George, Jason Y. Liu, Lina L. Feng, Robert J. Bryson-Richardson, Diane Fatkin and Merridee A. Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction", Nucleic Acids Research, Vol.34, No.19, pp.1-10, 2006

[51]. Gustavo Glusman, Shizhen Qin, Raafat El-Gewely, Andrew F. Siegel, Jared C. Roach, Leroy Hood and Arian F. A. Smit, "A Third Approach to Gene Prediction Suggests Thousands of Additional Human Transcribed Regions", PLOS Computational Biology, Vol.2, No.3, pp.160-173, March 2006

[52]. Raquel Quatrini, Claudia Lefimil, Felipe A. Veloso, Inti Pedroso, David S. Holmes and Eugenia Jedlicki, "Bioinformatic prediction and experimental verification of Fur-regulated genes in the extreme acidophile Acidithiobacillus ferrooxidans", Nucleic Acids Research, Vol. 35, No. 7, pp. 2153–2166, 2007

[53]. Thomas E. Royce, Joel S. Rozowsky and Mark B. Gerstein, "Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification", Nucleic Acids Research, Vol.35, No.15, 2007

[54]. Daniel Barker, Andrew Meade and Mark Pagel, "Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes", Bioinformatics, Vol.23, No.1, pp.14-20, 2007

[55]. Scott William Roy and David Penny, "Intron length distributions and gene prediction", Nucleic Acids Research, Vol.35, No.14, pp.4737-4742, 2007

[56]. Poonam Singhal, Jayaram, Surjit B. Dixit and David L. Beveridge, "Prokaryotic Gene Finding Based on Physicochemical Characteristics of Codons Calculated from Molecular Dynamics Simulations", Biophysical Journal, Vol.94, pp.4173-4183, June 2008

[57]. Manpreet Singh, Parminder Kaur Wadhwa, and Surinder Kaur, "Predicting Protein Function using Decision Tree", World Academy of Science, Engineering and Technology, Vol39, No. 66, pp.350-353, 2008

[58]. Shouguo Gao and Xujing Wang, "Predicting Type 1 Diabetes Candidate Genes using Human Protein-Protein Interaction Networks", J Comput Sci Syst Biol, Vol. 2, pp.133-146, 2009

[59]. Thao T. Tran, Fengfeng Zhou, Sarah Marshburn, Mark Stead3, Sidney R. Kushner and Ying Xu, "De novo computational prediction of non-coding RNA genes in prokaryotic genomes", Bioinformatics, Vol.25, No.22, pp.2897-2905, 2009

[60]. Said S. Adi and Carlos E. Ferreira, "Gene prediction by multiple syntenic alignment", Journal of Integrative Bioinformatics, Vol.2, No.1, 2005

[61]. Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio and Sandro Banfi, "MicroRNA target prediction by expression analysis of host genes", Genome Research, Vol.19, No.3, pp.481-490, March 2009

[62]. Zhenqiu Liu, Dechang Chen and Halima Bensmail, "Gene Expression Data Classification with Kernel Principal Component Analysis", J Biomed Biotechnol, Vol.2005, No.2, pp. 155–159, 2005.

[63]. Roberto Ruiz, Jose C. Riquelme and Jesus S. Aguilar-Ruiz,. "Incremental wrapper-based gene selection from microarray data for cancer classification," Pattern Recognition, Vol. 39 , No. 12, pp. 2383-2392, 2006

[64]. Yanxiong Peng, Wenyuan Li and Ying Liu,"A Hybrid Approach for Biomarker Discovery from Microarray Gene Expression Data for Cancer Classification", Cancer Inform., Vol.2, pp.301-311, 2007

[65]. Minca Mramor, Gregor Leban, Janez Demar and Bla Zupan,. "Visualization-based cancer microarray data classification analysis", Bioinformatics, Vol. 23, No.16, pp.2147-2154, 2007

[66]. Hau-San Wong and Hong-Qiang Wang,."Constructing the gene regulation-level representation of microarray data for cancer classification", Journal of Biomedical Informatics, Vol.41, No.1, pp.95-105, 2008

[67]. Ahmad m. Sarhan, "Cancer classification based on microarraygene expression data using DCT and ANN", Journal of Theoretical and Applied Information Technology, Vol.6, No.2, pp.207-216, 2009

[68]. Georgios Papachristoudis, Sotiris Diplaris and Pericles A. Mitkas,."SoFoCles: Feature filtering for microarray classification based on Gene Ontology", Journal of Biomedical Informatics, Vol.43, No.1, pp.1-14, 2010

[69]. Rameswar Debnath and Takio Kurita,' "An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories", BioSystems, Vol.100, No.1, pp.39-46, 2010

[70]. Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee and Byoung-Tak Zhang, "miTarget: microRNA target gene prediction using a support vector machine", BMC Bioinformatics, Vol.7, No.411, pp.1-14, 2006

[71]. Changchuan Yin and Stephen S.T. Yau, "Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence", Journal of Theoretical Biology, Vol.247, pp.687-694, 2007

[72]. Katharina J Hoff, Maike Tech, Thomas Lingner, Rolf Daniel, Burkhard Morgenstern and Peter Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach", BMC Bioinformatics, Vol. 9, No.217, pp.1-14, April 2008.

[73]. Poonam Singhal, Jayaram, Surjit B. Dixit and David L. Beveridge, "Prokaryotic Gene Finding Based on Physicochemical Characteristics of Codons Calculated from Molecular Dynamics Simulations", Biophysical Journal, Vol.94, pp.4173-4183, June 2008

[74]. Kai Wang, David Wayne Ussery and Søren Brunak, "Analysis and prediction of gene splice sites in four Aspergillus genomes", Fungal Genetics and Biology, Vol. 46, pp.14-18, 2009

[75]. Hany Alashwal, Safaai Deris and Razib M. Othman, "A Bayesian Kernel for the Prediction of Protein-Protein Interactions", International Journal of Computational Intelligence, Vol. 5, No.2, pp.119-124, 2009