International Journal of Advanced Technology and Engineering Exploration, Vol 5(46)
ISSN (Print): 2394-5443 ISSN (Online): 2394-7454
http://dx.doi.org/10.19101/IJATEE.2018.546018

Research Article

PSSM amino-acid composition based rules for gene identification

Heena Farooq Bhat[1*] and M. Arif Wani [2]
Research Scholar, Department of Computer Science, University of Kashmir, J&K, India[1]
Professor, Department of Computer Science, University of Kashmir, J&K, India[2]

Abstract
One of the major aspects in recognizing the molecular mechanism of the cell is to understand the significance or function of each protein encoded in the genome. For that purpose, genome annotation proves to be very supportive. One of the most obligatory phases of genome annotation is the prediction of the genes. Several methods or techniques have been developed in order to locate or predict the patterns of genes in genome sequence. However, still, the recognition of genes is found to be very complicated problem. Recognizing the corresponding gene of a given protein sequence by means of conventional tools is error prone. Hence, the recognition of genes is a very demanding task. In this paper, we first concentrate on the problem of gene prediction and its challenges. We then present a new method for identifying genes. This new method follows a two-step procedure. First, we present new features extracted from protein sequences and these features are derived from 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. The rules derived from an algorithm correspond to genes. This new method has been demonstrated on genome DNAset dataset. It is observed that the experimental results of new approach produces better results.

Keywords
Gene prediction, Classification, Feature extraction, Binding proteins, Rule induction, PSSM.

1.Introduction
With the advancement of genome sequence data, a large 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 [1−4], 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 [5].

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. One of the foremost problems in the process of gene forecasting is to discover 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. Deoxyribonucleic acid (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 ribonucleic acid (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 precise gene recognition is one of the fundamental steps in all meta-genomic sequencing projects [6]. Gene recognition also involves the use of support vector machines (SVMs). The SVM is a supervised learning algorithm used to categorize reserved blueprint of data [7−9]. SVM has been applied to various other domains which incorporate wind speed prediction [10], fingerprint recognition

*Author for correspondence

[11], face recognition [12], global solar radiation forecasting [13] and also evaluates various information retrieval algorithms with the use of linear algebra [14]. The functional proteins included in organisms at the upper level are not adjacent. These are frequently divided into coding and non-coding regions. These coding regions or segments are known as exons.

The exonomic sequences are then mixed together by non-coding section of exceedingly changeable length known as introns. The extensively used, move towards the genome annotation which consists of two methods namely extrinsic and intrinsic methods. The extrinsic methods are used for homology detection [15] and intrinsic methods are used for gene prediction [16]. The implementation of homology methods can predict only half 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 [17], HMMgene [18], and GenScan [19],[20] and GenomeScan [21]. 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 homology detection also takes an account of recognizing the patterns in multidimensional data [22].

Such a region is said to be an open reading frame (ORF) [23] which assembles like a gene but it has not been proved to be a gene yet. The objectives carried in this paper are as follows: We will first briefly review gene finding problem and then present the various approaches of gene finding algorithms. We will also provide the concept of PSSM where profiles of every protein sequence is converted into unvarying numeric representation. Paper also explains the new approach based on SAFARI algorithm for finding the induction rules which corresponds to genes. Finally, we will evaluate these approaches empirically and draw conclusions which will guide us for future research in this area.

2.Literature review

This section first describes the problem of finding genes.

2.1. Gene finding problem

Every living organism consists of cells. Each one of the cell holds a comprehensive replica of the RNA and DNA material. The genetic material is the draft that articulates all the proceedings in the human being and it is composed of the chemical substance known as DNA. A single-trapped DNA molecule is built of an elongated chain of miniature subunits called nucleotides (or bases). Each base is constructed of a sugar molecule, a phosphoric acid molecule, and one of four nitrogen bases: adenine (A), thymine (T), guanine (G), or cytosine (C), which concludes to the four-letter DNA code {A, T, G, C}. There are two major categories of cells, namely eukaryotic and prokaryotic cells based on the type of organism. The main discrepancy between eukaryotic and prokaryotic cells is that a eukaryotic cell involves a core part known as nucleus whereas a prokaryotic cell is without a nucleus. In eukaryotes, nucleus contains most of the genetic material. The same genetic material which resides on the chromosomes is structured in subunits. And these subunits are well-known as the genes of the living being. The genes are components of sequences of DNA which are being spread out all over the chromosomes. The genome of the living being is defined as a whole set of DNA/RNA, which involves both the coding and the non-coding sections of DNA.

The main complexity of recognizing genes is to recognize the purposeful and well-designed fractions of DNA sequence. Now the main motive that lies at the back of the crisis of recognizing genes, is to classify every component of DNA sequence correctly, as belonging to protein-coding section, RNA coding area, and non-coding or intergenic regions. The area of DNA sequence between the genes is known as intergenic region. The problem of predicting genes can be mathematically declared as follows.

Input*:* A sequence of DNA

$X = (x_1, \ldots, x_n) \in \Sigma^*$, *where* $\Sigma = \{A, T, C, G\}$

Output: The elements in X are properly categorized as it belongs to protein-coding part, RNA coding part, non-coding part, or intergenic part.

For both types of cells i.e. eukaryotes and prokaryotes, quite number of methods for recognizing genes have been developed. In prokaryotic types, genes are said to be found if it is build-up of stretched coding fragments, which are known as ORFs. On the other hand, genes in

eukaryotes also involve coding fragments but are broken up by long non-coding fragments. These coding fragments are said to be exons and non-coding fragments as introns. Only 3% of DNA sequence is coding part in case of human eukaryotes. Recognizing genes in prokaryotes is comparatively simple because of privileged DNA/RNA concentration and the non-existence of introns. One of the most important complexities in distinguishing genes for prokaryote is the occurrence of overlapping sections [24]. Due to large size of genome, the process appears to be more complex for eukaryotes and exons, which are shorter in length, are surrounded by lengthy introns. Additionally, less than 10% of coding region is involved in eukaryotic genomes while as in prokaryotic genomes; there is about 90% of coding sequence. Without a doubt, it is anticipated that more than 95% of human genes gives us an idea about at least one remarkable splice site [25].

For each program of gene estimation, two important characteristics need to be considered. The first characteristic represents the category of data source occupied by the program. The supplementary characteristic is the procedure that is used to coalesce the data used by program into a balanced and reliable forecast. In order to estimate the absolute composition of genes, three sorts of content are taken into general view: 1) Functional locations in the sequence, 2) content information, and 3) resemblance to known genes. The functional locations may include splice sites, start and stop codons, and a variety of transcription required sites. Such types of sites are normally called as signals. The technique that draws on signal is identified as ab-initio procedures of coding part prediction [26]. Even though, there is a significant increase in accuracy level of protein-coding gene prediction programs but still issues are inherent that need to get better and several concerns are as under:

(i) Recognition of exons which are shorter in length,
(ii) Detection of comprehensive gene formation,
(iii) Estimation of reduced and overlapping genes,
(iv) Subject sequences are not absolutely accurate.

The next part of this section describes the various approaches of gene prediction.

2.2. Gene Prediction Approaches

For the prediction of genes, a large number of approaches have been developed but concentrating on different types of issues. Several programs also exist that are most commonly used for this job. The gene finding approaches are broadly classified into three categories as: a) finding open reading frames (ORF), b) homology based approach which involves pair-wise alignment and c) ab-initio approach.

Here, we will illustrate some of the gene prediction approaches that are used most frequently for recognizing genes.

2.2.1. Gene prediction approach to find ORF

In case of prokaryotes, gene recognition is relatively dissimilar problem as compared to gene recognition in eukaryotic genomes. The prokaryotic genome involves two properties of having higher gene density and the absence of introns in their protein coding genes. That portion of coding genes are called as ORFs which may be longer that likely correspond to genes. The main problem which arises in this approach are frame-shift errors, i.e. a single insertion or deletion might lead to completely dissimilar amino-acid; very small genes may get missed; overlapping of long ORFs on opposite strands of DNA may lead to ambiguity.

2.2.2. Homology based approach

The rationale behind comparative or similarity based gene prediction methods is that the regions in the genome sequence coding for proteins are generally more conserved during evolution than non-functional regions. In actual fact, similarity based approaches are categorized into two classes for gene identification: the comparison of the DNA query sequence with a protein or cDNA sequence, or a database of such sequences, and the comparison of two or more genomic sequences. In both approaches; query and target sequences may be from the same or different species.

In these methods, genes can be predicted by aligning input sequences to the nearby homologous sequences in the database.

2.2.3. Ab-initio gene prediction methods

These methods instruct various factors on recognized annotations so that unidentified annotations can be predicted. Several gene prediction programs are briefly described as:

2.2.3.1.GenScan

GenScan [20], a gene recognizing program, employs a multifaceted probabilistic representation of the gene composition. In this program, the organization of the genomic data is modeled by an unambiguous state duration Hidden Markov Model. In addition, this program is used to detect the non-existence of genes or the occurrence of a single gene or multiple genes, which can be either comprehensive or fractional. It also predicts sub-optimal exons. Hence, GenScan

developed into the preferred method, for analyzing at the initial stage of long DNA sequences of eukaryotic genomes due to its high speed accuracy.

2.2.3.2.GenomeScan

GenomeScan [21] is based on GenScan. GenomeScan is a software program used to identify the coding and non-coding regions of genes in genomic DNA chain of characters, from a large diversity of living beings. The organism includes human and other vertebrates. This gene prediction program is based on two source of information: 1) representation of coding and non-coding regions and splice signal composition; and 2) DNA/RNA sequence resemblance data such as BLASTX hits. The program has been given an input which includes a DNA/RNA sequence formerly masked with RepeatMasker. The RepeatMasker is a file of parameters for suitable living beings and a 'Genoa file' which is full of review of accessible sequence comparison content. The positions of all detected coding gene sequences are published to an output file collectively with the comparable detected coding DNA sequence, nucleic acid sequences and an outline of the matching data used in the estimations.

2.2.3.3.GeneID

GeneID [19] was introduced as the first program to recognize full coding gene configuration of vertebrate genes in unstipulated DNA sequences. GeneID was intended by means of a tree formation: First, signals (splice sites and begin and end codons) defined by genes were estimated by the side of the query DNA sequence.

Then from these sites, assembling of potential coding regions occurred and as a final point, the scoring gene detection with the best possible measure was bringing together from the exons. In the novel GeneID, there was a heuristic scoring function which needs to be optimized. The sequence positions were forecasted and calculated using position weight matrices (PWMs). A huge amount of coding figures were computed on the estimated exons, and each exon keeps count as a purpose of the scores of the exon essential sites and of the coding figures.

2.2.3.4.GeneMark

GeneMark [27] was found to be the earliest means for finding prokaryotic genes. This tool has been used to engage a non-standardized Markov model to organize DNA sections into protein coding and non-coding but corresponding to coding. This process has been modified very recently to recognize gene composition in eukaryotic living beings. The best possible genes chosen by hidden Markov model and dynamic programming are practiced by a ribosomal position detection procedure. An up-to-date and modifying program is required repetitively, which then subsists in various parts for prediction of genes in prokaryotic, eukaryotic, and viral DNA sequences.

2.2.3.5.FGenesh

Fgenesh [28] proved to be the best ever and fast program i.e. 50-100 times faster than GenScan and for the most part, it is highly precise gene prediction program. This is HMM-based gene structure prediction (multiple genes, both chains). FGENESH capitulate the most precise and GeneMark is considered to be the second most exact prediction program. FGENESH recognized 11% more precise gene representation than GeneMark when tested on a set of 1353 test genes. This gene prediction program is considered to be the most accurate program for plant gene identification.

2.2.3.6.AUGUSTUS

Augustus [29] is a tool for recognizing genes in eukaryotic genomic DNA strand of characters. This program is based on a generalized hidden Markov model (GHMM). GHMM is a probabilistic representation of a sequence and its genetic material composition. Corresponding to the states in the model are the introns, exons, intergenic regions, etc. The Augustus network server permits to upload a DNA genetic material in FASTA format or multiple sequences in multiple FASTA format.

3.Position specific scoring matrix

PSSM was first introduced for detecting distantly related proteins. It was generated from a group of sequences previously aligned by structural or sequence similarity. Position-specific iterated BLAST (PSI-BLAST) [30,31] is the most commonly used application, which compares PSSM profiles for detecting remotely related homologous proteins or DNA. The original PSSM introduced by Gribskov et al. [32] consists of the following components:

- Position, which indicates the sequentially increased index of each amino acid residues in a sequence after multiple sequence alignment.
- Probe, which is a group of typical sequences of functionally related proteins that have been aligned by sequence or structural similarity.
- Profile, which is a matrix consisting of 20 columns corresponding to 20 amino acids; and consensus, which is a sequence of amino acid residues are the most similar to all the alignment residues of probes at each position. It is generated by selecting the highest score in the profile at each position.

4.PSSM based approach for gene finding

In this work, we will explore the idea of identifying genes by the following procedure as:

4.1.PSSM matrix representation of protein sequences

A position specific scoring matrix (PSSM) is a matrix containing probability of occurrence of each type of amino acid at each residue position of protein sequence. The data or record in PSSM is presented by a matrix of dimension L×20 (L rows and 20 columns) for a protein of length L, where 20 columns represents occurrence/substitution of each type of 20 amino acids.

The PSSM profile for each sequence is created using the following steps:

Step 1: The number of times an amino-acid '*i*' appeared in column '*j*' is calculated as

$$n_{ij} = n(i)\ at\ position\ j$$

Where '*i*' represent amino-acids ranging from 1 to 20 and '*j*' represents the column ranging from 1 to maximum length of a protein sequence.

Step 2: The frequency of amino acid '*i*' at position '*j*' is calculated as

$$F[i, j] = \frac{n_{ij}}{k}$$

Where n_{ij} is the number of instances of amino-acid '*i*' at position '*j*' and '*k*' is the total number of sequences.

Step 3: The PSSM is obtained by taking the logarithm of the values obtained above divided by the background frequency of the residues. To simplify, every amino acid appears equally in protein sequences, i.e. $f_i = 1/20 = 0.05$ for every amino acid '*i*'. To calculate the PSSM scores, we use log ratio as:

$$Score(i, j) = \log[f(i, j) / f(i)]$$

Where f*(i, j)* is the frequency of character '*i*' observed at position '*j*' and *f(i)* is the overall frequency of character '*i*'.

Step 4: Normalize the values of PSSM in range of 0 to1 by using the formula:

$$Nor\_Val = (Score(i, j) - MinVal) / (MaxVal - MinVal)$$

Step 5: The PSSM vector for each protein sequence is derived from the values obtained in step 4. The value of each amino-acid in a particular protein sequence is derived as:

$$PSSM\ Vector = Nor\_Val(i)\ at\ position\ j$$

Where '*i*' belong to amino-acid in a particular protein sequence.

Step 6: Add the values of same amino acids of a PSSM vector which reduces the same vector to 20 amino acids only.

$$\text{PSSM Vector New} = \sum_{i=1}^{L} (Symbols\ (i) == Symbols\ (i+1))$$

Where '*i*' belong to amino-acid in a particular protein sequence and ranges from 1 to L, where L is the length of particular protein sequence.

Step 7: Sort the value of each amino acid according to the original pattern of amino acid sequence (Symbols: ARNDCQEGHILKMFPSTWYVBZ). Generate PSSM vectors for all other sequences.

PSSM Vector Final = Sort (PSSM Vector New, 'Symbols')

The final PSSM profile is a matrix with dimension of M×20, which can depicted as follows:

$$PSSM = \begin{bmatrix} S_{1,1}, S_{1,2}, \ldots\ldots\ldots, S_{1,20} \\ S_{2,1}, S_{2,2}, \ldots\ldots\ldots, S_{2,20} \\ \vdots \quad \vdots \qquad\qquad \vdots \\ S_{M,1}, S_{M,2}, \ldots\ldots\ldots, S_{M,20} \end{bmatrix}$$

Where 'M' represents the number of sequences.

Now, each protein sequence, in each group is obtained as a vector which is already derived from step 5. The final dimension of each group will be (M ×20) where M is the number of protein sequences and 20 represents 20 amino-acids. All the groups are then merged and each group is represented as a class.

4.2.Structured approach for rule derivation

This approach uses a SAFARI algorithm [5] for obtaining rules which correspond to genes. This algorithm has an advantage over other decision based trees as it allows more than one attribute at the root node by which its performance is greatly enhanced. We implemented this approach for each PSSM of different protein sequences obtained from previous step. The instance values of each attribute are given symbolic names which are used in extraction of rules. This algorithm calculates the local probability and global probability i.e. a measure of classifying an example containing a specific symbol at each node using the subset of number of examples and a measure of classifying an example containing a specific symbol at each node using the entire set of examples respectively. Based on local and global probabilities, the function value 'F' is calculated for each symbol. The best symbol is calculated as the one that maximizes the function 'F' [5].

The algorithm below gives a procedure for our new approach of finding genes:

- The binding and non-binding protein chain sequences are processed to retrieve the homologous gene sequences.
- The gene sequences with the same gene id's obtained in step 1 are grouped together and each group contains N number of protein sequences.
- Each gene group is then converted into uniform numeric representations using PSSM.
- Each group consists of five classes.
- The final PSSM profile obtained is a matrix with dimension of M*20 where M is the total number of gene sequences and 20 are the amino-acids.
- Assign a symbol name 'S' to each value of an attribute represented by S_{ij} where '*i*' and '*j*' belongs to a number of values in each attribute.
- Calculate the local probability for each symbol.
- Calculate the global probability for each symbol.
- Calculate the function value 'F' for each symbol.
- Obtain the rules for each symbol which then correspond to genes.
- Delete the symbol for which rules are obtained from previous step.
- Repeat until 'F' becomes equal to 1.

5.Dataset

We have used gene datasets in this work for evaluating the gene prediction programs. The dataset that is used for this work is DNAset which involves definite genomic sequences. The DNAset [33] consists of 146 DNA-binding proteins and 250 non-binding proteins to develop various models for predicting DNA-binding proteins and for evaluating SVM models. 2435 DNA-Binding proteins from protein data bank (PDB) were extracted. Proteins with high similarity to other proteins were filtered. Finally, we got 146 non-redundant DNA-binding proteins and a non-redundant set of 250 non-binding proteins.

6.Results and discussions

The numerical execution of new approach for gene DNAset dataset is estimated and the rules are derived for each class. Each class contains a number of genes. The problem involves 20 attributes (20 amino acids): A(Alanine), R(Arginine), N(Asparagine), D (Aspartic Acid), C(Cysteine), E(Glutamic Acid), Q (Glutamine), G(Glycine), H(Histidine), I (Isoleucine), L(Leucine), K(Lysine), M (Methionine), F(Phenylalanine), P(Proline), S(Serine), T (Threonine), W(Tryptophan), V(Valine) and Y (Tyrosine). Each attribute has five values ranging from 0 to 1. There are five classes that consists of number of genes. The example set for the problem is given in *Table 1*.

We have obtained 30 groups based on same gene number, where each group consists of 'N' number of gene sequences. The PSSM is obtained for each group. The rules have been derived from each group for all the five classes as shown in *Table 2*. The text 'aa' in *Table 2* represents any of the 20 amino acid within range values. Each rule corresponds to a distinct gene.

Table 1 Normalized PSSM where each amino-acid is divided into five range intervals

Range	A		Range	R		Range	N		Range	D		Range	C	
	Min	Max		Min	Max		Min	Max		Min	Max		Min	Max
1	0.00	0.01	1	0.00	0.01	1	0.001	0.002	1	0.000	0.001	1	0.00	0.001
2	0.01	0.03	2	0.01	0.02	2	0.002	0.017	2	0.001	0.019	2	0.001	0.018
3	0.03	0.06	3	0.02	0.04	3	0.017	0.037	3	0.019	0.039	3	0.018	0.038
4	0.06	0.09	4	0.04	0.06	4	0.037	0.057	4	0.039	0.059	4	0.038	0.058
5	0.09	0.12	5	0.06	0.08	5	0.057	0.077	5	0.059	0.079	5	0.058	0.078

Range	E		Range	Q		Range	G		Range	H		Range	I	
	Min	Max		Min	Max		Min	Max		Min	Max		Min	Max
1	0 .001	0.002	1	0.00	0.011	1	0.00	0.01	1	0.00	0.002	1	0.00	0.01
2	0.002	0.019	2	0.011	0.022	2	0.01	0.02	2	0.002	0.012	2	0.01	0.02
3	0.019	0.041	3	0.022	0.044	3	0.02	0.04	3	0.012	0.027	3	0.02	0.04
4	0.041	0.063	4	0.044	0.066	4	0.04	0.06	4	0.027	0.042	4	0.04	0.06
5	0.063	0.085	5	0.066	0.088	5	0.06	0.08	5	0.042	0.057	5	0.06	0.08

Range	L		Range	K		Range	M		Range	F		Range	P	
	Min	Max		Min	Max		Min	Max		Min	Max		Min	Max
1	0.00	0.001	1	0.00	0.001	1	0.00	0.001	1	0.00	0.001	1	0.00	0.001
2	0.001	0.024	2	0.001	0.025	2	0.001	0.012	2	0.001	0.014	2	0.001	0.018
3	0.024	0.049	3	0.025	0.05	3	0.012	0.025	3	0.014	0.029	3	0.018	0.038
4	0.049	0.074	4	0.05	0.075	4	0.025	0.038	4	0.029	0.044	4	0.038	0.058
5	0.074	0.099	5	0.075	0.1	5	0.038	0.051	5	0.044	0.059	5	0.058	0.078

Range	S		Range	T		Range	W		Range	V		Range	Y	
	Min	Max		Min	Max		Min	Max		Min	Max		Min	Max
1	0.00	0.01	1	0.000	0.001	1	0.000	0.001	1	0.000	0.001	1	0.00	0.01
2	0.01	0.03	2	0.001	0.019	2	0.001	0.008	2	0.001	0.015	2	0.01	0.02
3	0.03	0.06	3	0.019	0.039	3	0.008	0.018	3	0.015	0.03	3	0.02	0.04
4	0.06	0.09	4	0.039	0.059	4	0.018	0.028	4	0.03	0.045	4	0.04	0.06
5	0.09	0.12	5	0.059	0.079	5	0.028	0.038	5	0.045	0.06	5	0.06	0.08

Table 2: Rule sets obtained by PSSM based approach

Index	Rule set obtained
1	IF 0.01<=aa<0.03 AND 0.012<=aa<0.027 AND 0.001<=aa<0.018 AND 0.018<=aa<0.028 AND 0.02<=aa<0.04 THEN gene is of class1
2	IF 0.011<=aa<0.022 AND 0.001<=aa<0.019 AND 0.019<=aa<0.041 AND 0.001<=aa<0.18 AND 0.017<=aa<0.037 THEN gene is of class1
3	IF 0.017<=aa<0.037 AND 0.011<=aa<0.022 AND 0.012<=aa<0.027 AND 0.03<=aa<0.06 THEN gene is of class1
4	IF 0.019<=aa<0.039 AND 0.024<=aa<0.049 AND 0.011<=aa<0.022 AND 0.049<=aa<0.074 AND 0.039<=aa<0.059 THEN gene is of class1
5	IF 0.039<=aa<0.059 AND 0.045<=aa<0.06 AND 0.038<=aa<0.05 AND 0.028<=aa<0.038 AND 0.075<=aa<0.1 THEN gene is of class1
6	IF 0.012<=aa<0.027 AND 0.028<=aa<0.038 AND 0.037<=aa<0.057 AND 0.057<=aa<0.077 THEN gene is of class1
7	IF 0.030<=aa<0.045 AND 0.037<=aa<0.057 AND 0.058<=aa<0.078 AND 0.019<=aa<0.039 AND 0.019<=aa<0.041 THEN gene is of class2
8	IF 0.041<=aa<0.063 AND 0.066<=aa<0.088 AND 0.03<=aa<0.045 THEN gene is of class2
9	IF 0.058<=aa<0.078 AND 0.014<=aa<0.029 AND 0.039<=aa<0.059 and 0.06<=aa<0.009 THEN gene is of class2
10	IF 0.02<=aa<0.017 AND 0.018<=aa<0.038 AND 0.038<=aa<0.051 AND 0.037<=aa<0.0057 AND 0.001<=a<0.015 THEN gene is of class2
11	IF 0.057<=aa<0.077 AND 0.06<=aa<0.09 AND 0.037<=aa<0.057 THEN gene is of class2
12	IF 0.001<=aa<0.018 AND 0.06<=aa<0.08 AND 0.014<=aa<0.029 AND 0.011<=aa<0.022 AND 0.022<=aa<0.044 THEN gene is of class2
13	IF 0.012<=aa<0.025 AND 0.045<=aa<0.06 AND 0.017<=aa<0.037 AND 0.025<=aa<0.038 AND 0.019<=aa<0.041 THEN gene is of class3
14	IF 0.041<=aa<0.063 AND 0.058<=aa<0.078 AND 0.001<=aa<0.024 AND 0.024<=aa<0.049 THEN gene is of class3
15	IF 0.044<=aa<0.066 AND 0.03<=aa<0.06 AND 0.06<=aa<0.08 AND 0.012<=aa<0.025 AND 0.028<=aa<0.038 THEN gene is of class3
16	IF 0.022<=aa<0.044 AND 0.044<=aa<0.059 AND 0.027<=aa<0.042 AND 0.025<=aa<0.038 THEN gene is of class3
17	IF 0.059<=aa<0.079 AND 0.001<=aa<0.018 AND 0.018<=aa<0.028 AND 0.024<=aa<0.049 THEN gene is of class3
18	IF 0.022<=aa<0.044 AND 0.02<=aa<0.04 AND 0.038<=aa<0.058 AND 0.06<=aa<0.09 THEN gene is of class4
19	IF 0.027<=aa<0.042 AND 0.002<=aa<0.012 AND 0.057<=aa<0.077 AND 0.017<=aa<0.037 AND 0.024<=aa<0.049 THEN gene is of class4
20	IF 0.024<=aa<0.049 AND 0.06<=aa<0.09 AND 0.001<=aa<0.018 AND 0.012<=aa<0.027 THEN gene is of class4
21	IF 0.01<=aa<0.02 AND 0.012<=aa<0.025 AND 0.041<=aa<0.063 THEN gene is of class4
22	IF 0.042<=aa<0.057 AND 0.044<=aa<0.066 AND 0.028<=aa<0.038 AND 0.03<=aa<0.06 THEN gene is of class4
23	IF 0.045<=aa<0.06 AND 0.012<=aa<0.027AND 0.039<=aa<0.059 AND 0.066<=aa<0.088 THEN gene is of class5
24	IF 0.01<=aa<0.03 AND 0.03<=aa<0.045 AND 0.012<=aa<0.025 AND 0.027<=aa<0.042 AND 0.057<=aa<0.077 THEN gene is of class5
25	IF 0.074<=aa<0.099 AND 0.037<=aa<0.057 AND 0.019<=aa<0.041 AND 0.058<=aa<0.078 AND 0.075<=aa<0.1 THEN gene is of class5
26	IF 0.06<=aa<0.09 AND 0.03<=aa<0.045 AND 0.019<=aa<0.039 AND 0.014<=aa<0.029THEN gene is of class5
27	IF 0.039<=aa<0.059 AND 0.042<=aa<0.057 AND 0.041<=aa<0.063 AND 0.045<=aa<0.06 THEN gene is of class5
28	IF 0.012<=aa<0.025 AND 0.022<=aa<0.044 AND 0.008<=aa<0.018 AND 0.001<=aa<0.015 THEN gene is of class5

7.Conclusion and future scope

In this work, we have presented a review on various gene recognizing programs. The various gene prediction programs were discussed, which are categorized into two main classes i.e. ab-initio methods and homology methods. In this paper, we have introduced a PSSM based approach for gene finding. The new approach performs the PSSM composition for each protein sequence. The PSSM profile is then given as an input to decision tree from where various rules were obtained. Each rule corresponds to a single gene. This new approach was implemented on DNAset dataset for predicting the genes. The accuracy is measured by considering the rules. It seems that new method generates the result better.

In recent years, diverse attention has been given for estimating the quantity of genes in the human genome. The genome annotation field reveals that on improving the exactness of gene detection algorithms, a number of challenging and difficult problems arise which requires substantial solutions. These methods are found to be exceptionally supportive in recognizing genes, but it still does not reflect the perfect results seeing that for the most part of job is done for definite genomes. Hence, supplementary exploration of gene prediction methods can be done.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] Wani MA. Incremental hybrid approach for microarray classification. In international conference on machine learning and applications 2008 (pp. 514-20). IEEE.

[2] Wani MA. Microarray classification using sub-space grids. In machine learning and applications and workshops 2011 (pp. 389-94). IEEE.

[3] Wani MA. Introducing subspace grids to recognise patterns in multidimensinal data. In international conference on machine learning and applications 2012 (pp. 33-9). IEEE.

[4] Wani MA, Yesilbudak M. Recognition of wind speed patterns using multi-scale subspace grids with decision trees. International Journal of Renewable Energy Research. 2013; 3(2):458-62.

[5] Wani MA. SAFARI: a structured approach for automatic rule. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2001; 31(4):650-7.

[6] Goel N, Singh S, Aseri TC. A comparative analysis of soft computing techniques for gene prediction. Analytical Biochemistry. 2013; 438(1):14-21.

[7] Bhat HF, Wani MA. Modified one-against-all algorithm based on support vector machine. International Journal of Advanced Research in Computer Science and Software Engineering. 2013.

[8] Bhat HF, Wani MA. A comparative study of five main support vector machine based multiclass classification algorithms. International Journal of Advance Foundation and Research in Science & Engineering. 2014; 1(2):1-6.

[9] Wani MA. Hybrid method for fast SVM training in applications involving large volumes of data. In international conference on machine learning and applications 2013 (pp. 491-4). IEEE.

[10] Wani MA, Bhat HF. Multiclass SVM algorithms for wind speed prediction. In international conference on

renewable energy research and applications 2017 (pp. 1139-43). IEEE.
[11] Khan AI, Wani MA. Efficient and rotation invariant fingerprint matching algorithm using adjustment factor. In international conference on machine learning and applications 2015 (pp. 1103-10). IEEE.
[12] Bhat FA, Wani MA. Performance comparison of major classical face recognition techniques. In international conference on machine learning and applications 2014 (pp. 521-8). IEEE.
[13] Mujtaba T, Wani MA. Daily global horizontal solar radiation forecasting using extreme learning machines. International conference on computing for sustainable global development (pp. 7290-5). IEEE.
[14] Bhat FA, Wani MA. Dropout technique based convolutional neural networks model for face recognition. Artificial Intelligent Systems and Machine Learning. 2017; 9(9):202-9.
[15] Bhat MR, Wani MA. Mixture weighted latent dirichlet allocation, an optimized and generalized probabilistic model for large corpus of data. Artificial Intelligent Systems and Machine Learning. 2018; 10(1):8-17.
[16] Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research. 2002; 30(19):4103-17.
[17] Xu Y, Mural RJ, Einstein JR, Shah MB, Uberbacher EC. GRAIL: a multi-agent neural network system for gene identification. Proceedings of the IEEE. 1996; 84(10):1544-52.
[18] Krogh A. Using database matches with HMMGene for automated gene detection in Drosophila. Genome Research. 2000; 10:523-8.
[19] Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA1. Journal of Molecular Biology. 1997; 268(1):78-94.
[20] http://genes.mit.edu/GENSCAN.html. Accessed 15 May 2018.
[21] Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. Genome Research. 2001; 11:803-16.
[22] Riyaz R, Wani MA. Local and global data spread based index for determining number of clusters in a dataset. In 15th IEEE international conference on machine learning and applications (ICMLA) 2016 (pp. 651-6). IEEE.
[23] Klasberg S, Bitard-Feildel T, Mallet L. Computational identification of novel genes: current and future perspectives. Bioinformatics and Biology Insights. 2016; 10:121-31.
[24] Goel N, Singh S, Aseri TC. A review of soft computing techniques for gene prediction. ISRN Genomics. 2013:1-8.
[25] Sleator RD. An overview of the current status of eukaryote gene prediction strategies. Gene. 2010; 461(1-2):1-4.
[26] Yandell M, Ence D. A beginners guide to eukaryotic genome annotation. Nature Reviews Genetics. 2012; 13(5):329-42.
[27] Guigo R, Knudsen S, Drake N, Smith T. Prediction of gene structure. Journal of Molecular Biology. 1992; 226(1):141-57.
[28] Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. Genome Research. 2000; 10:516-22.
[29] Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research. 2004; 32(suppl_ 2):309-12.
[30] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25(17):3389-402.
[31] Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST-a tool for discovery in protein databases. Trends in Biochemical Sciences. 1998; 23(11):444-7.
[32] Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences. 1987; 84(13):4355-8.
[33] Kumar M, Gromiha MM, Raghava GP. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics. 2007; 8.

Heena Farooq Bhat is a Ph.D Research Scholar in the Department of Computer Science, University of Kashmir, India. She completed her Bachelor's in Computer Science and Master's in Computer Science from University of Kashmir. Her Research areas include Data Mining, Machine Learning and Genome Analysis.
Email: heenafarooq14@gmail.com

Mohd Arif Wani is currently working as a Professor and Head in the Department of Computer Science, University of Kashmir, India. He has more than 100 publications in the International Journals and Conferences. His Research areas include Data Mining and Machine Learning.