STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720 (Online)
Volume 6, Issue 1
www.stmjournals.com

PSSM Amino-Acid Composition Based Gene Identification Using Support Vector Machines

Heena Farooq Bhat**, *M. Arif Wani
Department of Computer Science, University of Kashmir, Jammu and Kashmir, India

Abstract

The main characteristic of identifying 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 a 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. Firstly, 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. Finally, the PSSM vectors are given as an input to SVM for classification purpose. This new method has been demonstrated on genome DNA set dataset. It is shown that the experimental results of new approach produces better results.

Keywords: *Gene prediction, classification, feature extraction, binding proteins, rule induction, position specific scoring matrix*

****Author for Correspondence*** E-mail: heenafarooq14@gmail.com

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, where each sub-space grid represents a sub-problem [1–4]. 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. DNA binding proteins play 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 are then used to carry out a broad diversity of procedures in the unit. Other type of genes, known as non-coding genes, determines efficient genetic material (RNA) which is occupied in the guidelines of appearance of genes and production of proteins. This sequence of DNA is not allowed to transform into amino acids and hence be deficient in the distinctive sequence restriction of coding sequences.

The 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 [7]. The Support Vector Machines (SVM) is a supervised learning algorithm used to categorize reserved blueprint of data [8–10]. SVM has been applied to various other domains which incorporate wind speed prediction [11], fingerprint recognition [12], face recognition [7], 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 genome annotation 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) which assembles like a gene but it has not been proved to be a gene yet [23].

The remaining portion of the paper is organized as follows: In the next part of the paper, we first give the brief review of gene finding problem and then presents the various approaches of gene finding. Then we provide the concept of PSSM where profiles of every protein sequence is converted into unvarying numeric representation. The following part of the paper explains the approach based on PSSM and classifies using SVM, describes the dataset used, the evaluation of these approaches, and finally, conclusion and future research directions are given.

LITERATURE REVIEW

This describes the problem of finding genes.

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 deoxyribonucleic acid (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 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
$X=(x_1, \ldots ,_n) \in \Sigma^*$, where Σ={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 a 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 open reading frames (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 nonexistence 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 give 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 and 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 view in general: 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 incorporated 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, and
(iv) Subject sequences are not absolutely accurate.

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.

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 the two properties of having higher gene density and the absence of introns in their protein coding genes. That portion of coding genes, called as open reading frames (ORFs) which may be longer, is likely correspond to genes. The main problem which arises in this approach are frame-shift errors, that is a single insertion or

deletion might lead to completely dissimilar amino-acid; very small genes may get miss; overlapping of long ORFs on opposite strands of DNA may lead to ambiguity etc.

Homology Based Approach

The rationale behind comparative or similarity base 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.

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:

- *GenScan:* GenScan, a gene recognizing program that employs a multifaceted probabilistic representation of the gene composition [20]. 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 nonexistence 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 has developed into the preferred method for analyzing at the initial stage of long DNA sequences of eukaryotic genomes due to its high speed accuracy.
- *GenomeScan:* GenomeScan is based in part on GENSCAN. GenomeScan is a software program used to identify the coding and non-coding regions of g**e**nes in genomic DNA chain of characters from a large diversity of living being**s** [21]. The organism includes human and other vertebrates. This gene prediction program is based on two sources of information: 1) representation of coding and n**o**n-coding regions and splice signal composition; and 2) DNA/RNA sequence resemblance data such as BLASTX hits. The program is given an input which includes a DNA/RNA sequence formerly masked with RepeatMasker. The RepeatMasker is a file of parameters for the suitable living being and a 'Genoa file' which is full of a review of ac**c**essible 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.
- *GeneID:* Gen**e**ID was introduced as the first program to recognize full coding gene configuration of vertebrate genes in unstipulated DNA sequences [19]. 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 was 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 optimize. 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 kept count as a purpose of the scores of the exon essential sites and of the coding figures.
- *GeneMark:* GeneMark was found to be the earliest means for finding prokaryotic genes [27]. This tool has been used to engage a non-standardized Markov model to organize DNA sections into protein coding, non-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.

- *FGenesh:* FGenesh 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 [28]. This is HMM-based gene structure prediction (multiple genes, both chains). FGENESH capitulates 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.
- *AUGUSTUS:* AUGUSTUS is a tool for recognizing genes in eukaryotic genomic DNA strand of characters [29]. 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.

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. One of the main widespread used application is position-specific iterated BLAST (PSI-BLAST) [30, 31], which compares PSSM profiles for detecting remotely related homologous proteins or DNA. The original PSSM introduced by Gribskov *et al*. consists of the following components [32]:

- 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.
- Consensus, which is a sequence of amino acid residues that 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.

PSSM BASED APPROACH FOR GENE FINDING USING SVM

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

PSSM Matrix Representation of Protein Sequences

A position specific scoring matrix (PSSM) is a matrix that contains 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*' represents 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:

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

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 0–1 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.

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

Where, '*i*' belongs 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:

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

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.

Classification Using SVM

The support vector machine approach has been widely used in solving prediction problems in bioinformatics that can be represented in the form of a multiclass classification, such as gene identification, protein-protein interaction prediction and horizontally transferred gene detection. Selecting relevant features for machine learning approaches is important for a number of reasons such as generalization performance, running efficiency and feature interpretation. In this work, we have applied multiclass classification algorithm for classifying genes from each PSSM vector [33–39].

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 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 [40]. 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 (Table 1).

Table 1: *DNAset Dataset Used.*

Composition Type	# of Samples	
	# Binding Protein Sequences	# Non-Binding Protein Sequences
PSSM	146	250

RESULTS AND DISCUSSIONS

The numerical execution of new approach for gene DNA set dataset is estimated. 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). We divide the gene sequences into various groups according to same gene ids. We have obtained 30 groups based on same gene id where each group consists of 'N' number of gene sequences. The PSSM is obtained for each group containing 21 fields where 1–20 represent 20 amino acids and 21st column represents the genes. SVM is applied to each PSSM for classifying the correctly labeled genes.

After that, the training process is carried out where N/2 data is used to train the model and the remaining is used for testing. The same procedure is repeated for N times such that each sample is given an opportunity for testing the performance. Table 2 shows the accuracy rate evaluated for 30 groups where each group contains distinct number of gene sequences. The first column presents the number of groups, the second column represents the number of gene sequences for each group and the third column indicates the approach, i.e., SVM based PSSM approach where the accuracy rate for the new approach is estimated.

It is seen that the SVM produced better classification accuracy with the PSSM based approach. Compared to other methods, support vector machine yields the maximum percentage for correctly classified data with maximum number of genes (Figure 1).

CONCLUSIONS

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 SVM to classify the genes. This new approach was implemented on DNAset dataset for predicting the genes. The accuracy is measured by considering the number of genes classifying. It seems that new method generates better result.

Table 2: *The Performance of SVM Classifier Using PSSM for 30 Groups Each Containing 'N' Number of Gene Sequences.*

PSSM Group	Number of Gene Sequences per Group	Accuracy Rate
1	37	100
2	16	94
3	12	100
4	12	91.66
5	5	80
6	6	83.33
7	12	83.33
8	5	100
9	5	80
10	22	100
11	6	83.33
12	4	80
13	9	77.66
14	9	100
15	7	85.71
16	4	100
17	14	100
18	8	100
19	3	100
20	8	100
21	10	90
22	6	83.33
23	36	100
24	32	75
25	6	83.33
26	16	87.50
27	14	78.57
28	24	87.50
29	31	100
30	23	86.95

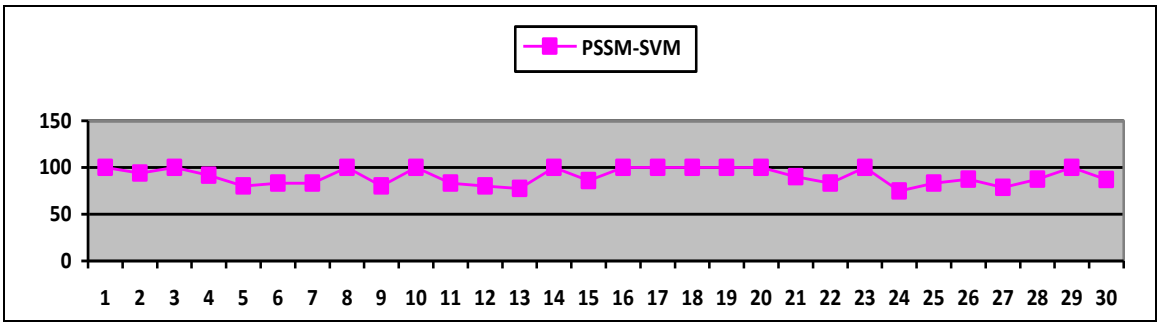

Fig. 1: *Accuracy Comparison of 30 Classes for PSSM Based Method Using SVM.*

A vast attention has been taken in the latest years for estimating the quantity of genes in the human genome. The genome annotation field put into practice that improves the exactness of gene detection is considered to be a very difficult and challenging task. 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.

REFERENCES

1. Wani MA. Incremental Hybrid Approach for Microarray Classification. *Proceedings of the 7th International Conference on Machine Learning and Applications*. 2008; 514–520p.
2. Wani MA. Microarray Classification Using Sub-Space Grids. *Proceedings of the 10th International Conference on Machine Learning and Applications*. 2011; 1: 389–394p.
3. Wani MA. Introducing Subspace Grids to Recognize Patterns in Multidimensional Data. *International Conference on Machine Learning and Applications*. 2012; 1: 33–39p.
4. Wani MA, Yesilbudak M. Recognition of Wind Speed Patterns Using Multi-Scale Subspace Grids with Decision Trees. *International Journal of Renewable Energy Research (IJRER)*. 2013; 3(2): 458–462p.
5. Wani MA. SAFARI: A Structured Approach for Automatic Rule. *IEEE Trans Syst Man Cybern Part B (Cybernetics)*. 2001; 31(4): 650–657p.
6. Goel N, Singh S, Aseri TC. A Comparative Analysis of Soft Computing Techniques for Gene Prediction. *Analytical Biochemistry*. 2013; 438(1): 14–21p.
7. Bhat FA, Wani MA. Performance Comparison of Major Classical Face Recognition Techniques. In *Machine Learning and Applications (ICMLA), 2014 13th International Conference on, IEEE*. 2014; 521–528p.
8. Bhat HF, Wani MA. Modified One-against-All Algorithm Based on Support Vector Machine. *Int J Adv Res Comput Sci Softw Eng*. 2013; 3(12): 972–975p.
9. 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 (IJAFRSE)*. 2014; 1(2): 35–45p.
10. Wani MA. Hybrid Method for Fast SVM Training in Applications Involving Large Volumes of Data. *12th International Conference on Machine Learning and Applications*, Miami, FL. 2013; 491–494p.
11. Wani MA, Bhat HF. Multiclass SVM Algorithms for Wind Speed Prediction. *International Conference on Renewable Energy Research and Applications (ICRERA)*. 2017; 1139–1143p.
12. Khan AI, Wani MA. Efficient and Rotation Invariant Fingerprint Matching Algorithm Using Adjustment Factor. In *Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on*. 2015; 1103–1110p.
13. Mujtaba T, Wani MA. Daily Global Horizontal Solar Radiation Forecasting Using Extreme Learning Machines. *4th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE*. 2017; 7290–7295p.
14. Bhat MR, Wani MA. Evaluating Algebraic Model Based Information Retrieval Algorithms for Small and Large Data Set. *4th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE*. 2017.
15. Bhat HF, Wani MA. Algorithms for Sequence Alignment. *4th International Conference on Computing for Sustainable Global Development (INDIACom),* IEEE. 2017; 4231–4236p.
16. Mathé C, Sagot MF, Schiex T, *et al.* Current Methods of Gene Prediction, Their Strengths and Weaknesses. *Nucleic Acids Res*. 2002; 30(19): 4103–4117p.
17. Xu Y, Mural RJ, Einstein JR, *et al.* GRAIL: A Multi-Agent Neural Network System for Gene Identification. *Proc. IEEE*. 1996; 84: 1544–1552p.
18. Krogh A. Using Database Matches with HMM Gene for Automated Gene

Detection in Drosophila. *Genome Res*. 2000; 10(4): 523–528p.
19. Burge C, Karlin S. Prediction of Complete Gene Structures in Human Genomic DNA. *J Mol Biol.* 1997; 268(1): 78–94p.
20. GeneScan Web Server. Available at www.genes.mit.pdf
21. Yeh RF, Lim LP, Burge CB. Computational Inference of Homologous Gene Structures in the Human Genome. *Genome Res*. 2001; 11(5): 803–816p.
22. Wani MA. Introducing Subspace Grids to Recognize Patterns in Multidimensional Data. *Proceedings 11th International Conference on Machine Learning and Applications, ICMLA 2012*. 2012; 1: 33–39p. 10.1109/ICMLA.2012.15.
23. Klasberg S, Bitard-Feildel T, Mallet L. Computational Identification of Novel Genes: Current and Future Perspectives. *Bioinform Biol Insights*. 2016; 10: 121–131p.
24. Goel N, Singh S, Aseri TC. A Review of Soft Computing Techniques for Gene Prediction. *ISRN Genomics*. 2013; Article ID 191206: 8p.
25. Sleator RD. An Overview of the Current Status of Eukaryote Gene Prediction Strategies. *Gene*. 2010; 461(1–2): 1–4p.
26. Yandell M, Ence D. A Beginner's Guide to Eukaryotic Genome Annotation. *Nat Rev Genet*. 2012; 13(5): 329–342p.
27. Guigo R, Knudsen S, Drake N, *et al.* Prediction of Gene Structure. *J Mol Biol.* 1992; 226(1): 141–157p.
28. Salamov AA, Solovyev VV. Ab initio Gene Finding in Drosophila Genomic DNA. *Genome Res*. 2000; 10(4): 391–393p.
29. Stanke M, Steinkamp R, Waack S, *et al.* A Webserver for Gene Finding in Eukaryotes. *Nucleic Acids Res.* 2004; 32: w309–w312p.
30. Altschul SF, *et al*. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res*. Sep 1997; 25(17): 3389–3402p.
31. Altschul SF, Koonin EV. Iterated Profile Searches with PSIBLAST: A Tool for Discovery in Protein Databases. *Trends Biochem Sci*. Nov 1998; 23(11): 444–447p.
32. Gribskov M, *et al.* Profile Analysis: Detection of Distantly Related Proteins. *Proc. Nat'l Academy of Sciences*, USA. Jul 1987; 84: 4355–4358p.
33. Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995; 20(3): 273–97p.
34. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min Knowl Discov.* 1997; 2(2): 121–67p.
35. Liu T, Zheng X, Wang J. Prediction of Protein Structural Class for Low-Similarity Sequences Using Support Vector Machine and PSI-BLAST Profile. *Biochimie*. 2010; 92(10): 1330–1334p.
36. Liu Y, Guo J, Hu G, *et al.* Gene Prediction in Meta-Genomic Fragments Based on the SVM Algorithm. In *BMC Bioinformatics*. Apr 2013; 14(5): S12p.
37. Muller K-R, Mika S, Ratsch G, *et al.* An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans Neural Netw.* 2001; 12(2): 181–201p.
38. Ma X, Wu J, Xue X. Identification of DNA-Binding Proteins Using Support Vector Machine with Sequence Information. *Comput Math Methods Med.* 2013; 2013: 8p. Article id 524502.
39. Muthukrishnan S, Puri M, Lefevre C. Support Vector Machine (SVM) Based Multiclass Prediction with Basic Statistical Analysis of Plasminogen Activators. *BMC Res Notes*. 2014; 7: Article 63.
40. Kumar M, Gromiha MM, Raghava GP. Identification of DNA-Binding Proteins Using Support Vector Machines and Evolutionary Profiles. *BMC Bioinformatics*. 2007; 8(1): 463p.

Cite this Article
Heena Farooq Bhat, Arif Wani M. PSSM Amino-Acid Composition Based Gene Identification Using Support Vector Machines. *Journal of Artificial Intelligence Research & Advances*. 2019; 6(1): 50–58p.