Journal of
Data Mining in Genomics & Proteomics

Sohpal et al,. J Data Mining Genomics Proteomics 2013, S3
DOI: 10.4172/2153-0602.S3-001

Research Article Open Access

N-W Algorithm and ANFIS Modeling on Alignment Similarity of Triplex Capsid Protein of Human Herpes Simplex Virus

Vipan Kumar Sohpal[1]*, Apurba Dey[2] and Amarpal Singh[3]

[1]*Department of Chemical and Bio Technology, Beant College of Engineering and Tech, Gurdaspur, Punjab, India*
[2]*Department of Bio Technology, National Institute of Technology, Durgapur, West Bengal, India*
[3]*Department of Electronics and Communication Engg, Beant College of Engineering and Tech, Gurdaspur, Punjab, India*

Abstract

Optimal sequence similarity of triplex capsid proteins of human herpes simplex virus (HHV) is a complex bioinformatics problem, which is controlled by alignment algorithms, substitution matrix, gap penalty and gap extension. A precise choice of mutation matrix is required to optimal the alignment similarity and appropriate computational approach required for similarity search. The present paper uses Adaptive Neuro-Fuzzy Inference System (ANFIS) approach to model and simulate the alignment similarity for PAM and Blosum substitution matrices. Mutation matrix and sequences of HHV-I and HHV-II were taken as model's input parameters. The model is the combination of fuzzy inference, artificial neural network, and set of fuzzy rules has been developed directly from computational analysis using N-W algorithm. The proposed modeling approach is verified by comparing the expected results with the observed practical results obtained by computational analysis under specific conditions. The application of ANFIS test shows that the substitution matrix predicted by a proposed model is fully in agreement with the experimental values at 0.5% level of significance.

Keywords: Sequence similarity; Substitution matrix; Triplex capsid protein; Human Herpes Simplex Virus (HHV); ANFIS

Introduction

Sequence alignment is the most fundamental technique in bioinformatics for establishing evolutionary relationship between different bio-molecules and biological species. Sequence comparison also offers bases for medical diagnosis and drug development. There are many computational models and approaches that can be applied to sequence alignment. These models can be classified on the basis of algorithms and alignment techniques. Dynamic programming methods using N-W and S-W approach is prominent in pairwise sequence alignment. Sequence alignment systems and variables (gap penalty, extension and substitution matrices) are intrinsically fuzzy, as their properties and behaviors contain uncertainty. Fuzzy logic and modeling are ideal to describe sequence alignment and provide robust tools for optimization of variables. In addition, it has been shown that exact or optimal solutions have significant limitations in the sequence alignment problems. The applications of fuzzy concepts and approaches have been also growing in the sequence alignment and phylogenetic analysis to overcome the randomness.

Zhang et al. [1] evaluated several validity measures in fuzzy clustering and develop a new measure for a fuzzy c-means algorithm, which uses a Pearson correlation in its distance metrics. They observed that newly developed measure could be used to assess the validity of fuzzy clusters produced by correlation-based fuzzy c-means clustering algorithm. Garcia et al. [2] proposed FISim, a similarity measure between PFMs, based on the fuzzy integral of the distance of the nucleotides, with respect to the information content of the positions. FISim provides excellent results when dealing with sets of randomly generated motifs. Mansoori et al. [3] proposed a fuzzy rule-based classifier for assigning amino acid sequences into different super-families of proteins. The obtained results show that the generated fuzzy rules are more interpretable, with acceptable improvement in the classification accuracy. Bidargaddi et al. [4] proposed a fuzzy profile HMM to overcome the limitations, and to achieve an improved alignment for protein sequences belonging to a given family. The strong correlations and the sequence preference involved in the protein structures make this fuzzy architecture based model as a suitable candidate for building profiles of a given family. Espadaler et al. [5] introduced a computational approach for annotation of enzymes, based on the observation that similar protein sequences are more likely to perform the same function, if they share similar interacting partners. They observed this method could increase 10% the specificity of genome wide enzyme predictions based on sequence matching by PSI-BLAST alone. Gomez et al. [6] described a new method that predicts the putative function for the protein, integrating the results from the PSI-BLAST program and a fuzzy logic algorithm. Collyda et al. [7] deal with phylogenetic analysis of protein and gene data, using multiple sequence alignments produced by fuzzy profile Hidden Markov Models. The results of the analysis are compared against those obtained by the classical profile HMM model, and depict the superiority of the fuzzy profile HMM in this field. Brylinski et al. [8] described a computational model that can be used to identify potential areas that are able to interact with other molecules (ligand, substrates and inhibitors). Samsonova et al. [9] proposed a rule-based characterization of olfactory receptors derived from a multiple sequence alignment of human GPCRs. They concluded that seven alignment sites are sufficient to characterize 99% of human olfactory GPCRs. Huang et al. [10] proposed an efficient nonparametric classifier for predicting enzyme subfamily class, using an adaptive fuzzy r-nearest neighbor (AFK-NN) method, where k and a fuzzy strength parameter m are adaptively specified. The accuracy of AFK-NN on the

***Corresponding author:** Vipan Kumar Sohpal, Department of Chemical and Bio Technology, Beant College of Engineering and Tech, Gurdaspur, Punjab, India, E-mail: vipan752002@gmail.com

Received March 04, 2013; **Accepted** March 27, 2013; **Published** April 04, 2013

Citation: Vipan Kumar S, Dey A, Singh A (2013) N-W Algorithm and ANFIS Modeling on Alignment Similarity of Triplex Capsid Protein of Human Herpes Simplex Virus. J Data Mining Genomics Proteomics S3: 001. doi:10.4172/2153-0602.S3-001

Copyright: © 2013 Vipan Kumar S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

new large-scale dataset of oxido-reductases family is 83.3%, and the mean accuracy of the six families is 92.1%. Collyda et al. [11] proposed a novel method for aligning multiple genomic or proteomic sequences, using a fuzzy field Hidden Markov Model (HMM). The resultant shows that Fuzzy HMMs increase the model capability of aligning multiple sequences, mainly in terms of computation time. Popescu et al. [12] introduced fuzzy measure similarity (FMS), that consider the context of both complete sets of annotation terms, when computing the similarity between two gene products. Liang [13] detected c-WINNOWER algorithm substantially improved the sensitivity of the winnower method of with fuzzy logic. Boeva et al. [14] focused on Fuzzy tandem repeats (FTRs). They obtained formulas for P-values of FTR occurrence, and developed an FTR identification algorithm implemented in Tandem-SWAN software.

Shen et al. [15] approached supervised fuzzy clustering by utilizing the class label information during the training process. It is anticipated that the current predictor may play a pivotal complementary role to other existing predictors in this area, to further strengthen the power in predicting the structural classes of proteins and their other characteristic attributes. Atchley and Fernandes [16] combined information theory with fuzzy logic search procedures to identify sequence signatures or predictive motifs for members of the Myc-Max-Mad transcription factor network. Jacob et al. [17] used fuzzy guided genetic algorithm-based approach makes it possible to use diverse biological information like genome sequence data, functional annotations and conservation across multiple genomes, to guide the organization process. Liang et al. [18] found that c-WINNOWER algorithm substantially improves the sensitivity of the winnower method of Pevzner and Sze, by imposing a consensus constraint, enabling it to detect much weaker signals. Brodie et al. [19] developed a software package, Base-By-Base, that provides visualization, tools to enable researchers to 1) rapidly identify and correct alignment errors in large, multiple genome alignments; and 2) generate tabular and graphical output of differences between the genomes at the nucleotide level. Huang and Li [20] introduced fuzzy k-nearest neighbors (k-NN) algorithm to predict proteins' sub cellular locations from their dipeptide composition. The result demonstrates the applicability of this relative straightforward method, and possible improvement of prediction accuracy for the protein sub cellular locations. Blankenbecler et al. [21] presented a structure-alignment method, where the problem mapped onto a cost function containing both fuzzy (Potts) assignment variables and atomic coordinates. The approach performs exceptionally well when compared with other methods, requires modest central processing unit consumption, and is robust with respect to choice of iteration parameters for a wide range of proteins. Heger and Holm [22] addressed the problem by a fuzzy alignment model, which probabilistically assigns residues to structurally equivalent positions (attributes) of the proteins. Torres and Nieto [23] introduced the space of fuzzy poly-nucleotides, and a means of measuring dissimilitude between them and establish mathematical principles to measure dissimilarities between fuzzy poly-nucleotides. Schlosshauer and Ohlsson [24] proposed a method based on a fuzzy recast of the dynamic programming algorithm for sequence alignment, in terms of mean field annealing. They demonstrate that the value of the reliability index can directly be translated into an estimate of the probability for a correct alignment.

From the prior studies, it reveals that protein sequence alignment algorithms in computational biology have been improved by different methods. In this paper, we propose protein N-W sequence alignment technique that provides quality information and a fuzzy inference method developed, based on the alignment similarity of triplex capsid protein sequence. A fuzzy logic developed in order to improve conventional protein sequence alignment method that uses protein sequence quality information.

Database and Substitution Matrices

In this work, protein datasets of human herpes simplex virus database are taken as the testing sets for alignment similarity from UniProt KB. Here we adopt the UniProt database for triplex capsid protein of HHV-I and HHV-II, and to reduce the redundant sequences in UniProt, and to analyze the similarity and score detection on the datasets with different sequence identities, the accession number (P32888, P22486 and P89461 of HHV-I and HHV-II) of UniProt are used. In this paper, BLOSUM and PAM matrices are used to find alignment similarity for relationship among sequences within different strains of HHV.

Methodology

In conventional algorithms, protein sequence alignment similarity is determined using the global sequence alignment algorithm proposed by N-W, which is established for using quality information of each protein fragment. However, there may be errors in the process of calculating protein sequence alignment scores when the quality of protein fragment tips is low, because multiple parameters (substitution matrices, gap, penalty and extension) is influencing the overall protein sequence quality information. In our proposed method, a quality sequence alignment can be achieved by improvement of conventional algorithms by back trace of N-W algorithms and fuzzy modeling as shown in figure 1.

Fuzzy modeling of computational analysis (Sequence similarity) based on ANFIS

The modeling of computational analysis for alignment similarity has been ANFIS, by considering the input parameters such substitution matrices, sequences of HHV-I and II and the alignment similarity. In ANFIS, parameters associated with the membership functions will change through the learning process. The computation of these parameters is facilitated by the gradient vector, which tells fuzzy working and modeling the FIS for input/output data for a set of parameters. Once the gradient vector is achieved, any of the several optimization routines could be applied in order to improve the parameters, so as to minimize some error measure. Some data points are located and have been used as an input for the training of the FIS.

Architecture of the ANFIS

Fuzzy logic approach contains a potential to produce a simplified control for various bioinformatics applications. The rule-based

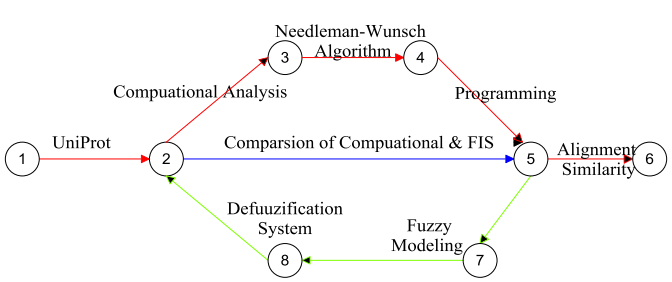

Figure 1: Framework of the proposed model for alignment similarity using computational & FIS analysis.

properties of fuzzy models allows for a model interpretation, in a way that is similar to the one humans use to describe reality. Conventional methods for statistical validation based on numerical data can be complemented by the human knowledge that usually involves heuristic knowledge and intuition. A multi-input single output (MISO) fuzzy model of computational analysis for sequence similarity has been developed using ANFIS, by considering two input parameters and one output variable, in order to predict the alignment similarity for Blosum. Similar environment setup for point accepted mutation matrices to evaluate the grid and sub clustering fuzzy system on computational data. This technique provides a method for the fuzzy modeling method to understand information about a data set, in order to evaluate the membership function parameters that best provide the associated FIS to track the given input/output data. This learning system works similarly to that of neural networks. The parameters associated with the membership functions will change through the learning process. This system is based on Sugeno-type fuzzy interface system, and can simulate and analyze the mapping relation between the input and output data through blended knowledge, to determine the optimal allocation of membership function. The ANFIS architecture of the type from Takagi and Sugeno is shown in figure 2 for alignment similarity of PAM.

It composes of five layers in this inference system. Each layer involves several nodes, which are described by the node function. The output signals from nodes in the previous layers will be accepted as the input signals in the present layer. After manipulation by the node function in the present layer, the output will be served as input signals for the next layer. To simply explain the mechanism of the ANFIS, we consider two inputs, x and y, and one output f in the FIS. Hence, the rule base will contain fuzzy "if–then" rules as follows:

Rule 1: If x is A1 and y is B1, then f=p1x+q1y+r1

Rule 2: If x is A2 and y is B2, then f=p2x+q2y+r2

Fuzzy inference system

The core of a fuzzy logic controller/modeling is the inference engine, which contains information of the control strategy in the form of "if–then" rules. Since the fuzzy logic, rules require linguistic variables. Inputs and outputs of a process are generally continuous crisp values, the conversion of crisp values into fuzzy values and vice versa are required. The initial step of the fuzzy modeling approach is to determine the input and output variables of the fuzzy logic controller. Sugeno type FIS is taken for the purpose. A typical direct fuzzy logic control system is shown in figure 2. The ANFIS editor is used to create, train and assess the Sugeno fuzzy logic. This FIS system is designed for the MISO system. This MISO system includes two inputs and one output.

Identification of input and output variables

The fuzzy logic is based on the identification of the fuzzy set that represents the possible values of the variables. In figure 3, a block diagram of the fuzzy control process along with the computational system of sequence alignment and score is shown. In particular, figure 4 shows real inputs and real output. Fuzzy model described in this article is a MISO system with two input parameters substitution matrices, and sequences and output as alignment similarity. Possible universe of discourse for the input parameters is given below:

Input parameters: Substitution Matrices=PAM (10-300) and BLOSUM (30-80)

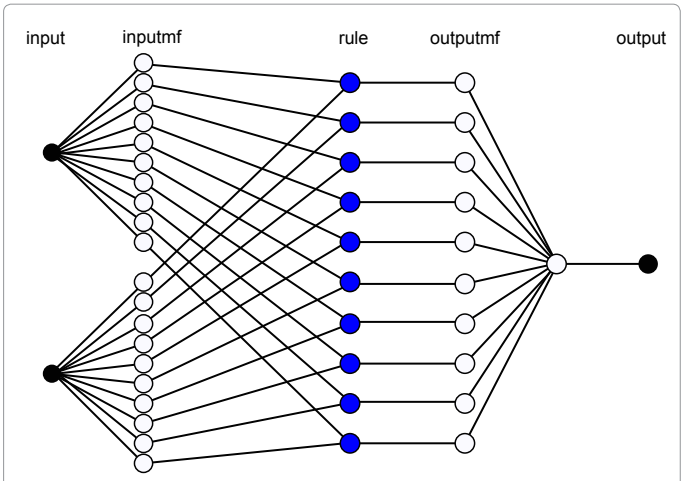

Figure 2: ANFIS architecture of (Takagi and Sugeno) for alignment similarity of PAM substitution matrices.

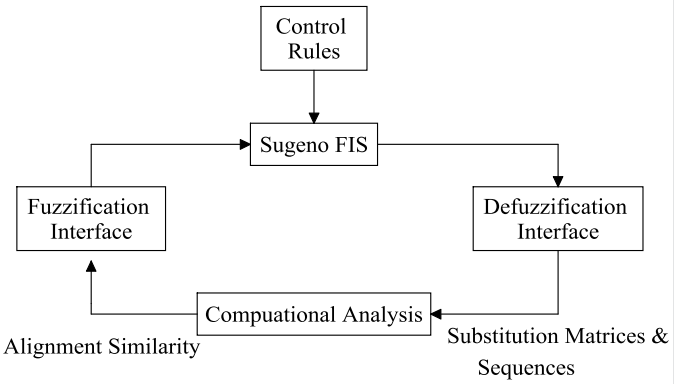

Figure 3: Block diagram for fuzzy control system on computational analysis of similarity.

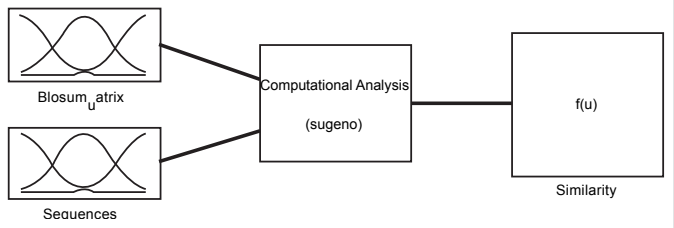

Figure 4: Fuzzy model of sequence similarity showing inputs and output.

Sequences=3 (Triplex capsid protein of HHV-I and II)

Output parameter: Alignment Similarity=Predicted as percentage of aligned pairs

(Depending upon input parameters)

Membership functions for the input and output variables for ANFIS modeling with sub-clustering

In this process, linguistic values assigned to the variables, and that was performed using fuzzy subsets and their associated membership functions. A membership function assigns numbers between 0 and

1. Zero membership value represents that it is not a member of the fuzzy set, while one represents a complete member of fuzzy set. A membership function can have any shape, and the standard shapes of membership functions include trapezoidal, triangular and bell shaped. Modeling with grid partition involves three membership functions that were produced for each input variable of substitution matrices, and sequences based on ANFIS. The in1mf1, in1mf2, in1mf3 are three linguistic levels for substitution matrix, and in2mf1, in2mf2, in2mf3 are for sequences are shown in figures 5a-c, that are applicable for both the matrices. Modeling sub-clustering for BLOSUM involves eight membership functions that were produced for each input variable of substitution matrices (BLOSUM), and sequences based on ANFIS. The in1cluster1, in1cluster2, in1cluster3, in1cluster4, in1cluster5, in1cluster6, in1cluster7 and in1cluster8 are eight linguistic levels for variable mutation matrix and in2cluster1, in2cluster2, in2cluster3, in2cluster4, in2cluster5, in2cluster6, in2cluster7, in2cluster8, are for sequences variable over a given universe of discourse as shown in figures 6a-c. On the other hand, ANFIS modeling with sub-clustering for PAM, having ten membership functions were generated for same two input variable of PAM matrices, and ten membership functions for sequences of protein data based on ANFIS. The in1cluster1, in1cluster2, in1cluster3, in1cluster4, in1cluster5, in1cluster6, in1cluster7, in1cluster8, in1cluster9 and in1cluster10 are ten linguistic levels for variable PAM matrix and in2cluster1, in2cluster2, in2cluster3, in2cluster4, in2cluster5, in2cluster6, in2cluster7, in2cluster8, in2cluster9 and in2cluster9 are for biological sequences of protein variable over a given universe of discourse. Output variable alignment similarity is also having ten membership functions. The span of each function was tuned within the specified range. Tests were conducted to evaluate the response parameters, and the span was varied accordingly for improvement. After a few iterations, the final membership functions for the system were determined as shown in the respective figures.

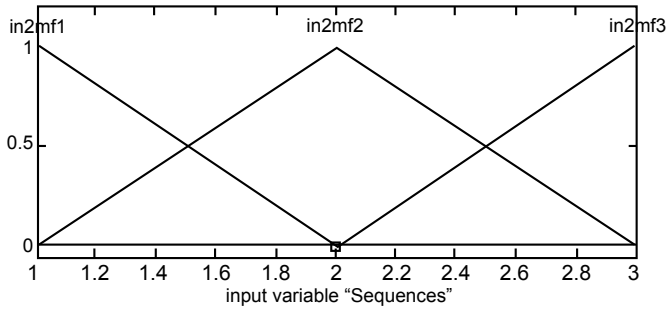

Figure 5a: Membership function plots of input variable sequence.

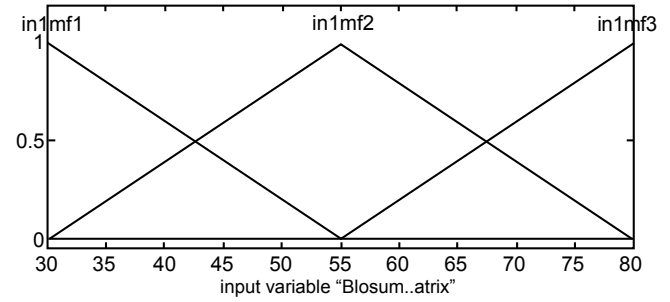

Figure 5b: Membership function plots of input variable BLOSUM matrix.

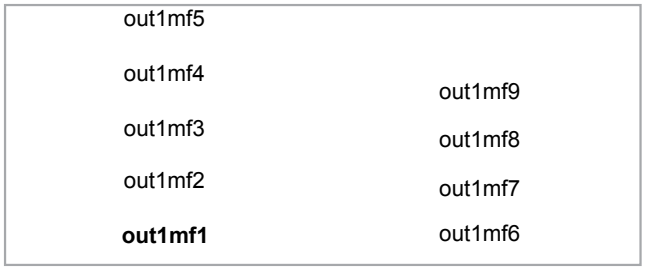

Figure 5c: Membership function plots of output variable similarity.

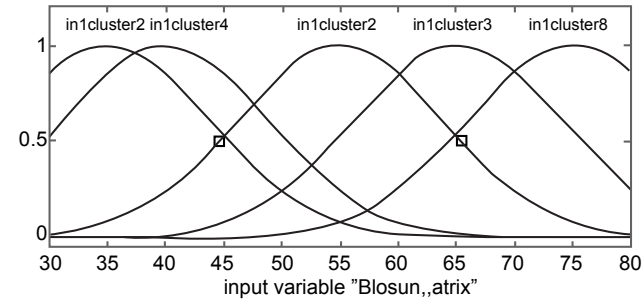

Figure 6a: Membership function plots of input variable BLOSUM matrix.

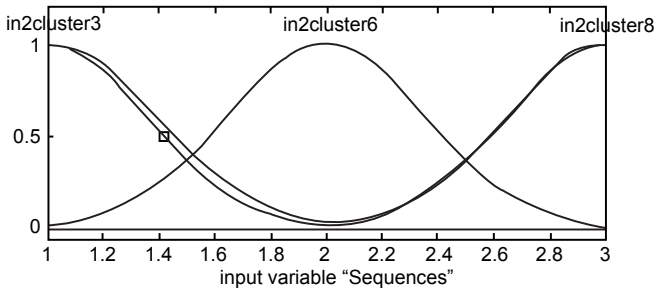

Figure 6b: Membership function plots of input variable sequences.

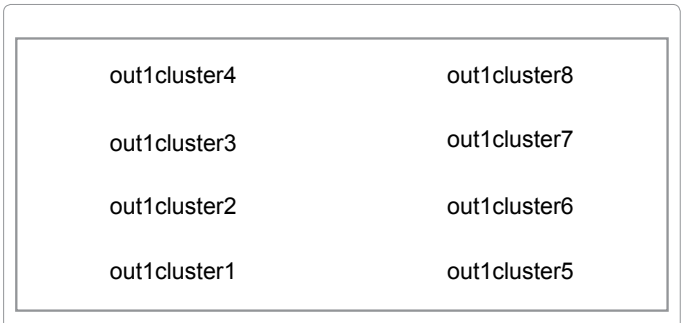

Figure 6c: Membership function plots of output variable similarity.

FIS rules employed in sequence similarity

The fuzzy modeling of the sequence similarity has been undertaken as design of the MISO fuzzy model with two inputs, i.e. substitution matrix and sequences, each of which being determined for 3 linguistic variables using ANFIS modeling with grid partition. These variables generate 9 numbers of conditional statements, as "if-and-then" rules of the model. Formulated set of rules of the model are outlined below:

1. IF (BLOSUM_Matrix is in1mf1) and (Sequences is in2mf1), then (Similarity is out1mf1) (1)
2. IF (BLOSUM_Matrix is in1mf1) and (Sequences is in2mf2), then (Similarity is out1mf2) (1)
3. IF(BLOSUM_Matrix is in1mf1) and (Sequences is in2mf3), then (Similarity is out1mf3) (1)
4. IF(BLOSUM_Matrix is in1mf2) and (Sequences is in2mf1), then (Similarity is out1mf4) (1)
5. IF(BLOSUM_Matrix is in1mf2) and (Sequences is in2mf2), then (Similarity is out1mf5) (1)
6. IF(BLOSUM_Matrix is in1mf2) and (Sequences is in2mf3), then (Similarity is out1mf6) (1)
7. IF(BLOSUM_Matrix is in1mf3) and (Sequences is in2mf1), then (Similarity is out1mf7) (1)
8. IF(BLOSUM_Matrix is in1mf3) and (Sequences is in2mf2), then (Similarity is out1mf8) (1)
9. IF(BLOSUM_Matrix is in1mf3) and (Sequences is in2mf3), then (Similarity is out1mf9) (1)

Similarly the substitution matrices and sequences, both of variables uses for linguistic variables using ANFIS modeling with sub-clustering. These variables generate 8 numbers of conditional statements as "if-and-then" rules of the model for BLOSUM matrix. Formulated set of rules of the model are outlined below:

1. IF (BLOSUM _Matrix in1cluster1) and (Sequence 2inclutser1), then (Similarity is out of cluster 1)(1)
2. IF (BLOSUM _Matrix in1cluster2) and (Sequence 2inclutser2), then (Similarity is out of cluster 2)(1)
3. IF (BLOSUM _Matrix in1cluster3) and (Sequence 2inclutser3), then (Similarity is out of cluster 3)(1)
4. IF (BLOSUM _Matrix in1cluster4) and (Sequence 2inclutser4), then (Similarity is out of cluster 4)(1)
5. IF (BLOSUM _Matrix in1cluster5) and (Sequence 2inclutser5), then (Similarity is out of cluster 5)(1)
6. IF (BLOSUM _Matrix in1cluster6) and (Sequence 2inclutser6), then (Similarity is out of cluster 6)(1)
7. IF (BLOSUM _Matrix in1cluster7) and (Sequence 2inclutser7), then (Similarity is out of cluster 7)(1)
8. IF (BLOSUM _Matrix in1cluster8) and (Sequence 2inclutser8), then (Similarity is out of cluster 8)(1)

Same input variables generate 10 numbers of conditional statements, as "if-and-then" rules of the model for PAM matrix. Formulated set of rules of the model are outlined below.

1. IF (PAM_ Matrix in1cluster1) and (Sequence 2inclutser1), then (Similarity is out of cluster 1)(1)
2. IF (PAM _Matrix in1cluster2) and (Sequence 2inclutser2), then (Similarity is out of cluster 2)(1)
3. IF (PAM _Matrix in1cluster3) and (Sequence 2inclutser3), then (Similarity is out of cluster 3)(1)
4. IF (PAM _Matrix in1cluster4) and (Sequence 2inclutser4), then (Similarity is out of cluster 4)(1)
5. IF (PAM _Matrix in1cluster5) and (Sequence 2inclutser5), then (Similarity is out of cluster 5)(1)
6. IF (PAM _Matrix in1cluster6) and (Sequence 2inclutser6), then (Similarity is out of cluster 6)(1)
7. IF (PAM _Matrix in1cluster7) and (Sequence 2inclutser7), then (Similarity is out of cluster 7)(1)
8. IF (BLOSUM _Matrix in1cluster8) and (Sequence 2inclutser8), then (Similarity is out of cluster 8)(1)
9. IF (BLOSUM _Matrix in1cluster9) and (Sequence 2inclutser9), then (Similarity is out of cluster 9)(1)
10. IF (BLOSUM _Matrix in1cluster10) and (Sequence 2inclutser10), then (Similarity is out of cluster 10)(1)

Figures 7a and 7b indicates rule viewer that shows the values of the various input to the model and computed outputs. Here, the similarity (output) can be predicted by varying the input parameters substitution matrix (BLOSUM) and sequences. Figure 7a shows a particular instance with grid partition having input values given to the system 55 for BLOSUM matrix, and sequence 2 (sequence 13) for analysis. The output generated by the system for sequence similarity is shown as 86%; while figure 7b stands for sub clustering ruler view having sequence similarity 85.8%.

Similarly, the output generated with sub clustering for the PAM matrix of 150 and sequence 2 (sequences 13) having similarity 92.3%, as shown in figure 7c. Likewise, this fuzzy model has generated other

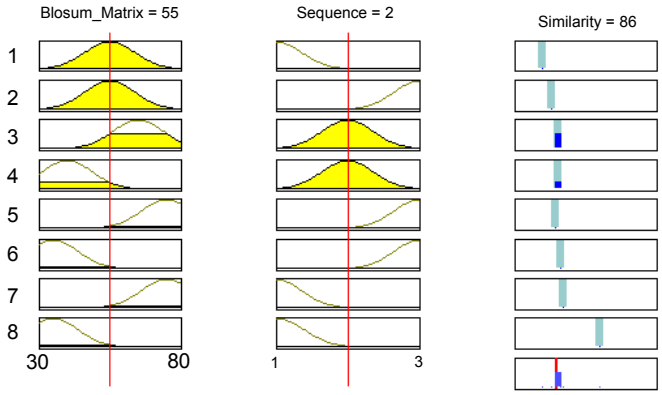

Figure 7a: Ruler view of input variable (BLOSUM matrix and sequences) with output variable alignment using grid partition.

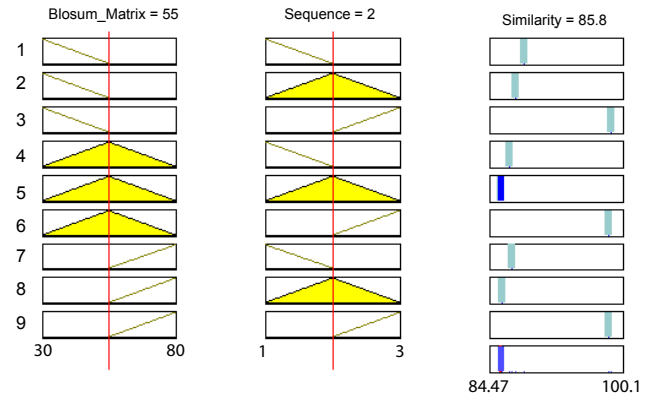

Figure 7b: Ruler view of input variable (BLOSUM matrix and sequences) with output variable alignment using sub clustering.

values of output variable for different set of data points in the specified range of input variables.

Figures 8a and 8b are showing two different views of control surfaces, which are indicating the results predicted by the fuzzy model for different sets of data points. These control surfaces as shown give the interdependency of input and output parameters guided by the various rules in the given universe of discourse. It has already been finalized that there are eight rules predicting the similarity, in case of BLOSUM matrix for MISO fuzzy model. In addition to that, ten rules are also set up for PAM matrix selection in sub clustering system. These rules were implemented in MATLAB environment, using the Sugeno type of FIS in fuzzy logic toolbox. Results predicted from this fuzzy model of alignment similarity have been compared with the experimental results, for its validation in the latter part of the article.

From the Fuzzy modeling, it has been observed FIS system with sub clustering results are more closely to computational biology output. So, the Fuzzy C means (FCM) an option to assess the similarity data. FCM is data clustering techniques which assign each data point in the dataset, a degree of membership to each cluster. It effectively allows for sharing of objects between clusters. FCM depends upon the exponent, maximum iteration and minimal improvement, in addition to number of clusters.

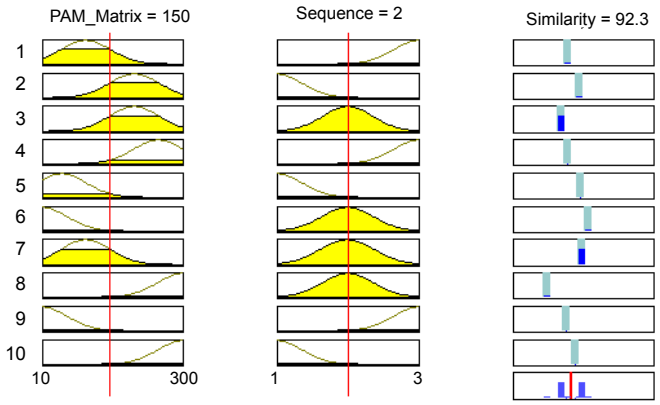

Figure 7c: Ruler view of input variable (PAM matrix and sequences), with output variable alignment using sub clustering.

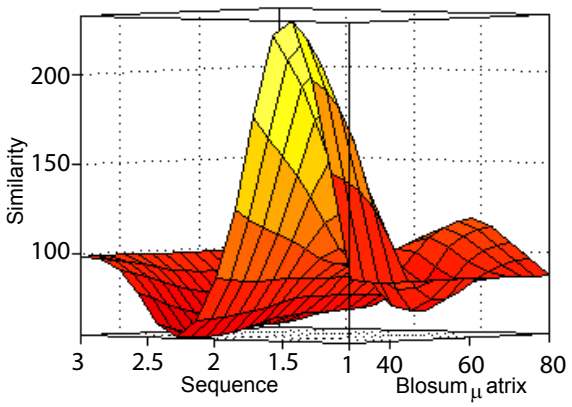

Figure 8a: Control surface view of the fuzzy model with sub clustering of BLOSUM Matrix.

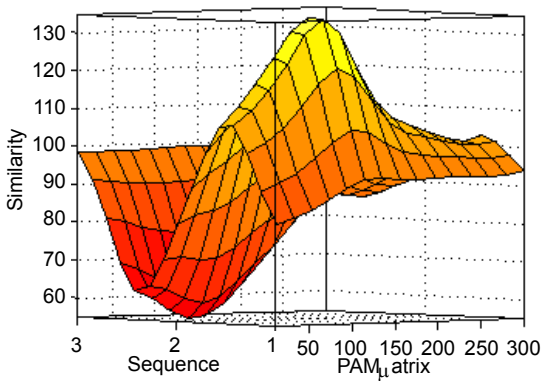

Figure 8b: Control surface view of the fuzzy model with sub clustering of PAM Matrix.

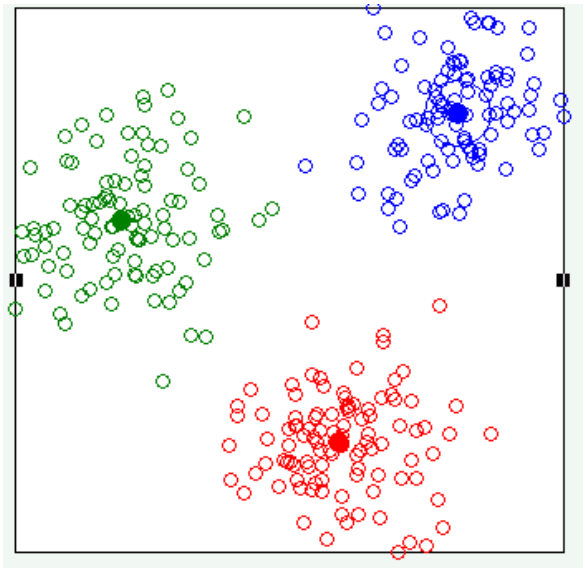

Figure 9a: Fuzzy C Mean (FCM) surface view of the similarity analysis of BLOSUM Matrix.

The results suggest that individual parameter have membership greater than 1.0 among the cluster value of similarity, and shown in figures 9a and 9b. It justifies the close relationship among statistical data of HHV-I and HHV-II for N-W alignment.

Results and Discussion

The alignment similarity of three sequences (namely Sequence 12, Sequence 13 and Sequence 23), with the substitution matrices, are shown in figure 10. Figures 10a and 10b show the curves of alignment similarity (percentage) versus the substitution matrices of these three sequences for PAM and Blosum matrices. The alignment similarity for sequence 23 is highest among three sequences, irrespective substitution matrices. However, as shown in figures 10a and 10b, the similarity percentage of the two closely related sequences is close (sequences 12 and 13). It is worth noting that when the value of substitution matrices is changing, the respective variation in these two sequences is small. When there are higher substitution matrices, the alignment similarity of protein using BLOSUM40 is better than

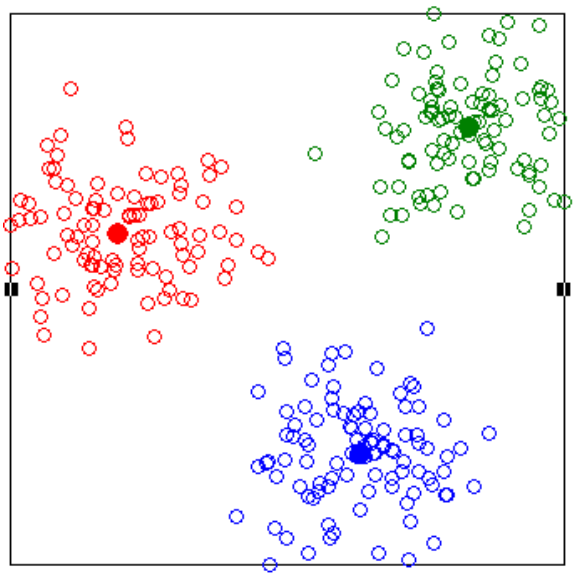

Figure 9b: Fuzzy C Mean (FCM) surface view of the similarity analysis of PAM Matrix.

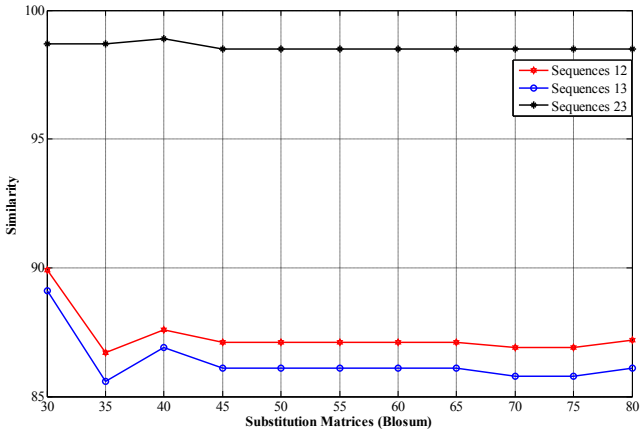

Figure 10a: Alignment similarities versus substitution matrices (BLOSUM) for different pair wise sequences using N-W alignment.

that of BLOSUM62 for all sequences. When the number substitution matrices increased further, the alignment similarity detection using BLOSUM75 is lower than those using other BLOSUM 40. But when substitution matrices value is at the lowest, the alignment similarity of the detection using BLOSUM30 increases remarkably. It seems that the number of the correctly recognized pairs is much larger than that of the incorrect recognitions, virtue of that the alignment observed using BLOSUM30 is high. Moreover, it is authenticated from the results that less divergent species have more correctly pair at lowest BLOSUM matrices. Thus, when a large number of correct recognitions are allowed, alignment score using BLOSUM30 may be more sensitive to pairs. When comparing alignment similarity predictions obtained using the standard BLOSUM matrix versus PAM, we first noticed that similarity obtained with PAM at low matrices, have overall smaller variance than the ones obtained with BLOSUM. Sequences 23 again produce significant in term of similarity (%). This again justifies that PAM substitution matrices is conversely true for BLOSUM for closely related strains. This is also confirmed by alignment similarity (Figures 11a and 11b), which again gives similar results, as would be expected in reverse substitution matrices. The results show BLOSUM to be the better matrix series, PAM. The PAM series is a subsidiary inferior to BLOSUM, in all the tests (similarity) reported. The highest percentage of alignment similarity observed in all three sequences of triplex capsid proteins in lower BLOSUM and higher PAM substitution matrices.

The winning path is represented through dots in the scoring space, which itself is a heat map displaying the highest scores for all the partial alignments of HHV-I and HHV-II sequences. The color of each (z1, z2) coordinate in the scoring space represents the best score for the pairing of subsequences Seq1 (1:z1) and Seq2 (1:z2), where z1 is a position in Seq1 and z2 is a position in Seq2. The winning path represents pairing of positions in the optimal global alignment. The color of the last point (lower right) of the winning path represents the optimal global alignment score for the two sequences, and is the score output returned

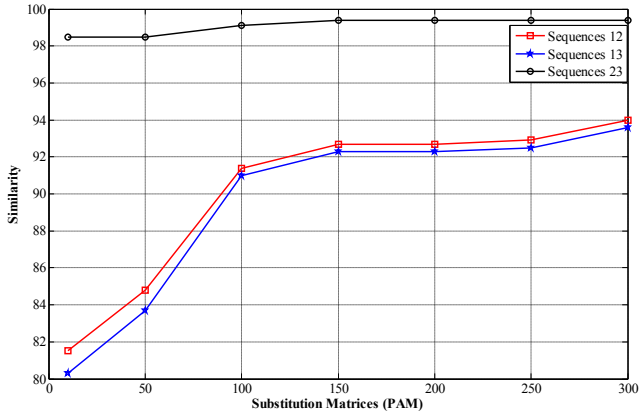

Figure 10b: Alignment similarities versus substitution matrices (PAM) for different pair wise sequences using N-W alignment.

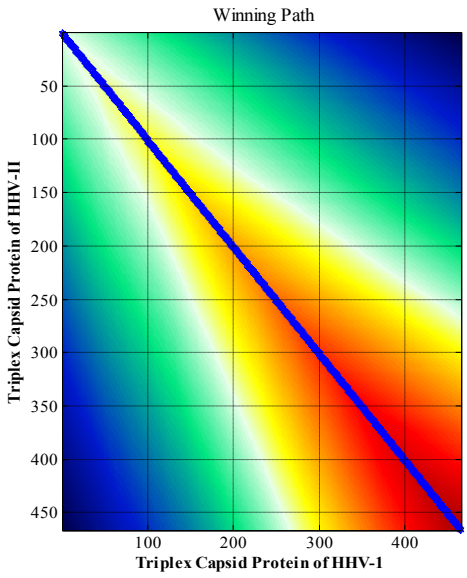

Figure 10c: N-W Algorithms back propagation for winning path of triplex capsid protein of HHV-I and HHV-II for BLOSUM 30 substitution matrix.

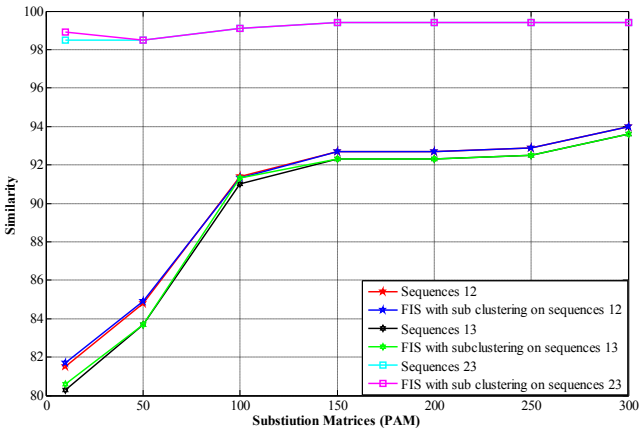

Figure 11a: Alignment similarities versus substitution matrices (PAM) for different pair wise sequences using N-W algorithms and FIS with sub clustering.

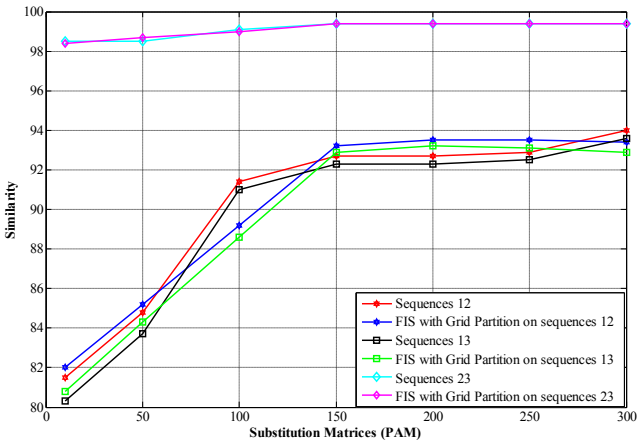

Figure 11b: Alignment similarities versus substitution matrices (PAM) for different pair wise sequences using N-W algorithms and FIS with grid partition.

by N-W align. The diagonal line shown in figure 10c represents the combination of several dots due to generous protein data of human herpes simplex virus. The diagonal line shows for N-W algorithm with BLOSUM 30 for sequences 1 and 2.

Figures 11a and 11b gives the comparison of the predicted alignment similarity of pairwise sequence using developed fuzzy model, and the data reported for computational analysis of triplex capsid protein of HHV-I and II for PAM substitution matrices. Figure 11a shows the plot between alignment similarity and PAM substitution matrices (FIS with sub clustering), and figure 11b shows the plot between alignment similarity and PAM substitution matrices (FIS with grid partition). The exponential increase in alignment similarity during the initial stage of the curve, with the increase of substitution matrices value of PAM, is due to correctly aligned pair at lowest penalty cost, and after that, the linear variation in similarity rate with an increase substitution matrix. There are slight increases in alignment similarity % at highest substitution matrices, instead of the highest cost of penalty. It confirms the PAM evaluation that higher substitution matrices are suitable for closely species. This trend of alignment similarity curve is also closely followed by the outcome of the designed fuzzy model. Out of the various outputs generated by the fuzzy model, only 1.5% data cross the experimental results, when FIS system with grid partition approaches. FIS with sub clustering have only 0.15% point deviation from the experimental analysis. With the total average error being 0.50%, the average accuracy of the model comes out to be 99.50%. In the present work, the total number of data points involved was sixty three points. Thus, it can be concluded that there is a close relation between the simulated results, and the practical results obtained at similar computation technique conditions for predicting alignment similarity, as shown in figures 10a and 10b.

Figures 12a and 12b gives the comparison of the predicted alignment similarity of HHV sequence, using developed fuzzy model and the data reported for computational analysis of triplex capsid protein of HHV-I and II for BLOSUM substitution matrices. Figure 12a shows the plot between alignment similarity and BLOSUM substitution matrices (FIS with grid partition), and Figure 12b shows the plot between alignment similarity and BLOSUM substitution matrices (FIS with sub clustering). The highest alignment similarity observed during the

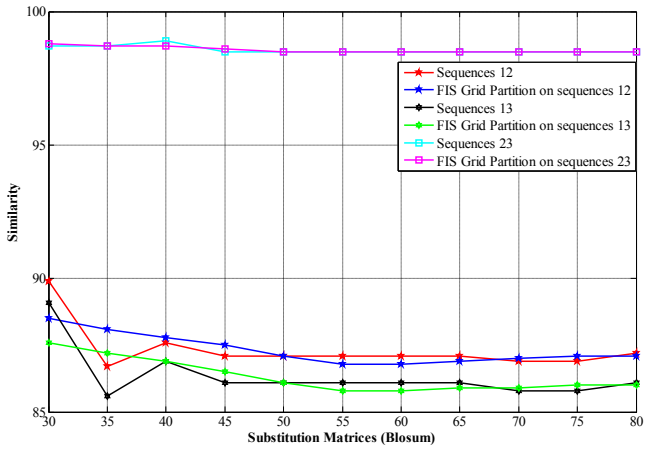

Figure 12a: Alignment similarities versus substitution matrices (BLOSUM) for different pair wise sequences using N-W algorithms and FIS with grid partition.

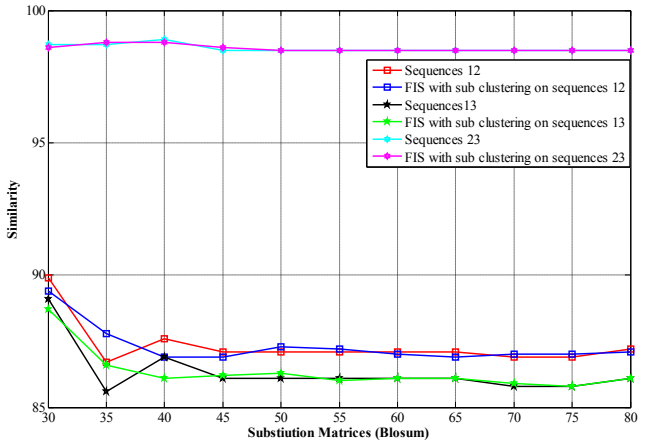

Figure 12b: Alignment similarities versus substitution matrices (BLOSUM) for different pair wise sequences using N-W algorithms and FIS with sub clustering.

initial stage of the curve due to correctly aligned pair at lowest penalty cost. Linear variation in similarity rate, with an increase substitution matrix found after the value of BLOSUM 40. It justifies the BLOSUM matrices have lower substitution matrices are suitable for less divergent species, as compared to PAM. This trend of rate of similarity curve is also closely followed by the outcome of the designed fuzzy model. With the total average error being 0.055%, the average accuracy of the model comes out to be 99.95%.

Conclusion

This paper is concerned with the application of fuzzy logic, to predict the alignment similarity using N-W algorithms for triplex capsid proteins of HHV-I and II. In this article, the MISO fuzzy model is developed using ANFIS, and validated with experimental results for given conditions. With more than 99% average accuracy, it has been found that results generated by the designed both fuzzy model are close to the computational results. After investigating the significance of the developed model, it has been concluded that the maximum differences between the sequences similarities is 1.50% on compared experimental data with FIS in grid partition. With this lowest deviation, accuracy of the sub clustering model can be used, to get quick answers for online intelligent control and optimization. The substitution matrix required in these sequences is BLOSUM 30 and PAM 300, and in its current state, the model is limited to number of sequences and N-W algorithm. This study supports that the fuzzy logic technique can be introduced as a viable alternative, to carry out sequence similarity analysis. Moreover, Fuzzy logic allowed the modeling and optimization to be treated simultaneously.

Author Contribution

Sohpal VK performed N-W simulation and comparative analysis of viral capsid proteins of HHV with ANFIS modeling, and drafted the manuscript. Apurba Dey and Amarpal Singh supervised field of work and revised the manuscript. Both authors read carefully and approved the manuscript.

References

1. Zhang M, Zhang W, Sicotte H, Yang P (2009) A new validity measure for a correlation-based fuzzy c-means clustering algorithm. Conf Proc IEEE Eng Med Biol Soc 2009: 3865-3868.
2. Garcia F, Lopez FJ, Cano C, Blanco A (2009) FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. BMC Bioinformatics 10: 224.
3. Mansoori EG, Zolghadri MJ, Katebi SD (2009) Protein superfamily classification using fuzzy rule-based classifier. IEEE Trans Nanobioscience 8: 92-99.
4. Bidargaddi NP, Chetty M, Kamruzzaman J (2008) Hidden Markov models incorporating fuzzy measures and integrals for protein sequence identification and alignment. Genomics Proteomics Bioinformatics 6: 98-110.
5. Espadaler J, Eswar N, Querol E, Avilés FX, Sali A, et al. (2008) Prediction of enzyme function by combining sequence similarity and protein interactions. BMC Bioinformatics 9: 249.
6. Gómez A, Cedano J, Espadaler J, Hermoso A, Piñol J, et al. (2008) Prediction of protein function improving sequence remote alignment search by a fuzzy logic algorithm. Protein J 27: 130-139.
7. Collyda C, Diplaris S, Mitkas P, Maglaveras N, Pappas C (2007) Enhancing the quality of phylogenetic analysis using fuzzy hidden Markov model alignments. Stud Health Technol Inform 129: 1245-1249.
8. Bryliński M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, et al. (2007) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3: e94.
9. Samsonova EV, Krause P, Bäck T, IJzerman AP (2007) Characteristic amino acid combinations in olfactory G protein-coupled receptors. Proteins 67: 154-166.
10. Huang WL, Chen HM, Hwang SF, Ho SY (2007) Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. Biosystems 90: 405-413.
11. Collyda C, Diplaris S, Mitkas PA, Maglaveras N, Pappas C (2006) Fuzzy Hidden Markov Models: a new approach in multiple sequence alignment. Stud Health Technol Inform 124: 99-104.
12. Popescu M, Keller JM, Mitchell JA (2006) Fuzzy measures on the Gene Ontology for gene product similarity. IEEE/ACM Trans Comput Biol Bioinform 3: 263-274.
13. Liang S (2003) cWINNOWER algorithm for finding fuzzy DNA motifs. Proc IEEE Comput Soc Bioinform Conf 2: 260-265.
14. Boeva V, Regnier M, Papatsenko D, Makeev V (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. Bioinformatics 22: 676-684.
15. Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577-581.
16. Atchley WR, Fernandes AD (2005) Sequence signatures and the probabilistic identification of proteins in the Myc-Max-Mad network. Proc Natl Acad Sci U S A 102: 6401-6406.
17. Jacob E, Sasikumar R, Nair KN (2005) A fuzzy guided genetic algorithm for operon prediction. Bioinformatics 21: 1403-1407.
18. Liang S, Samanta MP, Biegel BA (2004) cWINNOWER algorithm for finding fuzzy dna motifs. J Bioinform Comput Biol 2: 47-60.
19. Brodie R, Smith AJ, Roper RL, Tcherepanov V, Upton C (2004) Base-By-Base: single nucleotide-level analysis of whole viral genome alignments. BMC Bioinformatics 5: 96.
20. Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21-28.
21. Blankenbecler R, Ohlsson M, Peterson C, Ringner M (2003) Matching protein structures with fuzzy alignments. Proc Natl Acad Sci U S A 100: 11936-11940.
22. Heger A, Holm L (2003) Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. Bioinformatics 19: i130-i137.
23. Torres A, Nieto JJ (2003) The fuzzy polynucleotide space: basic properties. Bioinformatics 19: 587-592.
24. Schlosshauer M, Ohlsson M (2002) A novel approach to local reliability of sequence alignments. Bioinformatics 18: 847-854.

This article was originally published in a special issue, **Special Issue on Genome Annotation** handled by Editor: Dr. Jorge Cancela, University of Florida, USA