IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. III (May - June 2017), PP 01-07
www.iosrjournals.org

Hybrid Fuzzy Algorithm for Protein Clustering Using Secondary Structure Parameters

Murugananthi. C[1], Ramyachitra. D[2]
[1] *(Department of Computer Technology, Vellalar College for Women, India)*
[2] *(Department of Computer Science, Bharathiar University, India)*

Abstract : *Clustering of proteins is important in the field of bioinformatics. Clustering of proteins is used for analyzing the proteins to determine their functions and structure. The number of partitioning techniques, hierarchical methods and graph-based methods are available for clustering protein sequences. In this paper, we propose a hybrid fuzzy technique for clustering proteins based on its secondary structure elements. The algorithm works in two stages. In the first stage, initial number of clusters is determined using k-nearest neighbor distances. The second stage comprises membership calculation and cluster construction. The performance of the hybrid fuzzy clustering was evaluated by comparison with other existing methods on four data sets. Experimental results show that proposed approach performs better in terms of validity indices and execution time as well.*

Keywords: *Protein sequence, Secondary structure, Fuzzy clustering, K-nearest neighbor*

I. Introduction

Bioinformatics is the use of computer technology for managing biological data and solving complex biological problems. Genomics and proteomics projects are currently producing massive amount of new gene and protein sequences. Proteins are important organic molecules composed of amino acids arranged in a linear chain and held together by peptide bonds. Proteins are an essential part of organisms, perform all necessary functions and participate in all processes within the body.

Clustering is partitioning of objects into different groups, so that the objects in each group share some common features [1]. Data points within group are similar and different between groups [2]. Computational tools are needed to manage and analyze bioinformatics datasets. Clustering techniques have been used for managing bioinformatics datasets, including gene and protein sequence analysis [3]. Clustering increases efficiency of data analysis and used to identify the relationships between biological objects such as protein sequences and structures [4]. Determining the relationship between protein sequence and structure is important in structure and function prediction.

Protein classification and clustering have been used in many bioinformatics applications as sequence and structure analysis, drug design, molecular dynamics, biophysics, and computational chemistry. Many clustering algorithms and methods are available for clustering proteins. Most of the previous works uses the hierarchical, partitioning, graph based techniques, and clustering the proteins by sequences. Large scale hierarchical clustering for protein sequences was used by Antje Krause et. al., [5]. Hybrid hierarchical and k-median based clustering technique was proposed by Vijaya et. al., [6]. Hierarchical clustering mostly based on the local decision making method that joins the two closest proteins or clusters without considering the data as a whole.

Sondes Fayech et. al., [7] developed partitioning clustering algorithms Pro-Kmeans, Pro-LEADER, Pro-CLARA, Pro-CLARANS using Smith-Waterman algorithm for protein sequence clustering. These methods fail to expose nonlinear relationships between proteins and thereby fail to correctly represent a dataset with non-linear structure. The graph based methods existed on clustering protein sequences are as follows: Graph based parallel clustering [8]; spectral clustering [9] and markov clusterin [10]. In graph based clustering, number of clusters not specified in advance, but can change cluster granularity with parameters and cannot find overlapping clusters. It is computationally expensive and not suitable for clusters with large diameter.

Sung Hee Park et al., [11] used k-means clustering algorithm for clustering protein secondary structures and distances between proteins are calculated using dynamic programming. But this method needs a number of clusters which are to be used in an initial state. In this paper, we present hybrid fuzzy clustering algorithm for clustering proteins using its secondary structure elements. It integrates the concepts of fuzzy c-means and k-nearest neighbor algorithms. We use secondary structure information for clustering proteins in a simple manner instead of using normal protein sequence. Fuzzy clustering provides the chance to deal with proteins belonging to more than one group at a time. Normally it is sensitive to noise and outliers, but this is handled by integration of k-nearest neighbor algorithm. It automatically determines the cluster centers (number of cluster) and outliers.

In this paper, method of extracting secondary structure and distance matrix computation are given in Section 2.1 and 2.2 respectively. Clustering process of hybrid fuzzy clustering is presented in Section 2.3. The validity indices are described in Section 2.4. Experimental results and the dataset used for performance evaluation are given in Section 3.

II. Methods and Materials

The proposed work consists of the following three parts: Secondary structure extraction, distance matrix computation, and clustering of protein secondary structures. The working flow is illustrated in Fig. 1.

2.1. Secondary structure extraction

The A protein sequence determines its structure and the structure determines functions. Protein secondary structure plays a significant role in modeling and analyzing protein structures because it represents the amino acids sequence into regular structures [12].

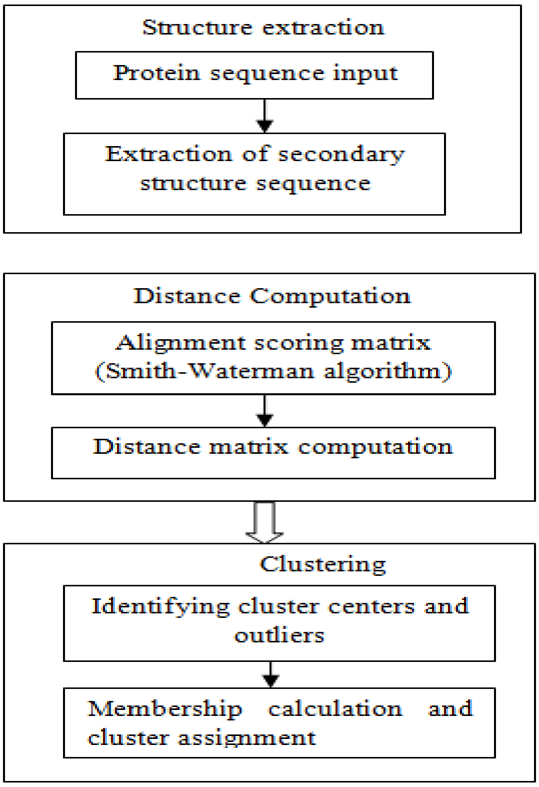
Structure extraction
Protein sequence input
Extraction of secondary structure sequence
Distance Computation
Alignment scoring matrix (Smith-Waterman algorithm)
Distance matrix computation
Clustering
Identifying cluster centers and outliers
Membership calculation and cluster assignment

Fig. 1. The working flow of proposed work

Protein tertiary structure is formed by packing secondary structure elements into one or several units [4]. Secondary structure prediction is used in genome analysis, protein function prediction. It is used to identify domains, classify proteins and recognize functional motifs [13]. Since secondary structure elements are general representation of protein structure, it is used to cluster a set of proteins at the abstraction level [11]. The amount of data required to abstract protein structure is reduced by representing it with secondary structure element sequence. The basic secondary structure elements are α−helix (H), β−sheet (E), turn (T) and coil (C). Given a protein sequence with amino acids, the secondary structure extraction is to predict whether each amino acid is in a α−helix, a β−sheet, a turn, or a coil. In this work, we input the protein sequence and extract the secondary structure element sequence using GOR V method [14].

2.2. Distance matrix computation

We used Smith-Waterman local alignment algorithm [15] for calculating alignment score. This method compares all sequences with each other and computes the alignment score. The distance matrix can be computed after finding the alignment score matrix. Distance between two protein sequences can be derived from its similarity score [16]. For a given set of protein sequences, distance between any two sequences is calculated as

$$D(A,B) = -\ln S_n(A,B) \quad (1)$$

where *A* and *B* are protein sequences, $D(A,B)$ is the distance between *A* and *B*, *ln* is natural logarithm, $S_n(A,B)$ is the normalized similarity score between *A* and *B*. Here $0 \leq S_n(A,B) \leq 1$ for any protein sequences *A* and *B*, and $S_n(A,B) = 1$ if sequences *A* and *B* are same. The normalized similarity score is obtained by using the below formula

$$S_n(A,B) \cong \frac{S(A,B)}{L.Q} \tag{2}$$

where $S(A,B)$ is the similarity score of *A* and *B*, *L* denote the length of the local alignment of *A* and *B*, and *Q* is normalization parameter. The normalization parameter *Q* is computed as a value when two residues are matched with each other. This value depends on the distribution of residues in the local alignment of *A* and *B*, and the scoring matrix between residues.

2.3. Clustering

In hard clustering or crisp clustering, each object belongs to exactly one cluster. In soft clustering (fuzzy clustering), objects can belong to more than one cluster [17]. Fuzzy clustering is a process of assigning membership and then using them to assign objects to one or more clusters. One of the widely used fuzzy clustering algorithms is Fuzzy C-Means (FCM) Algorithm [18]. In this algorithm membership is assigned to each object corresponding to each cluster based on the distance between an object and cluster center.

This work proposes a hybrid fuzzy clustering algorithm for protein sequences that combines simplicity with good performance and robustness. Our proposed method combines the concepts of k-nearest neighbor and fuzzy c-means algorithms. The proposed Hybrid Fuzzy Clustering (HFC) consists of two main parts. This method takes number of nearest neighbors k and threshold value α for outliers as initial parameters.

2.3.1. Identifying cluster centers and outliers

This part goes through the following steps

a) Construct neighborhood graph using distance matrix to connect each protein to its *k-NN*.

b) Calculate average of k-nearest neighbors (*AKNN*) distance for each protein.

$$AKNN(x) = \sum_{y \epsilon KNN(x)} d_{xy} \tag{3}$$

where protein y is nearest neighbor protein of x and d_{xy} is the distance between x and y.

c) Proteins with minimum *AKNN* distance among their neighbors are identified as cluster centers (C_j). Proteins with maximum *AKNN* and larger than the threshold value α are identified as outliers (O_i).

2.3.2. Membership calculation and cluster construction

This part goes through the following steps

a) Calculate cluster membership u_{ij} for each protein by the below formula except the cluster centers and outliers.

$$u_{ij} = 1/\sum_{l=1}^{c}\left(\frac{d_{ij}^2}{d_{il}^2}\right)^{\frac{1}{m-1}} \tag{4}$$

where $1 \leq i \leq N$ and $1 \leq j \leq C$. *N* is the number of proteins and *C* is the number of clusters, u_{ij} is the membership of *i*th protein in the *j*th cluster. d_{ij} is the distance between the *i*th protein and *j*th cluster center. *m* is the fuzzification parameter and it should be more than one. If *m*=1, then the problem is a crisp clustering. $m \in [1,\infty]$ and usually *m* is set to 2 (Hathaway & Bezdek, 2001).

b) Assign each protein to the cluster center in which it has highest membership. The cluster center V_j is updated as follows:

$$V_j = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{for } j = 1,2,\ldots, C \tag{5}$$

where V_j is the *j*th cluster center and h is number of proteins in *j*th cluster.

c) Repeat these two steps until the cluster center remains the same.

Like FCM, the HFC also minimize the objective function [18] in Eq. (6) and integrates the properties of fuzzy c-means algorithm. Summation of membership of each object should be equal to one in Eq. (7).

$$J(U,V) = \sum_{j=1}^{c}\sum_{i=1}^{N}(u_{ij})^m\,(d_{ij})^2 \tag{6}$$

$$\sum_{j=1}^{c} u_{ij} = 1 \tag{7}$$

2.4. Validity indices

To assess the performance of HFC and compare it with the other existing algorithms, we used two validity indices silhouette index and partition index.

2.4.1. Silhouette index

The silhouette index [19] is a cluster validity index used to assess the quality of any clustering. The silhouette index of a protein defines its closeness to its own cluster relative to its closeness to other clusters. The silhouette width *s(x)* of the protein *x* is defined as

$$S(x) = \frac{b(x) - a(x)}{\max\,[b(x), a(x)]} \qquad (8)$$

where $a(x)$ is the average distance between protein x and all other proteins in its cluster and $b(x)$ is the minimum of the average distances between protein x and the proteins in the other clusters. The silhouette index $s(c)$ of cluster c is defined as the average silhouette width of its all proteins. Finally, silhouette index of the whole clustering is the average silhouette width of all clusters. It reflects the compactness and separation of clusters. The value of the silhouette index varies from -1 to 1 and higher values indicate a better clustering result.

2.4.2. Partition index

The partition index $p(c)$ [20] is defined as the ratio between the overall within-cluster variability and the overall between-cluster distance. Based on this validation index, a good data clustering results in low intra cluster variation and high inter cluster variation. To find the overall within-cluster variation, the variation within each cluster is calculated as the average distance between each pair of proteins in the cluster and then averaged for all clusters. The between-cluster variation is obtained by averaging the distance between each pair of clusters. Each single between-cluster distance is calculated by averaging the distance between each pair of protein from the two clusters. The lower partition index value indicates the better clustering result.

III. Experimental Results

3.1. Protein sequence datasets

The experiment was conducted on four different protein data sets: Dengue virus proteins, Human Leukocyte Antigen (HLA) proteins, Globins proteins and Saccharomyces cerevisiae (Yeast) proteins. Dengue virus protein sequences are extracted from Protein Data Bank [21] and named as DS1. Sequences of Globins protein family and Human Leukocyte Antigen (HLA) proteins were collected from European Bioinformatics Institute (EMBL-EBI) database [22] and named as DS2, DS3 respectively. Yeast proteins are collected from Saccharomyces Genome database [23] and named as DS4.

IV. Results And Discussion

The performance of the proposed HFC was evaluated by comparison with other existing algorithms, namely k-means from partitioning method, hierarchical clustering and self-organization map. The experiments were conducted on Intel pentium-5 processor with 2GB RAM. Secondary structure sequences of datasets given in Section 3.1. are extracted using GOR V method [14]. After that, alignment scoring matrix was obtained by Smith-Waterman algorithm [15]. Then, the normalized similarity scores are calculated by Eq. (2). Distance matrix of protein sequences are calculated using similarity scores. After completing all these processes, conventional and our proposed methods are initialized, and run with the secondary structure element sequences and above predicted distance matrix.

Our proposed HFC does not suffer from the random initialization. It automatically determines the number of clusters with the help of number of nearest neighbors. The execution of HFC on protein datasets carried out with different values of parameter *k*, and it partitioned the sequences into various number of clusters based on the *k* value. The value of *k* increases, then the number of clusters decreases and vice versa. The values of parameter *k* and the number of cluster are given in Table 1.

Table 1. K values and the number of clusters for four datasets

Datasets	Number of clusters				
	K=8	K=7	K=6	K =5	K=4
Dengue virus proteins	9	13	16	20	24
HLA proteins	11	15	22	30	34
Globins proteins	10	16	20	25	31
Yeast proteins	12	21	28	33	40

Table 2. Average validity indices of clustering methods on four data sets

Algorithms	S(c)	P(c)
Dengue virus proteins	0.2653	0.4557
HLA proteins	0.4224	0.3481
Globins proteins	0.3518	0.3359
Yeast proteins	0.5067	0.2754

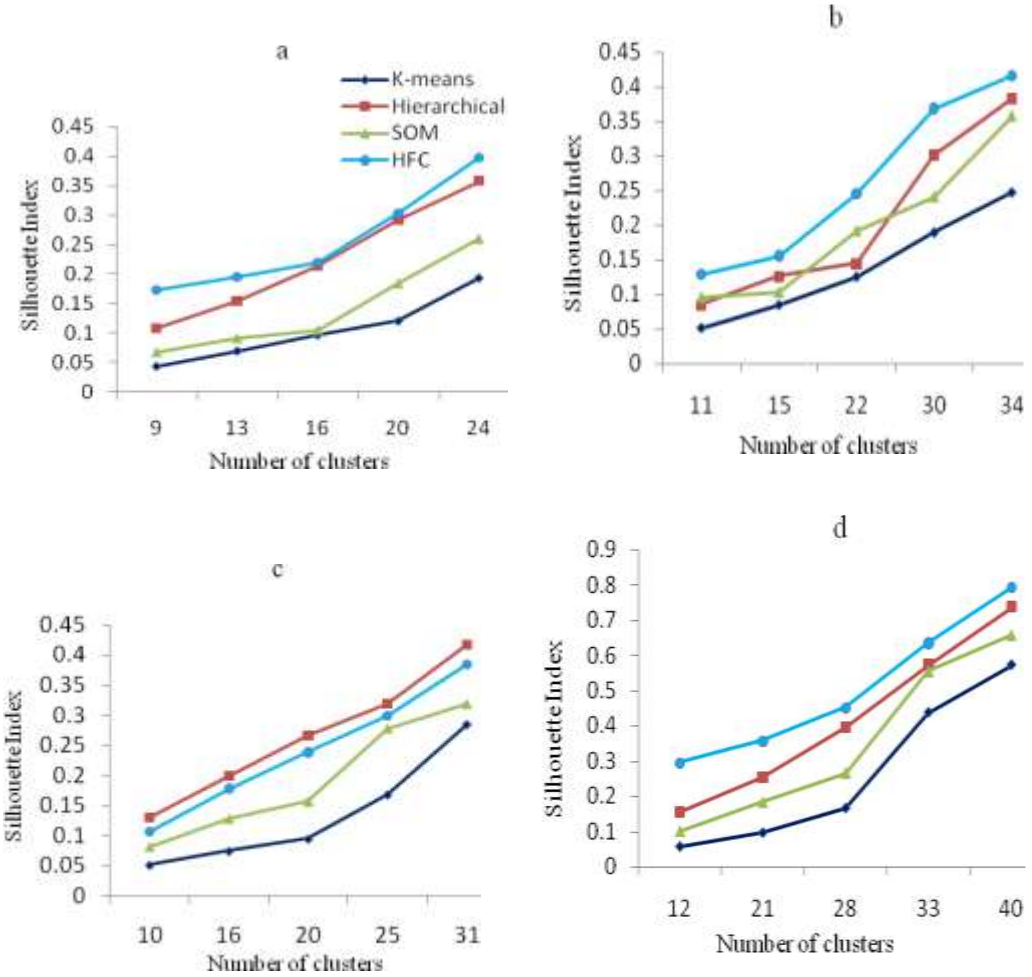

Fig. 2. Clustering validation and comparison by silhouette index a) Silhouette index on dengue virus dataset b) Silhouette index on HLA dataset c) Silhouette index on globins dataset d) Silhouette index on yeast dataset.

We calculate validity indices given in Section 2.4 for different number of clusters on four datasets. Fig. 2 shows silhouette index on four datasets. Fig. 3 shows partition index on four datasets. Table 2 gives the average values of validity indices on four datasets for each clustering methods. From Table 2 it is known that, the silhouette index of HFC is 17%, 31% and 48% higher than SOM, hierarchical and k-means clustering respectively. The partition index of HFC is 18%, 21% and 40% lower than SOM, hierarchical and k-means clustering respectively. According to both of the validity index analysis, HFC is the best algorithm in three out of four datasets. Fig. 4 shows the execution time of clustering methods on four datasets. Execution times of HFC also lower than three existing clustering methods. From the results, it is inferred that proposed approach HFC performs better in terms of validity indices and execution time as well.

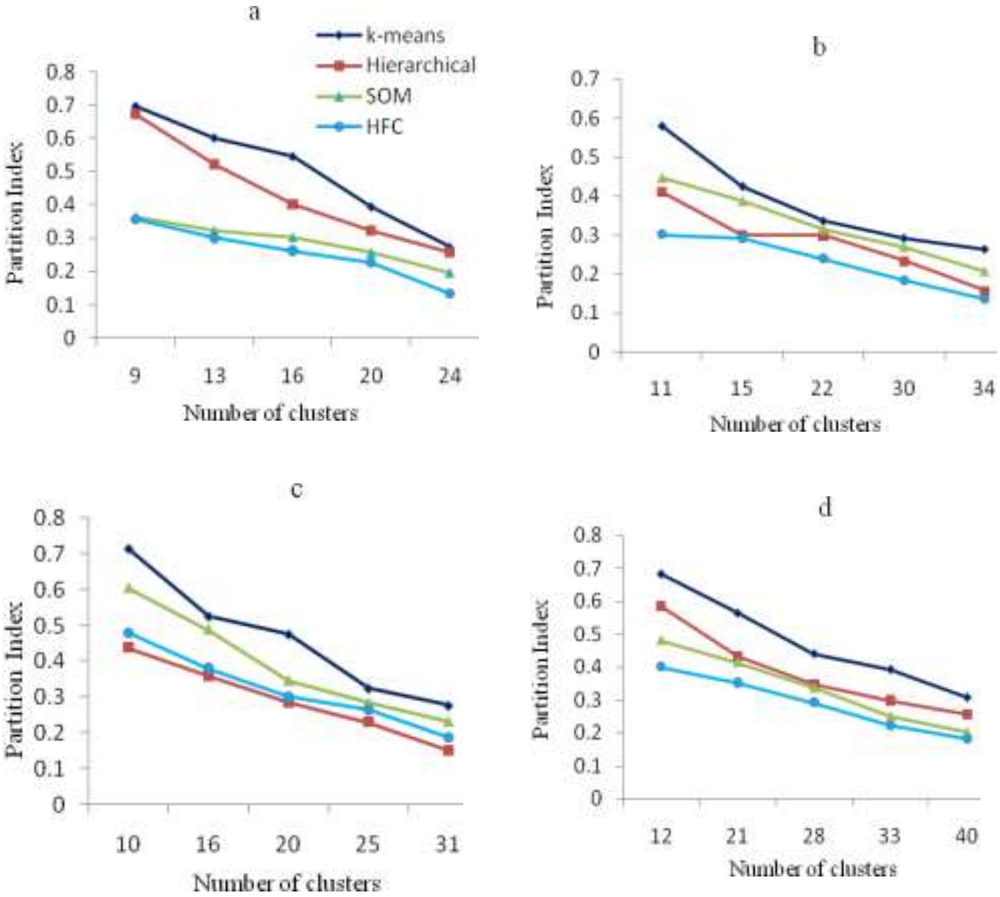

Fig. 3. Clustering validation and comparison by partition index a) Partition index on dengue virus dataset b) Partition index on HLA dataset c) Partition index on globins dataset d) Partition index on yeast dataset.

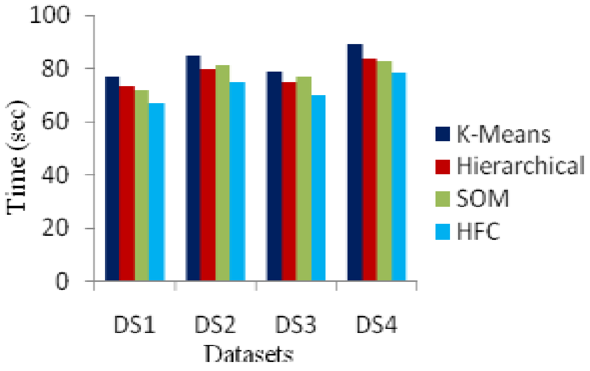

Fig. 4. Execution time of algorithms on four datasets

V. Conclusion

In this paper, hybrid fuzzy clustering algorithm has been proposed for clustering proteins using its secondary structure element sequences. In this work, structure information also taken into account for clustering proteins effectively. Since secondary structure elements are general representation of protein structure, it is used to cluster a set of proteins at the abstraction level. The amount of data required to abstract protein structure is reduced by representing it with secondary structure element sequence. Results show that HFC outperforms the other existing algorithms on four real datasets in terms of silhouette index, partition index and run time. The experiments conducted on four data sets shows that HFC can be used effectively for clustering protein data sets.

References

[1] Yonghui Chen, Kevin D Reilly, Alan P Sprague, and Zhijie Guan, SEQOPTICS: a protein sequence clustering system, BMC Bioinformatics, 7(Suppl 4), S10, 2006.

[2] Efendi Nasibov, and Cagin Kandemir-Cavas, OWA-based linkage method in hierarchical clustering, Application on phylogenetic trees, Expert Systems with Applications, 38, 2011, 12684–12690.

[3] Chan, Z. S. H., Collins, L., and Kasabov, N., An efficient greedy k-means algorithm for global gene trajectory clustering, Expert Systems with applications, 30(1), 2006, 137–141.

[4] Piotr Lukasiak, Jacek Blazewicz, and Maciej Milostan, Some Operations Research Methods for Analyzing Protein Sequences and Structures, Annals of Operations Research, 175, 2010 9-35.

[5] Antje Krause, Jens Stoye, and Martin Vingron, Large scale hierarchical clustering of protein sequences, BMC Bioinformatics, 6:15, 2005.

[6] Vijaya, P.A., Narasimha Murty, M., and Subramanian, D.K., Efficient bottom-up hybrid hierarchical clustering techniques for protein sequence classification, Pattern Recognition, 39, 2006, 2344–2355.

[7] Sondes Fayech, Nadia Essoussi, and Mohamed Limam, Partitioning clustering algorithms for protein sequence data sets, BioData Mining, 2:3, 2009.

[8] Assayony, M., and Rashid, N.A., Design of a parallel graph-based protein sequence clustering algorithm, IEEE Symposium on Information Technology, 3, 2008, 1-8.

[9] Paccanaro, A., Chennubhotla, C., Casbon, J.A., and Saqi, M.A.S., Spectral clustering of protein sequences, Proceedings of IEEE International Joint conference on Neural Networks, 4, 2003, 3083-3088.

[10] Lehel Medves, Laszlo Szilagyi, and Sandor M. Szilagyi, A Modified Markov Clustering Approach for Protein Sequence Clustering. Lecture Notes in Bioinformatics, Springer-Verlag Berlin Heidelberg, 5265, 2008, 110–120.

[11] Sung Hee Park, Chan Yong Park, Dae Hee Kim, Seon Hee Park, and Jeong Seop Sim, Protein Structure Abstraction and Automatic Clustering Using Secondary Structure Element Sequences, Proceedings of the International Conference on Computational Science and its Applications, Springer-Verlag Berlin Heidelberg, 3481, 2005, 1284-1292.

[12] Hsin-Nan Lin, Ting-Yi Sung, Shinn-Ying Ho, and Wen-Lian Hsu, Improving protein secondary structure prediction based on short subsequences with local structure similarity, BMC Genomics, 11(Suppl 4):S4, 2010.

[13] Haitao Cheng, Taner Z. Sen, Andrzej Kloczkowski, Dimitris Margaritis, and Robert L. Jernigan, Prediction of protein secondary structure by mining structural fragment database. Polymer, 46, 2005, 4314–432.

[14] Kloczkowski, A., Ting, K.L., Jernigan, R.L., and Garnier, J., Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence, Proteins, 49(2), 2002, 154–166.

[15] Smith, T.F., and Waterman, M.S., Identification of common molecular subsequences. Journal of Molecular Biology, 147, 1981, 195-197.

[16] Matsuda, H., Ishihara, T., and A. Hashimoto, Classifying molecular sequences using a linkage graph with their pairwise similarities, Theoretical Computer Science, 210, 1999, 305–325.

[17] Zadeh, L., Fuzzy sets, Information and Control, 8, 1965, 338–353.

[18] Bezdek, J.C., Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.

[19] Rousseeuw, P.J., Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 1987, 53-65.

[20] Limin Fu, and Enzo Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data, BMC Bioinformatics, 8:3, 2007.

[21] Protein Data Bank, 2014. (http://www.rcsb.org).

[22] The European Bioinformatics Institute (EMBL-EBI) database, 2014. (http://srs.ebi.ac.uk).

[23] Saccharomyces Genome database, 2014. (http://www.yeastgenome.org).