Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 814 – 820

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

Extraction of Motif Patterns from Protein Sequence Using Rough- K-Means Algorithm

E. Elayaraja [a] ,K. Thangavel[a] , B. Ramya[a] , M.Chitralegha[a], a*

[a]*Department of Computer Science,Periyar University,Salem-636 011*

Abstract

Bioinformatics is the application of computer technology to the management of biological information. In Bioinformatics, Motif finding is one of the most popular problems, which has many applications. It is the process of locating the meaningful patterns in the sequence of Deoxyribo Nucleic Acid (DNA), Ribo Nucleic Acid (RNA) or Proteins. Motifs vary in lengths, positions, redundancy, orientation and bases. Finding these short sequences (motifs or signals) is a fundamental problem in molecular biology and computer science with important applications such as knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology. In this work, the clustering system is used to predict local protein sequence Motifs. Since clustering algorithms can provide an automatic, unsupervised discovery process for sequence motifs, the K-Means clustering algorithm and Rough-K-means algorithm proposed are chosen as the motif discovery method for this study and the results are compared. The structural similarity of the clusters discovered by the proposed approach is studied to analyze how the recurring patterns correlate with its structure. Also, some biochemical references are included in our evaluation.

Keywords: Clustering; Motif; Protein Sequence; HSSP;DSSP; HSSP-BLOSUM62.

1. Introduction

Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich, but lacks a comprehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel Knowledge Discovery in Databases (KDD) methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.

Proteins are large, complex molecules that play many critical roles in the body. They do most of the work in cells and are required for the structure, function, and regulation of the body's tissues and organs. Proteins are made up of hundreds or thousands of smaller units called amino acids, which are attached to one another in long chains. There are 20 different types of amino acids that can be combined to make a protein [6]. In a group of related proteins, there exists some highly conserved region across a subset of proteins that share the same function. Such conserved sequence patterns are denoted as sequence motifs. These patterns are vital for Understanding gene function, human disease, and may serve as therapeutic drug target.

The popular databases for sequence motifs are PROSITE [1], PRINTS [2], BLOCKS [3], SBASE. The commonly used tools for protein sequence motif discovery include MEME, Gibbs Sampling, and Block Maker.

In this paper Protein sequences are converted into sliding sequence segments by applying sliding window technique on HSSP file [4]. Rough-K-Means and K-means clustering are adopted for extracting PROTEIN SEQUENCE MOTIF [5]. These sliding sequence segments are classified into different groups with the clustering algorithms. The structural similarity of these groups is evaluated using the secondary structure information obtained from the DSSP file. The recurrent groups with high

* E. Elayaraja..
E-mail address: raja_e2001@yahoo.co.in.

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.932

structural similarity will become the candidate to generate sequence motifs representing common structure. Identified sequence motifs are represented by frequency profiles.

This paper has been organized into nine sections. In Section 2, the dataset of our work is introduced. In Section 3, and 4, the Homology-Derived Secondary Structure of Proteins (HSSP) and Dictionary of Secondary Structure of Proteins (DSSP) are explained**.** In Section 5, the HSSP-Blosum62 measure used to evaluate the quality of the clusters is presented. In Section 6, and 7, the K-Means and Rough-K-Means clustering algorithms are discussed. Section 8 gives the experimental results. The representation of motif patterns is presented in section 9. Finally section 10 summarizes the paper with direction for further research.

2. Dataset

Since the major purpose of this work is to obtain protein sequence motif information across protein family boundaries, the dataset of our work is supposed to collect all known protein sequences. However, without a systematic approach, it is very difficult to extract useful knowledge from an extremely large volume of data.

The dataset used in this work includes 2710 protein sequences obtained from Protein Sequence Culling Server (PISCES) [12]. No sequence in this database shares more than 25% sequence identity. The frequency profile from the HSSP is constructed based on the alignment of each protein sequence from the Protein Data Bank (PDB) where 300 sequences are considered homologous in the sequence database.

3. Homology-Derived Secondary Structure of Proteins

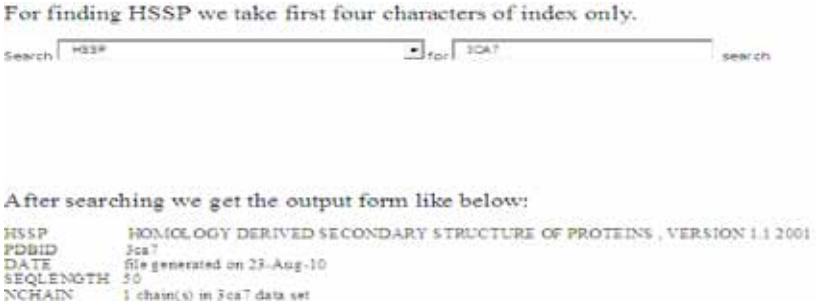
HSSP (Homology-derived Secondary Structure of Proteins) is a derived database merging information from three-dimensional structures and one-dimensional sequences of proteins. One can therefore group sequence-similar proteins into families of structural homologues.

Description of HSSP files. One HSSP file contains a structural protein family: The file is divided into four blocks, HEADERS, PROTEINS, ALIGNMENTS and SEQUENCE PROFILE.

3.1 Sequence Profile Block

Relative frequency for each of the 20 amino acid residue in a given sequence position, from counting the residue at that position in each of the aligned sequences including the test sequence. A value of 100 means in that position only one type of amino acid is found.

For finding HSSP we take first four characters of the index only.



3.2 Representation of Data Segment

The sliding windows with ten successive residues are generated from protein sequences. Each window represents one sequence segment of ten continuous positions. Fig. 1 show how we apply the sliding window technique on the HSSP file, each window corresponds to a sequence segment, which is represented by a 10 x 20 matrix [7]. For the frequency profiles representation for sequence segments, each position of the matrix represents the frequency for a specified amino acid residue in a sequence position for the multiple sequence alignment.

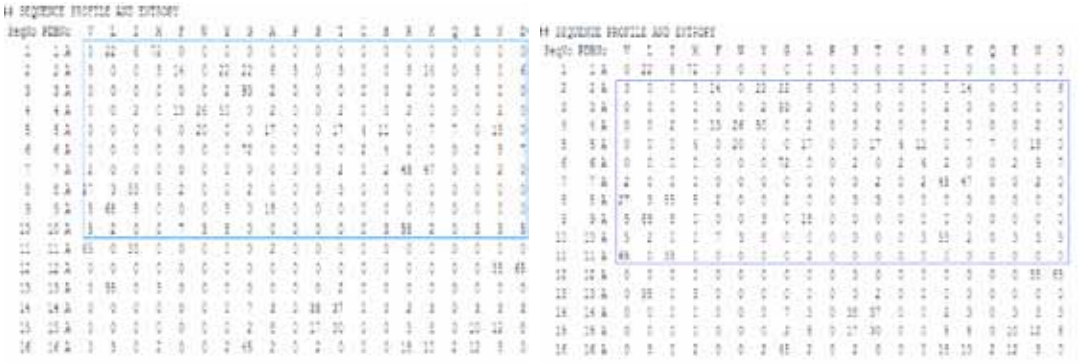


Fig. 1. Sliding Window techniques with a window size of 10 applied on 1b25 HSSP file

Thus by applying the sliding window technique we generate n number of sequence segments (10 X 20 matrices).

4. Dictionary of Secondary Structure of Proteins

We also obtained secondary structure from DSSP, which is a database of secondary structure assignments for all protein entries in the Protein Data Bank, for evaluation purposes. The Dictionary of Secondary Structure of Proteins is commonly used to describe the protein secondary structure with single letter codes. The secondary structure is assigned based on hydrogen bonding patterns as those initially proposed by Pauling et al. in 1951 (before any protein structure had ever been experimentally determined). There are eight types of secondary structure that DSSP defines:

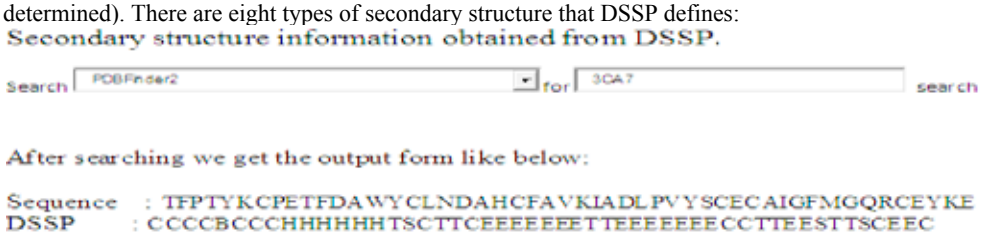


Fig. 2. Secondary Structure Information Obtained From DSSP

5. HSSP-BLOSUM62 Measure

BLOSUM62 (Fig. 3.) is a scoring matrix based on known alignments of diverse Sequences.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Fig. 3. BLOSUM62 Matrix

By using this matrix, we may access the consistency of the amino acids appearing in the same position of the motif information generated by our method. Because different amino acids appearing in the same position should be close to each other, the corresponding value in the BLOSUM62 matrix will give a positive value. Hence, the measure is defined as the following

If k = 0: HSSP-BLOSUM62 measure = 0
Else If k = 1:
If $HSSP_i > 10\%$: HSSP-BLOSUM62 measure = $BLOSUM62_{ii}$
If $8\% \le HSSP_i < 10\%$: HSSP-BLOSUM62 measure = $\frac{1}{2} BLOSUM62_{ii}$
Else: HSSP-BLOSUM62 measure = $\frac{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} HSSP_i \cdot HSSP_j \cdot BLOSUM\ 62_{ij}}{\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} HSSP_i \cdot HSSP_j}$

6. K-Means Clustering Algorithm

K-Means algorithm [11] is a prototype-based, partitional clustering technique that attempts to find user-specified number of clusters, which are represented by their centroids. In most of the cases Euclidean distance measure is chosen as a common measure. A set of *n* objects $x_i, i = 1, 2, \ldots, n$, are to be partitioned into *K* groups. The cost function, based on the Euclidean distance between a vector *x* in group *j* and the corresponding cluster centroid c_j, can be defined by

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i - C_j \right\|^2$$

6.1 City Block Distance

Euclidean distance is probably the most commonly used distance function between feature vectors; it is not always the best metric. The City block metric is used for calculating the difference between a sequence segment and the centroid of a given sequence cluster [8]. Each centroid of a sequence cluster is represented by 10 x 20 matrixes.

The following formula is used to calculate the distance between two sequence segments [9].

$$\text{Dissimilarity} = \sum_{i=1}^{L} \sum_{j=1}^{N} \left| F_k(i,j) - F_c(i,j) \right|$$

Where L is the window size and N is 20.

$F_k(i, j)$ is the value of the matrix at row i and column j used to represent the sequence segment k. $F_c(i,j)$ is the value of the matrix at row i and column j used to represent the centroid of a given sequence cluster.

7. Rough K-means Algorithm

In rough clustering each cluster has two approximations, a lower and an upper approximation. The lower approximation is a subset of the upper approximation. The members of the lower approximation belong certainly to the cluster; therefore they cannot belong to any other cluster. The data objects in an upper approximation may belong to the cluster. Since their membership is uncertain they must be a member of an upper approximation of at least another cluster.

7.1 Rough properties of the cluster algorithm

- Property 1: a data object can be a member of one lower approximation at most.
- Property 2: a data object that is a member of the lower approximation of a cluster is also member of the upper approximation of the same cluster.
- Property 3: a data object that does not belong to any lower approximation is member of at least two upper approximations.

This algorithm can also be interpreted as two layer interval clustering approach with lower and upper approximation.
The rough K-means algorithm [9, 10] can be stated as follows:

1. Select initial clusters of n objects into k clusters.
2. Assign each object to the Lower bound (L(x)) or upper bound (U(x)) of cluster/ clusters respectively as:
   For each object v, let d (v,xi) be the distance between itself and the centroid of cluster xi. The difference between d (v,xi) / d(v,xj), 1≤ i, j ≤ k is used to determine the membership of v as follows:
   - If d (v,xi) / d(v,xj) ≤ thershold, then v ∈U(xi) & v ∈ U(xj). Furthermore, v will not be a part of any lower bound.
   - Otherwise, v∈L(xi),such that d(v,xi) is the minimum for 1≤ i ≤ k. In addition, v∈U(xi).
3. For each cluster xi re-compute center according to the following equations the weighted combination of the data points in its lower_bound and upper_bound.

$$x_i = \begin{cases} w_{lower} \times \dfrac{\sum_{v \in L(x)} v_j}{|L(x)|} + w_{upper} \times \dfrac{\sum_{v \in U(x)-L(x)} v_j}{|U(x)-L(x)|} & \text{if } |U(x)-L(x) \neq \phi| \\ w_{lower} \times \dfrac{\sum_{v \in L(x)} v_j}{|L(x)|} & \text{otherwise} \end{cases}$$

Where 1≤ j ≤ k. The parameters w_{lower} and w_{upper} correspond to the relative importance of lower and upper bounds. If convergence criterion is met, i.e. cluster centers are same to those in previous iteration, then stop; else go to step2.

Fig. 4. Rough-k-Means algorithm

8. Experimental Results

In this work, 300 protein sequences are extracted from the Protein Sequence Culling Server (PISCES) as the dataset. In this protein database, the percentage identity cutoff is 25%, the resolution cutoff is 2.2, and the R-factor cutoff is 1.0. With these protein sequences, sliding windows with ten consecutive residues are obtained. Each window contains one sequence segment of ten continuous positions. This sliding window approach generates 67,186 segments. K-Means and Rough-K-means algorithm are applied to these segments and they are clustered into 200 clusters. The threshold value is set as 1, wlower= 0.7, wupper= 0.3 for Rough-K-Means algorithm. The secondary structure information is used as biological evaluation criteria. The higher HSSPBLOSUM62 value indicates more significant motif information.

Table 1. Comparison of HSSP-BLOSUM62 measure and percentage of sequence segments belonging to clusters with high structural similarity.

Different Methods	>60%	>70%	H-B Measure
K-means	90%	84%	0.9216
Rough- K -Means	93%	86%	1.3278

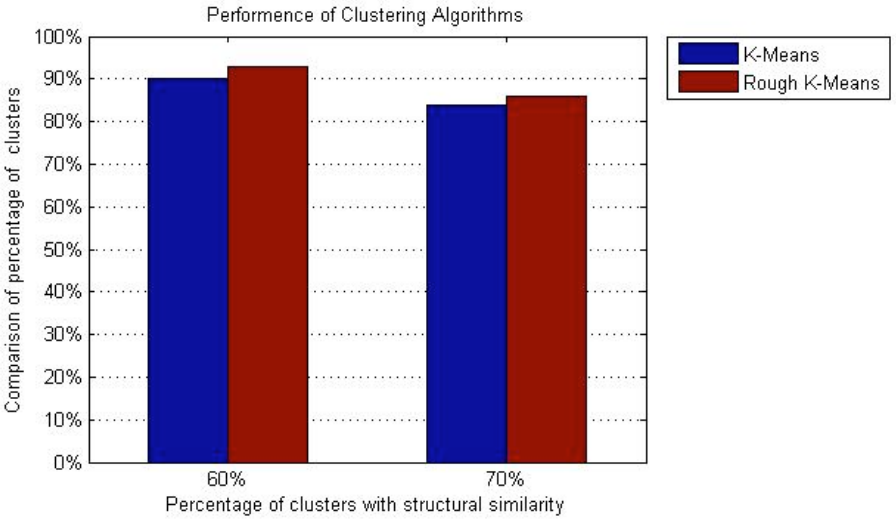


Fig. 5. Comparison of percentage of clusters with structural similarity > 60% and 70%

9. Representation of Motif Patterns

The following format is used for the representation of each motif table.

- The first row represents the number of members belonging to this motif, the secondary structural similarity and the average HSSP-BLOSUM62 value.
- The first column stands for the position of amino acid profiles in each motif with window size ten.
- The second column expresses the type of amino acid frequently appearing in the given Position. If the amino acids are appearing with the frequency higher than 10%, they are indicated by upper case; if the amino acids are appearing with the frequency between 8% and 10%, they are indicated by lower case.
- The third column corresponds to the hydrophobicity value, which is the summation of the Frequencies of occurrence of Leu, Pro, Met, Trp, Ala, Val, Phe, and Ile.
- The fourth column indicates the value of the HSSP-BLOSUM62 measure.
- The last column indicates the representative secondary structure to the position.

Table 2. Coil-Helices-coil Motif with conserved aST

Number of segments:207
Structure Homology:75.1%
AvgHssp-Blosum62: 0.5649

#	Amino Acid	H	B	S
1	gSt	28.82	0.01	C
2	GaSTn	54.57	0.12	C
3	gaST	49.54	0.16	C
4	aST	41.22	0.72	C
5	aST	66.99	0.72	H
6	AST	34.96	0.70	C
7	aST	53.73	0.73	C
8	AST	26.06	0.74	C
9	aST	3.38	0.71	C
10	ST	27.05	1.0	H

Table 3. Coil-Helices Motif with conserved VLI

Number of segments:399
Structure Homology:.74%
AvgHssp-Blosum62: 1.6659

#	Amino Acid	H	B	S
1	A	74.02	4.00	C
2	G	30.09	6.00	C
3	Vlia	42.90	0.79	C
4	sknD	47.26	0.15	C
5	VA	31.22	0.00	H
6	VLI	37.93	2.13	H
7	VLI	28.19	1.95	H
8	VLI	36.15	2.00	H
9	ApSt	85.46	-0.08	H
10	gaSd	40.06	-0.29	C

10. Conclusion

Proteins are involved in every body functions including nutrient transportation, muscle building, metabolism regulation, etc. Understanding the functions and structures of proteins encourages cellular process discovery. In this paper we have obtained the data set from the Protein Sequence Culling Server (PISCES). The sliding windows with ten successive residues were generated from protein sequences. These sequence segments of ten continuous positions were clustered into different groups with K-Means and Rough-K-Means algorithm.

Acknowledgement

The first and second author would like to thank UGC, New Delhi for the financial support received under UGC Major Research Project No. F-34-105/2008.

References

[1] Bairoch A, Sucher P, and Hofmann K, "PROSITE: New Developments" , Nucleic Acids Res 1996, 24:189-196.
[2] Attwood TK, Beck ME, Bleasby A J, Degtyarenko K, Smityh DJP, " Progress with the PRINTS protein fingerprint database", Nucleic Acids Res 1996, 24:182-183.
[3] Pietrokovski S, Henikoff JG, Henikoff S, "The BLOCKS Database - a system for protein classification", Nucleic Acids Res 1996, 24:19?-200.
[4] C. Sander and R. Schneider, "Database of homology-derived protein Structures and the structural meaning of sequence alignment," *Proteins Struct. Funct. Genet.* 1991, vol. 9, no. 1, pp. 56–68,.
[5] Bhattacharya, S. , " Gibbs Sampling Based Bayesian Analysis of Mixtures with Unknown Number of Components", *Sankhya. Series B*, 2009 To appear.
[6] G. Karp, *Cell and Molecular Biology (Concepts and Experiments)*, 3rd Ed. New York: Wiley, 2002, pp. 52–65.
[7]Zhong, W., Altun, G., Harrison, R., Tai, P. C. & Pan, Y. (2005) "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property", *NanoBioscience, IEEE Transactions on. 4*, 255-265.
[8] K. F. Han and D. Baker, "Recurring local sequence motifs in proteins,"*J. Mol. Biol.*, 1995, vol 251, no. 1, pp. 176–187,.
[9] P. Lingras, C. West, "Interval set clustering of web users with rough *k*-means", *J. Intell. Inform. Syst.* 23 (2004) 5–16.
[10] P. Lingras, R. Yan, C. West, "Comparison of conventional and rough *k*-means clustering, in: International conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing", *Lecture Notes in Artificial Intelligence*, vol. 2639, Springer, Berlin, 2003, pp. 130–137.
[11] Margaret H. Dunham, Data Mining- "Introductory and Advanced Concepts", Pearson Education, 2006.
[12] G. Wang and R. L. Dunbrack, Jr., "PISCES: a protein sequence-culling server," *Bioinformatics*, 2003, vol, 19, no. 12, pp.1589-1591.