IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Detection of Motif in Protein Sequence Using K-Means and Fuzzy C-Means Algorithms

Geethamani. R[1*], Kalaivani. B[2]

[1, 2] Department of Computer Science K.R College of Atrs And Science, Kovilpatti, Tamilnadu, India

DOI: https://doi.org/10.26438/ijcse/v7si8.8790 | Available online at: www.ijcseonline.org

Abstract**—** Finding the Motif in biological sequences of protein synthesis is a basic problem in determining the protein structure. Detection of the Motif is used with many applications in gene regulation, protein family identification and determination of functionally and structurally important identities. Large amount of biological data is used to resolve the problem of discovering patterns in biological sequences computationally. In this research, we have designed an approach using a system of clustering in data mining to detect frequently occurring informative motifs that are high in information content. We have proposed a comparative approach for Skin Melanin associated problems(SMA) detection in preliminary stages using protein sequence. We have used the protein sequence with normal and abnormal data as the trained dataset. Test instances are classified into normal to abnormal by comparing it with the fundamental dataset. In this paper, We compare and evaluate the performance of two clustering algorithms namely K-means and fuzzy c-means clustering for protein sequences.

Keywords— clustering ; k-means and fuzzy c-means; SMA

I. INTRODUCTION

Sequence motifs in protein are signatures of protein families and can be used as key for the prediction of protein function. The analysis and adaptation of already known motifs leads to major part of innovation; even new motifs are still being discovered at an approximately linear rate. The emphasis of motif analysis appears to be shifting from metabolic enzymes, in which motifs are associated with catalytic functions and thus often recognizable, to structural and regulatory proteins, which contain more divergent motifs. The resemblance of structural information increasingly contributes to the identification of motifs and their sensitivity. Genome sequencing provides the basis for a systematic analysis of all motifs that are present in a exacting organism. A systematically derived motif database is therefore feasible, allowing the classification of the common of the newly appearing protein sequences into known families.

Motifs are believed to be particularly important in biology and Bioinformatics. The discovery of information determined in biological sequences is assuming a eminent role in identifying genetic diseases and in deciphering biological mechanisms. This information is usually encoded in patterns frequently occurring in the sequences.

Motif discovery is the difficult step to understand the regulatory functions of genes. The motifs can represent patterns which trigger the transcription process and are responsible for gene expression regulations. In Bioinformatics, motif discovery is very important because they represent preserved sequences which can be biologically meaningful. It could be essential to the analysis and understanding of the biological data. If a pattern occurs frequently, it ought to be important or meaningful in some way. Motifs are frequent patterns in biological data that are presumed to have a biological function. Often they indicate sequence specific binding sites for proteins. Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing and transcription termination, growing usefulness in defining regulatory networks of genetic and deciphering the regulatory program of individual genes make them important tools for computational biology in the post-genomic era. Also, Motifs are important in understanding the fundamental cause of amyloid illnesses, pharmaceutical and industrial purposes.

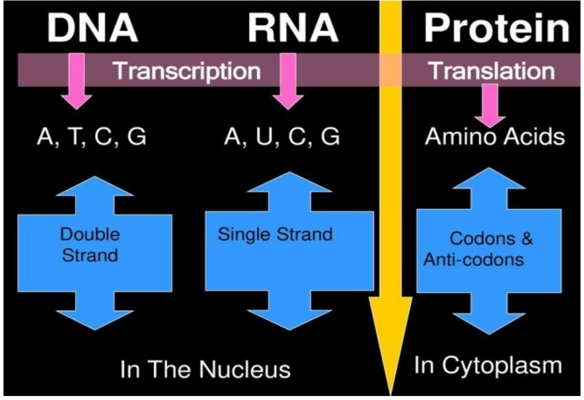

Fig:1.1 Transcription and Translation

Protein type is identified by protein sequences classifiers, but they then need to be characterized further, to consign biological function. The challenge is in this application of useful biological understanding to particular protein sequences. There are many reasons to select distinguish the proteins rather than DNA sequences. These include: the larger alphabet (21 amino acids versus 4 bases); the lower signal-to-noise ratio in protein sequence searches; the closeness between protein sequence and function; and the availability of good, well annotated databases of protein sequences and protein sequence signatures. Proteins can be characterized at different levels in a cell that perform a function, but this function is also performed within a particular perspective, for example as part of a complex pathway, as well as at a defined cellular place. At the functional level, this may come down to analysis of the protein sequence along its entire length, at the level of single domains or motifs, or at the finest level, single important amino-acid residues. With the increased availability of completely sequenced genomes, and using the correct tools and resources, there is range for protein characterization on all these levels.

The initial step in the analysis of new or uncharacterized protein sequences is usually to seek the protein databases for related sequences. The main protein sequence databases available are SWISS-PROT and TrEMBL, the Protein Information Resource (PIR), which is a translation of GenBank. If the similarity to proteins in a database is major, information from the proteins in the search results can be conditional to apply to the query sequence. This relies on the quality of the explanation in the protein sequence databases, and, more generally, on the availability of experimental results in the scientific literature. But problems arise during similarity of sequence finds when more than one domain is present in a protein. A large number of matches to one domain in a sequence may mask 'hits' that match a second domain in the sequence, and it makes important information is lost. It is also possible for sequences to be evolutionarily related but for their sequences to have diverged to such an extent that they are not taken in a formation of similarity sequence search. And, with the increase in population of protein sequence databases, the number of related sequences rises, so when a search is performed it identifies a large set of highly related sequences and the less related sequence hits may be lost. It is for these reasons that protein signature databases evolved and have become increasingly useful tools for protein sequence analysis; they aim to identify domains, or classify proteins into families, and thereby infer function. A signature refers to the diagnostic entity used to identify a domain or family; it may be derived using a number of methods, like patterns and profiles. This article explains the main signature databases available for protein sequence analysis, their methods, and their individual uses.

The human genome contains 3164 million chemical nucleotide bases (A, C, T, and G). The average gene consists of more than 3000 bases, but sizes vary greatly, with the largest known human gene (2.4 million bases). Almost all (99.9%) nucleotide bases are closely same in all people. The functions are unknown for over 50% of discovered genes. Less than 2% of the genome codes for proteins. Frequent sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome. Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. The human genomes gene-dense are predominantly composed of the DNA building blocks G and C. In contrast, the gene-poor are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes. Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between. Chromosome 1 contains most of the genes (2968), and the Y chromosome has the fewest (231). In GenBank, freely available and updated by the Human genome DB.

II. PROBLEM OBJECTIVE AND METHODOLOGY

Clustering the proteins is used to mine the relationship between protein sequences and structures. Two techniques namely K-means and fuzzy c-means clustering are used for clustering proteins. The performance of two algorithms are analyzed and compared for finding the proficient technique. Clustering can be divided into two sub groups:

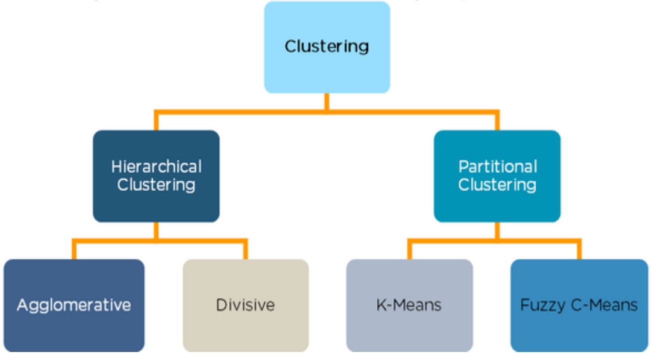

Fig:1.2 Types of clusteing

K-means algorithm

The K-means clustering program used to find frequent local sequence motifs for proteins. In their work, a set of initial cluster point centers is chosen randomly. Since the performance of K-means clustering is very sensitive to initial point selection, their technique may not give expected results. The k-means algorithm receives *k* number of input parameters and performs the partition on a set of *n* objects in the multidimensional space. The method of *k*-means works with the random selection of *k* number of objects and is represented as cluster means (cluster centers). Depending on the distance measure between the object and the mean of cluster, for each of the residual objects, a similar object is being assigned which helps to compute a new cluster mean. This process is continued till the convergence of criterion function. Hence, *k*-means is able to find the best cluster center points in the space. The general steps can be realized in Fig.1.2.

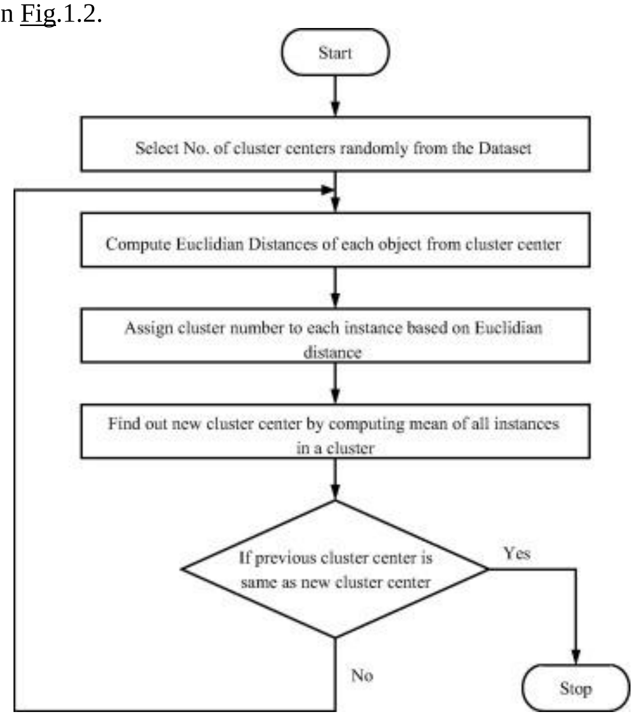
Start
Select No. of cluster centers randomly from the Dataset
Compute Euclidian Distances of each object from cluster center
Assign cluster number to each instance based on Euclidian distance
Find out new cluster center by computing mean of all instances in a cluster
If previous cluster center is same as new cluster center
Yes
No
Stop

Fig 1.2 Steps of K means Algorithm

1) *Fuzzy c-means (FCM) algorithm*

The FCM algorithm makes use of fuzzy membership function. The fuzzy membership function is used to assign a degree of membership for each class. FCM is able to form new clusters having close membership values to existing classes of the data points. The technique of FCM implies on three basic operators such as fuzzy membership function, partition matrix and the objective function. FCM is used to partition a set of *N* clusters through minimization of the objective function with reference to the fuzzy partition matrix.

$$J_m = \sum_{i=0}^{N} \sum_{j=1}^{c} u_{ij}^{m} \parallel x_i - c_j \parallel^2, 1 \leq m < \infty$$

where x_j denotes the j^{th} cluster point, and v_i represents the i^{th} cluster center. $u_{i,j}$ is the membership value of x_j withy reference to cluster *i*. *m* denotes the fuzzy controlling parameter i.e. for the value 1, it will tend to hard partition and for the value of ∞, it tends toward the complete fuzziness. ‖ ‖ represents the norm function.

The iterative method is used to compute the membership function and cluster center as follows:

The steps of FCM algorithm are as follows:

1. Specify the number of cluster centers C.
2. Select an inner product metric using the formula of Euclidean norm and the weighting metric (fuzziness).
3. Calculate *U* (partition matrix) using Eq. (2).
4. Update the fuzzy cluster centers using Eq. (3).
5. Calculate the new objective function *J* using Eq. (1).
6. If $\|J_{new} - J_{old}\| \leq \epsilon$ then stop.
7. Else repeat steps 3-5.

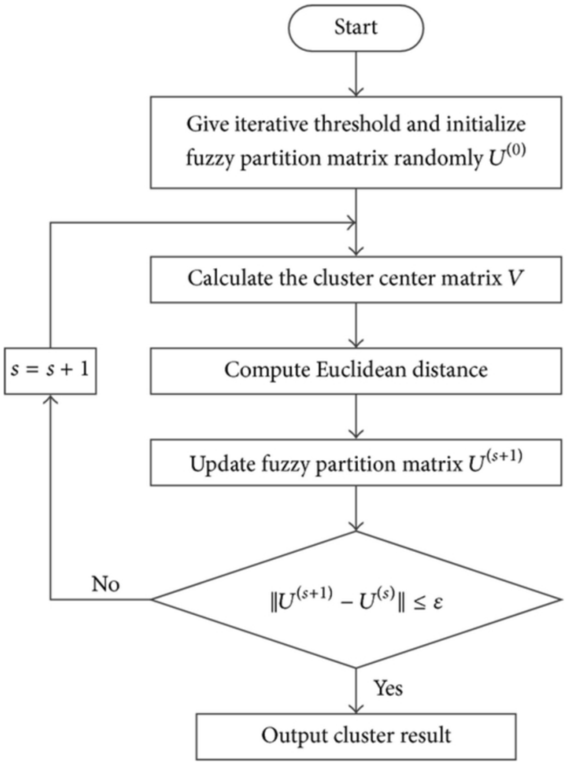
Start
Give iterative threshold and initialize fuzzy partition matrix randomly $U^{(0)}$
Calculate the cluster center matrix V
Compute Euclidean distance
Update fuzzy partition matrix $U^{(s+1)}$
$s = s + 1$
No
$\|U^{(s+1)} - U^{(s)}\| \leq \varepsilon$
Yes
Output cluster result

Fig: 1.3 Fuzzy c means clustering algorithm

Dataset:

The dataset used in this work includes 2290 protein sequences obtained from the Protein Sequence Culling Server (PISCES). Since PISCES uses PSI-BLAST alignments to distinguish many underlying patterns below 40% identity, PISCES produces a rigorous non redundant database.

this paper, we have proposed a probabilistic model for hair keratin associated proteins (SAP) are a major component of the hair fiber, and play crucial roles in forming a strong hair shaft through a cross-linked network with keratin intermediate filaments (KIF), which are produced from hair keratins. Recently, the study of human SAP has advanced significantly. So far, five clusters of human *SAP* genes have been characterized, leading to the identification of more than 80 individual human *SAP* genes. *In situ* hybridization studies have demonstrated sequential and spatial expression patterns of these SAP members in differential portions of the hair fiber cortex and cuticle. Furthermore, several human *SAP* genes have size polymorphisms that are mainly because of variable numbers of cysteine-rich repeat segments, and the patterns of some of these size variants are distinct between different human populations. This scheme takes supervised protein sequences from normal and abnormal category as the trained data set and is compared against a protein sequence from the user. This scheme has two phases, the Training Phase and the Testing Phase.

1. In Training Phase, the system will be trained with well studied and identified protein sequence affected by hair proteins. This phase takes 500 samples as the training data set with four fields (protein sequence, protein ID, HPV Type and Status) Training is carried out with supervised learning method. The status field in the training dataset indicates the normal and abnormal protein sequence.
2. The Testing Phase takes an input protein sequence and compares it with the trained data set. Each trained protein sequence is split into a number of small blocks and calculates the probability value for the occurrence of a particular block in the given input test data. This process is carried out for both Normal and Abnormal categories.

III. CONCLUSION

The risk identification is one of the most wanted research area in the field of bio-computation. In this paper, we have proposed a probabilistic approach for hair keratin associated problems detection in early stages using protein sequence. We have used the protein sequence with normal and abnormal instances as the trained dataset. Test instances are classified into normal or abnormal by comparing it with the training dataset. Threshold value is used for finding the normal and abnormal test instance. The threshold value will be changing for every new probability value. The proposed scheme needs minimum computational facility and it takes minimum time for classification. In future, a proper dataset can be developed from the available information of the CCDB database and clustering/classification algorithms can be used for getting better result.

In this study, the new initialization method for the Fuzzy C-means algorithm has been proposed to solve problems associated with random selection. In the new initialization method, we try to choose suitable initial points, which are well separated and have the potential to form high-quality clusters. Many biochemical tests published in the literature indicate that discovered sequence motifs are biologically meaningful. Analysis of sequence motifs also shows the improved K-means algorithm may detect some very subtle sequence motifs overlooked by the traditional algorithm. The discovered sequence motifs across protein families may overcome the shortcomings of other popular sequence motifs. The most updated dataset from PISCES is used for the first time to create sequence motifs. Because the dataset from PISCES has several advantages over other existing databases, sequence motifs discovered in this process can reveal more patterns that are meaningful during the process of evolution than other studies. Since the Fuzzy c-means algorithm is a very powerful tool for data mining problems, algorithm may be useful for other important bioinformatics applications.

REFERENCES

[1] Jipkate, BR & Gohokar, VV 2012 'A Comparative Analysis of Fuzzy C-Means Clustering and K Means Clustering Algorithms'. Int. J. of Computational Engineering, vol. 2, no. 3, pp. 737-739

[2] Bezdek, JC 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York., doi: 10.1007/978-1-4757-0450-1

[3] Bora, DJ & Gupta, AK 2014 'A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm'. Int. J. of Computer Trends and Technology, vol. 10, no. 2, pp. 108-113

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Protein and Nucleic Acids. Cambridge,U.K.: Cambridge Univ. Press, 1998.

[5] J. Finer-Moore and R. M. Stroud, "Amphipathic analysis and possible formation of the ion channel in an acetocholine receptor," Proc. Nat.Acad. Sci. USA, vol. 81, no. 1, pp. 155–159, 1984.

[6] D. Frishman and P. Args, "Knowledge-based protein secondary structure assignment," Proteins Struct. Funct. Genet., vol. 23, pp. 566–579, 1995.

[7] S. K. Gupta, K. S. Rao, andV. Bhatnagar, "K-means clustering algorithm for categorical attributes," in Proc. Data Warehousing and Knowledge Discovery (DaWaK-99), pp. 203–208

AUTHORS PROFILE

My name is Geethamani,R. I did M.Sc (CS& IT) in S.R.N.M College, sattur and M.Phil (CS) in Ayya nadar Janaki Ammal College. I Joined as a Assistant professor in ANJAC at 2005.I finished M.Tech in IT at M.S University in the year of 2012. Now I am working in K.R College of Arts and Science , Kovilpatti.I am having teaching experience of 13 years in the department of computer Science.