Refined Markov clustering Algorithm for Mycobacterium Tuberculosis Protein Sequence analysis

Dr.D.Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University
Coimbatore, Tamil Nadu
jaichitra1@yahoo.co.in

R.Geetha
Research Scholar
Department of Computer Science
Bharathiar University,
Coimbatore, Tamil Nadu
geethachitra90@gmail.com

Abstract— Clustering of proteins is an essential as it helps to infer biological function of a new sequence. In this paper, the protein sequences of Mycobacterium Tuberculosis have been clustered based on its space group using Refined Markov Clustering algorithm. The proposed technique reduces the overlapping clusters and performs better than other algorithms. This approach minimizes the proceeding time for the protein sequence effectively. The proposed work was evaluated by comparative analysis with k-medoids, spectral normalized cut and fast connected component algorithm. According to the clustering validation and comparison results, the proposed algorithm performs better than other algorithms.

Keywords- Data mining; Protein Sequence; Spectral Meli-shi; Markov Clustering

I. INTRODUCTION

With the rapid growth of biological sequences databases, extracting useful information from biological sequences is an emerging problem. As proteins are functionally essential in life, among the biological sequences, protein sequences are very interesting. The protein structure and function can be better studied with more accuracy and efficiency using many computational methods and one of the most important one is sequence clustering. Clustering the protein sequences helps to infer the biological function of the new sequence as well as it can be used to protein 3-dimensional structure discovery (1). A significant number of methods have addressed the clustering of protein sequences. COG uses a hierarchical merging of clusters (2), SYSTERS combines hierarchical clustering with graph based clustering (3), N- cut uses graph based clustering approaches (4). (5) used k means and rough k means clustering algorithms to predict local protein sequence motifs. (6) focused on generalization of k medoid style clustering algorithms on four different data sets. Also, several clustering algorithms such as Pro-k means, Pro-Leader, Pro-CLARA and Pro-CLARANS (7) have been proposed for clustering protein sequences. Since, very large number of protein sequences are deposited into the database in the recent years, an efficient clustering algorithm is needed to group the similar sequences based on their characteristics in minimum time.

Large number of protein sequences of different species is deposited into the protein database day by day. As a consequence of the increase in the human population, deficiency and overcrowding in big cities, the efficient control of tuberculosis (TB) remains a major problem both for developing and developed countries. Tuberculosis control has been complicated because of the development of resistance of the microorganism to first-line anti-tuberculosis drugs (8). (9) stated that extensively drug resistant tuberculosis (XDR-TB) has become a new threat for the control of TB in many countries including India. Hence, in this paper, the protein sequence of tuberculosis has been analyzed based on its characteristics. In this paper, we describe a Refined Markov Clustering algorithm for grouping the protein sequences based on its space group. Methods and clustering process is presented in section 2. Experimental results and datasets are given in section 3. Finally section 4 gives the conclusion.

II. METHODS AND MATERIALS

The proposed work consists of (A) System architecture (B) Clustering process and (C) Refined markov clustering algorithm

A. System Architecture

Our proposed work consists of the following parts:

- Extraction of Space Group from protein Data Bank
- Sequence Alignment
- Grouping of proteins based on parameters
- Implementation of clustering process

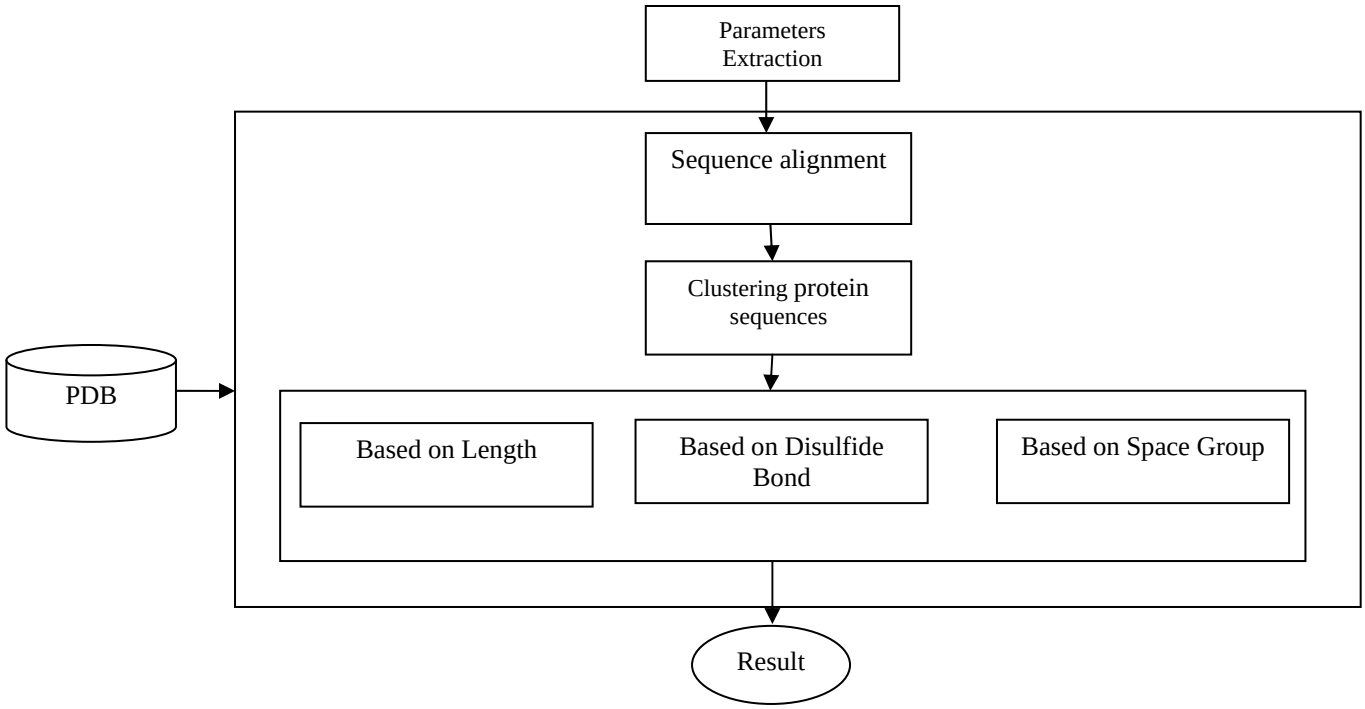

Fig.1. System Architecture

Figure 1 shows the system architecture. The input sequence is given and the length, disulfide bond, space group are extracted for the input sequence from PDB. Grouping of proteins is performed by clustering techniques. In this work, a Refined Markov clustering algorithm is proposed for clustering the mycobacterium tuberculosis protein sequences.

B. Clustering process

The algorithm is intended to find a sequence of object *K* called medoids that are centrally located in clusters. The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object *K*. A concept is pertinent, if and only if, it minimize the Shannon entropy. The Shannon entropy of a concept $C_i = (A_i, B_i)$ is defined by:

$$h(C_i) = -\sum_{i=1}^{n} \frac{n^k}{n_i} * \log\left(\frac{n^k}{n_i}\right) \quad (1)$$

Where *n* is the number of the sequences, n_i is the number of data of *Ai*, n^k is the number of data of A_i that belongs to the sequence $f_{k,}$ where $1 \leq k \leq n$ (10). Medoids for each cluster are calculated by finding object *i* within the cluster that minimizes

$$\sum_{j \in c_i} d(i,j) \quad (2)$$

Where *Ci* is the cluster containing objects *i* and *d* (*i; j*) is the distance between objects *i* and *j* (11). Each node in the graph corresponds to a protein sequence and the weight on each edge corresponds to the similarity between two protein sequences it connects (12). W_{ij} is the similarity between two points S_i and S_j then

$$W_{ij} = exp\left(\frac{-d^2(S_i, S_j)}{\sigma^2}\right) \quad (3)$$

Where $d = (S_i, S_j)$ denotes the Euclidean distance between two points S_i and S_j. The parameter σ controls the width of the neighborhood (13). The average F-score of the entire clustering C is defined as the average of the F-scores of all the clusters, weighted by their sizes (14).

$$Avg.F(C) = \frac{\sum_i |C_i| * F(C_i)}{\sum_i |C_i|} \quad (4)$$

Recently, large data sets of protein-protein interactions (PPI) which can be modeled as PPI networks are generated through high-throughput methods. The neighborhood graph of $v \in V$ consists of v which denotes all its neighbours and the edges among them. It is defined as $Dv = (V', E')$, in which $V' = \{v\} \cup \{u | u \in V, (g, v) \in E\}$, and $E' = \{(gi, uj) | (gi, uj) \in E, gi, uj \in V'\}$ (15). The neighbor affinity NA (A, B) of two clusters i.e. A = (V_A, E_A) and B = (V_B, E_B) is defined in Equation (5), for measuring their overlapping degree (16).

$$NA(A, B) = \frac{|V_A \cap V_B|2}{|V_A| \times |V_B|} \quad (5)$$

C. Refined Markov clustering Algorithm

Markov clustering (MCL) has been used for clustering biological networks—for instance clustering protein–protein interaction (PPI) networks to identify functional modules. However a limitation of MCL and its variants (e.g. regularized MCL) is that it only supports hard clustering often leading to an impedance mismatch given that there is often a significant overlap of proteins across functional modules. In this paper Refined Markov clustering Algorithm (RMCA) is proposed to reduce the overlapping clusters.

The parameters are same as MCL except introducing the penalty ratio β including the usage of parameters *l* and *p*. The resulting clustering from iterative execution of R-MCL contained several redundant and low-quality clusters and those clusters have been removed and tested using quality function denoted by q. We remove all clusters whose *q* value is below a user-specified threshold ω. The value of ω depends on *q* and the network's property. Furthermore all clusters whose size is ≤2 are also removed.

After removing low-quality clusters, it is examined whether each cluster is redundant or not in the descending order of its *q* value. A cluster c_j is removed if there exists a cluster *xi* that $qf(x_i) >= q\ (cj)$ and DA $(x_i, x_j) > p$, where *p* is another user-specified threshold and DA is neighborhood affinity.

$$DA(X_I, X_J) = \frac{|X_I \cap X_J|}{|X_I| * |X_J|} \quad (6)$$

Thus, *b* is used to control the degree of overlap among clusters. The higher *b* produces higher overlapped clusters and vice versa. If ω becomes larger and *b* is decreased, the post-processing removes more clusters and hence so the remaining high-quality clusters can precisely mismatch. On the other hand, less ω and larger *b* result in more clusters. The resulting clusters can identify more functional modules; however, the result contains relatively redundant and low-quality clusters and therefore so a number of resulting clusters cannot precisely mismatch functional modules.

In proposed RMC method, there is no need to specify the predefined number of clusters. It automatically determines the numbers of cluster and outliers. Outliers are the proteins that are not assigned to any cluster i.e. too different from other proteins. Clusters partitioned by proposed method RMC are having high intra cluster similarity and high inter cluster variation.

III. EXPERIMENTAL RESULTS

A. Datasets

To assess the performance of the proposed work and compare it with the existing methods, the dataset used here is Tuberculosis proteins. These proteins were divided into three types of datasets based on sequence length. First Dataset includes the proteins with the sequence length 100 to 199. Second Dataset includes the proteins with length 200 to 299. Third Dataset includes the proteins with length 300 to 399. Fourth Dataset includes the proteins with length 400 to 499.

B. Results and discussion

The clustering performance was assessed using two different methods and execution time. The first method named as similarity index is defined as the ratio between the overall within-cluster similarity and the overall between-cluster similarity. Based on this method, a good data clustering results in high intra cluster similarity and low inter cluster similarity. To find the overall within-cluster similarity, the similarity within each cluster is calculated as the average distance between each protein in the cluster and then averaged for all clusters. The between-cluster similarity is obtained by averaging the distance between each pair of clusters. Each single between-cluster similarity is calculated by averaging the distance between each pair of protein from the two clusters.

The correlation coefficient is denoted as *r* and calculated by using the Eq. 7. Where *n* is the number of data points, *x* and *y* are the pair of data points.

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n (\sum x^2) - (\sum x)^2} \sqrt{n (\sum y^2) - (\sum y)^2}} \quad (7)$$

Second validation index named as partition index is defined as the average 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 of standard deviation between each protein in the cluster and then averaged for all clusters. The between-cluster variation is obtained by averaging the standard deviation between each pair of clusters. Each single between-cluster distance is calculated by averaging the standard deviation between each pair of protein from the two clusters.

The standard deviation is denoted as *S* and calculated by using the below formula. Where *n* is the number of data points, $\bar{x}$ is the data point and x_i is the average of data points.

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (8)$$

According to the similarity index analysis, the proposed method RMC outperforms the other three algorithms. Also, in partition index analysis, RMC produces the best result than other algorithms. Fig. 2 shows Similarity index based on correlation co-efficient. Fig. 3 shows Similarity index based on standard division. Fig.4 Shows Partition index based on correlation co-efficient. Fig. 5 shows Partition index based on standard division.

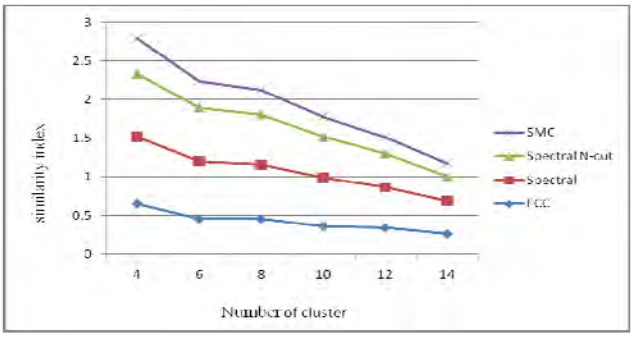

Fig. 2. Similarity index based on correlation co-efficient

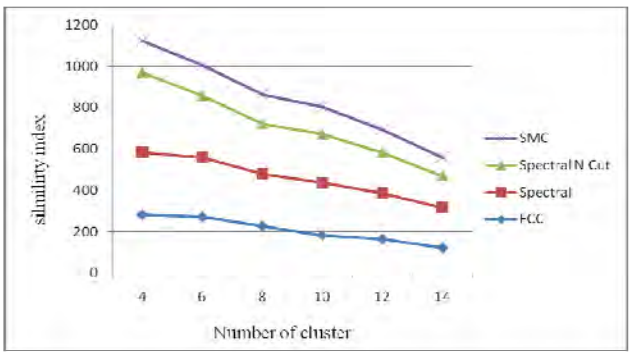

Fig. 3. Similarity index based on standard division

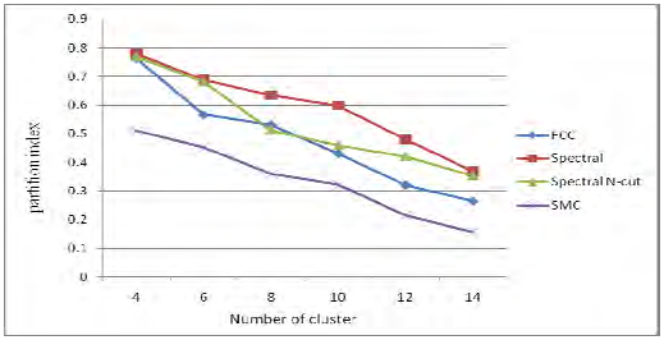

Fig. 4. Partition index based on correlation co-efficient

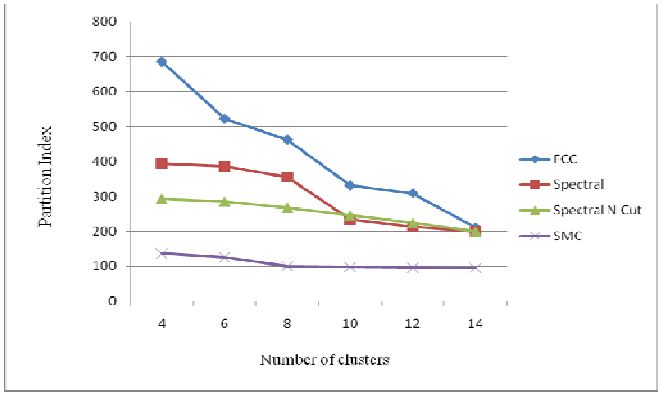

Fig. 5. Partition index based on standard division

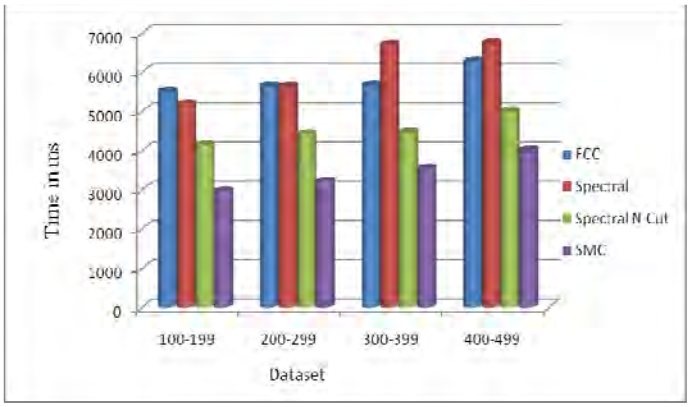

Fig. 6. Execution time (ms) of clustering algorithms for four datasets

The execution time of protein sequence using RMC algorithm was compared with other algorithms. In few cases the execution time of RMC increases when compared with Fast Connected Component, Spectral Meli-Shi, and Spectral Normalized-cut algorithm. Especially RMC performs better when compared to other clustering algorithms.

IV. CONCLUSION

The analysis of protein sequence is a kind of computation driven science which rapidly increases the size of biological data. The proposed method reduces the execution time for the analysis of Mycobacterium tuberculosis protein. According to the clustering validation and comparison results, the proposed algorithm performs better than other algorithms. Implementation of the clustering process offers the fast execution time and better performance. The proposed clustering technique can be easily extended to any other applications different from protein sequence analysis. Further research involves the improvement of the disulfide bond and space group for the Mycobacterium tuberculosis protein.

REFERENCES

[1] Yonghui Chen, Kevin D Reily, Alan P Sprague and Zhijie Guan, SEQOPTICS: a protein sequence clustering system, BMC Bioinformatics, 2006.

[2] R .Tatusov, N .Fedorova, J .Jackson, A. Jacobs, B .Kiryutin, E .Koonin, D .Krylov, R .Mazumder, S.Mekhedov,A .Nikolskaya, B .Rao, S . Smirnov, A .Sverdlov, S .Vasudevan, Y .Wolf, J .Yin , D .Natale: The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 2003**.**

[3] A .Krause, J .Stoye, M .Vingron: Large scale hierarchical clustering of protein sequences**.** *BMC Bioinformatics* 2005.

[4] Shi J, Malik J: Normalized cuts and image segmentation**.** Proceedings of the IEEE conference on Computer Vision Pattern Recognition, pp.731-737, 1997.

[5] E.Elayaraja, K.Thangavelu, B.Ramyaa, M.Chitraleghaa, Extraction of motif patterns from protein sequence using rough k-means algorithm, pp. 814-820, 2012.

[6] Nidal Zeidat, Christoph, F., Eick, K-medoid-style Clustering Algorithms for Supervised Summary Generation, pp.932-938, 2005.

[7] Sondes Fayech, Nadia Essoussi, Mohamed Limam, Partitioning clustering algorithms for protein sequence data sets, 2009.

[8] K .Penuelas-Urquides, L .Gonzalez-Escalante, L .Villarreal-Trevino, B .Silva-Ramırez, D. J .Gutierrez-Fuentes, R .Mojica-Espinosa, C .Rangel-Escareno, L .Uribe-Figueroa, G. M .Molina-Salinas, J .Davila-Velderrain, F .Castorena-Torres, M .Bermudez de Leon, S .Said-Fernandez, Comparison of Gene Expression Profiles Between Pansensitive and Multidrug-Resistant Strains of Mycobacterium tuberculosis, 2013.

[9] Joy Sarojini Michael & T .Jacob John, Extensively drug-resistant tuberculosis in India: A review, 2012, pp.599-604.

[10] M .Maddouri, and M .Elloumi, A data mining approach based on machine learning techniques to classify biological sequences, 2002, pp.217-223.

[11] A .Yoshida, E .Freese, Purification and chemical characterization of alanine dehydrogenase from Bacillus subtilis, Biochem. Biophys. Acta, Vol. 1964. pp.33–43.

[12] S .Inderjit, Dhillon, Yuqiang Guan, Brian Kulis, Kernel kmeans, Spectral Clustering and Normalized Cuts, 2004, pp.22–25.

[13] Xianchao ZHANG, Quanzeng YOU, An improved spectral clustering algorithm based on random walk, 2011, pp.268–278.

[14] Venu Satuluri, Srinivasan Parthasarathy, Duygu Ucar, Markov Clustering of Protein Interaction Networks with Improved Balance and Scalability2010.

[15] M .Wu, X .Li, C-K .Kwoh, S-K .Ng, A Core-Attachment based Method to Detect Protein Complexes in PPI Networks. BMC Bioinformatics, 2009.

[16] Shuliang Wang, Fang Wu, Detecting overlapping protein complexes in PPI networks based on robustness, 2013.

[17] C .Diogo Stelle Maria, P .Barioni Luis, Scott, Using data mining to identify structural rules in proteins, Applied Mathematics and Computation, 2011, pp.1997–2004.

[18] J .Han, M .Kamber, Data Mining: Concepts and Techniques. 1st edition. Morgan Kaufmann Publishers, 2000.

[19] J .Han, M .Kamber, Data Mining: Concepts and Techniques, second ed., Morgan Kaufmann, New York, 2006.

[20] S .Kane, PK .Dewan, D .Gupta, A .Das, A .Singh, G .Bitra, LS .Chauhan, G .Dallabetta, Large-scale public-private partnership for improving TB-HIV services for highrisk groups in India, 2010.

[21] Khalid Raza, Application of data mining in bioinformatics, 2012, pp.114-118.

[22] B .Lodish, et al, Biologia Celular Molecular, five ed., Artmed, Porto Alegre, 2005.

[23] C .Michely, L .Diniz, Ana Carolina, T .Pacheco Karen, F .Girao Fabiana, A .Araujo Cezar, M .Walter Diana, The tetratricopeptide repeats (TPR)-like superfamily of proteins in Leishmania spp., as revealed by multi-relational data mining, 2010,pp. 2178–2189.

[24] Severine Ferdinand, Georges Valetudie, Christophe Sola, Nalin Rastogi, Data mining of Mycobacterium tuberculosis complex genotyping results using mycobacterial interspersed repetitive units validates the clonal structure of spoligotyping-defined families, 2004,pp.647–654.

[25] Sivakumar Shanmugam, N.Selvakumar, SujathaNarayanan, Drug resistanceamongdifferentgenotypesof Mycobacteriumtuberculosis isolated from patientsfromTiruvallur,SouthIndia, Infection, Genetics and Evolution, 2011,pp.980–986.

[26] P .Themis, G .Exarchos Markos, b.Tsipouras, D .Costas Papaloukas, I.Dimitrios, Fotiadis, A two-stage methodology for sequence classification based on sequential pattern mining and optimization, Data & Knowledge Engineering, 2008,pp.467–487.

[27] Xin-Xu Li, and Xiao-Nong Zhou, Co-infection of tuberculosis and parasitic diseases In humans: a systematic review, Li and ZhouParasites&Vectors, 2013,pp. 6-79.

[28] Zhiyong lou1, Xiaoxue Zhang, protein targets for structure-based anti-mycobacterium tuberculosis drug discovery, 2012, pp. 111-117.