Tapered Assessment on Distributed Clustering vital in Protein Sequence Environment
Assessment on Distributed Clustering with Protein Data

K.Thenmozhi 1,M.Pyingkodi 2,S.Kumaravel 3
1Assistant Professor, Department of Computer Applications, Selvam College of Technology, Namakkal, TamilNadu, India,
2Assistant Professor, Department of Computer Applications, Kongu Enginner College, Erode. Tamilnadu, India,
3Assistant Professor, Department of Computer Applications, Selvam College of Technology, Namakkal, TamilNadu, India,

Abstract — The ever-increasing size of data sets, poor scalability, space and time of execution of clustering algorithm has haggard attention to distributed clustering for partitioning large data sets. Protein sequence prediction is one of the vital roles in bioinformatics, which is used to analyze the biological data. The combination of Distributed clustering algorithm and soft computing techniques used to discover the gene/protein structure or sequence. Soft computing is a collection of algorithms that are employed for finding a solution because of their ability to handle imprecision, uncertainty in large and complex problem. The vital role of distributed clustering algorithm is to cluster the distributed datasets without collecting all the data into single site. Cluster the data locally and extract the representatives of these clusters and send to global site where the cluster based on local representative. It deals with large homogenous/heterogeneous data for any application using soft computing approaches. Clustering is the process of similar object grouped into one cluster and dissimilar object grouped in other. The effort is being taken to progress the efficiency of distributed combining algorithm using different soft computing techniques for protein data.

Keywords: Distributed Clustering, Soft computing, Protein sequence, Clustering.

I. INTRODUCTION

Biological data is a great challenge for clustering algorithm. The time and space complexity is great holdup for large-scale data sets. Parallel Propagation performance is apt for clustering huge gene data and detects the families in large protein families [1]. In Distributed clustering, the Objects are clustered into different sites. It’s used for instead of transmitting all the data into single site. Clustered the data locally and extract the representatives of these clusters send to global site where the clusters based on local representatives.

A protein sequence motif is a vital role in bioinformatics. A protein consists of a chain of amino acids connected by peptide bonds. It is a linking of amino acids that is the combination of amine group and acid group. The combination of two or more amino acids is peptides. Protein formed by linking of amino acids and peptides. Enzymes, hormones, transcription factors, pumps and antibodies are examples for the diverse functions fulfilled by proteins in a living organism. The protein sequence motif discovered by Smith-Waterman, FASTA, BLAST, PSI-BLAST. Three main attributes of proteins such as sequence, structure, and function [2].

Prediction of Protein structure and sequence is an essential task in bioinformatics research. Sequence motif denotes the pattern repeatedly occurred. The protein structure and function predicted by motif patterns. PROSITE, PRINTS and BLOCKS are popular motif DB. It is discovering from MEME,
Gibbs, Sampling and Block Maker. The input data is huge so granular FCM used predict the motif sequence [3].

The relationship among protein sequence and structure is one of the vital tasks in bioinformatics. The motif search is the sequence alignment from same protein family. Han and Banker used k-means clustering to discover the protein sequence motif. Initial random selection is not suitable to get an optimal result. In k-means, the greedy algorithm is used to select an initial point to overcome this problem.

The protein sequence is segmented by sliding techniques. Improved k-mean clustering algorithm is used to predict the motif sequence. The input data is huge so granular FCM used predict the motif sequence Granular SVM take more time for segmenting the process. Accordingly we introduced the SVD (Singular Value Decomposition). The FCM, AFCM (Adaptive FCM), RK-means (Rough K-means), Granular FCM, Granular AFCM and Granular RK-mean used to identify the motif sequence [3-5]. We look forward to Distributed cluster instead of sliding segment techniques, it show the true positive result in terms of space, time and pattern sequence.

II. CLUSTERING AND DISTRIBUTED CLUSTERING
Cluster analysis is an unsupervised method of huge data analysis process. It is the processes of grouping objects into clusters such that the objects in the same cluster are similar where as objects in different clusters are different.

Objects are clustered into different sites. The Distributed clustering algorithms used for instead of transmitting all the data into single site. Clustered the data locally and extract the representatives of these clusters send to global site where the clusters based on local representatives [6]. A "clustering" is a set of clusters that containing all objects in the data set. It can be classified as,

i) Hard Cluster: Each object Belongs to cluster or not

ii) Soft Cluster (Fuzzy): One object can occur more than one cluster [7].

It follows the various algorithms,

<table>
<thead>
<tr>
<th>Types</th>
<th>Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connectivity-based(hierarchical) clustering</td>
<td>Divisive, Hierarchical clustering</td>
</tr>
<tr>
<td>Distribution-based clustering</td>
<td>DBCLASD</td>
</tr>
<tr>
<td>Density-based clustering</td>
<td>Partitioning - Wave Cluster, Den Clue, Clique, DBSCAN, Hierarchical - Grid Clustering, Bang, OPTICS, Chameleon</td>
</tr>
<tr>
<td>Optimization / Distance / Centroid-based Clustering</td>
<td>Partitioning Clustering - k-means, k-modes-medoid, PAM, CLARA, CLARANS, Hierarchical Clustering - single link, CURE, BIRCH</td>
</tr>
</tbody>
</table>

A) Evaluation Assessment:
Validate the cluster results by Internal, External, Manual, and Indirect Evaluation [8].
### Distance Evaluation

The distance is preferred to recognize the relationship among data.

- Minkowski distance
- Standardized Euclidean distance
- Cosine distance
- Pearson correlation distance
- Mahalanobis distance
- City block Distance, etc.

To predict the protein sequence motif, mostly the city block distance used instead of rest of the distance measured.

### III. CLUSTERING ALGORITHMS CONTRIBUTE IN PROTEIN SEQUENCE

TRIBE-MCL clustering is an efficient optimal grouping of protein sequence with matrix square to improve the execution speed and time [9]. K-Means clustering algorithm and Rough-K-means algorithm, used to predict the protein sequence motif. Clustering is emerging field to find a pattern in the motif sequence of DNA (Deoxyribo-Nucleic Acid), RNA (Ribo-Nucleic Acid) and Proteins [10]. Ultra fast CD-HIT (Cluster Database at High Identity with Tolerance) can be efficiently cluster a large protein database with millions of sequences. CD-HIT and version of CD-HIT compares two protein datasets to find similar matches among them. Its more times faster than BLAST (Basic Local Alignment Search Tool.) [11]. UBLAST and USEARCH algorithms are faster than BLAST and CD-HIT. It improves the speed and sensitivity, and lower cluster identifiers and memory use, and classification of huge datasets [12].

The prediction of protein functions for a subset from the Gene Ontology. It’s categorized by transcription factors, receptors, ion channels, stress and immune response proteins, hormones and growth factors. The protein sequence can be identified in web interface [13]. The main objectives in structural and functional genomics are predicting the protein families in huge database. The Markov cluster (MCL) algorithm used for proteins into families based on pre computed sequence similarity information. It tested and validated on SwissProt, InterPro, SCOP (Structural Classification of Proteins) and the draft human genome in huge database. This method is exactly suitable for protein families [14].

Template-based protein structure prediction is used to protein sequence alignment [15]. The machine-learning method of SVM used to find the four structural classes in CATH database. Among four classes, multi class gives the accuracy in SVM [16]. Novel PCA of Hierarchical PCA-EELM (principal component analysis-ensemble extreme learning machine) model to predict protein-protein interactions in protein sequences. It focuses on dimension reduction and improves the prediction performance [17].

<table>
<thead>
<tr>
<th>Name</th>
<th>Formula</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davies–Bouldin index</td>
<td>$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{j \neq i} \left( \frac{\sigma_i + \sigma_j}{d(c_i,c_j)} \right)$</td>
<td>$n$ - number of clusters, $c_x$ - the centroid of cluster $x$, $\sigma_x$ - average distance of all elements in cluster $x$ to centroid $c_x$, $d(c_i,c_j)$ - distance between centroids $c_i$ and $c_j$</td>
</tr>
<tr>
<td>Dunn index</td>
<td>$D = \frac{\min_{i \neq j} d(i,j)}{\max_{k \neq l} d(k,l)}$</td>
<td>$d(i,j)$ - distance between clusters $i$ and $j$, $d'(k)$ - measures intra-cluster distance of cluster $k$</td>
</tr>
<tr>
<td>Silhouette coefficient</td>
<td>$(b - a) / \max(a, b)$</td>
<td>Mean intra-cluster distance $(a)$, mean nearest-cluster distance $(b)$</td>
</tr>
</tbody>
</table>
The protein sequence-to-structure relationship is a vital task in bioinformatics research. SVM is not suitable for huge datasets. Thus, the novel method of Clustering Support Vector Machines (CSVMs) is strong enough to recognize the complicated pattern of sequence-to-structure relationships with granular computing [18]. A new method is described. Hydrophobic cluster analysis (HCA) is used for comparing and aligning the protein sequences. It can efficiently evaluate the closest features of protein sequence [19].

The clustering of protein sequences can be categorized in three leading groups: hierarchical, graph-based and partitioning methods. An analysis of literature view, the hierarchical and graph-based approaches has been widely used. Even if partitioning clustering techniques are enormously used in other fields, few applications have been found in the field of protein sequence clustering. The four partitioning clustering approaches is using Smith-Waterman local-alignment algorithm to determine pair-wise similarities of sequences. Few applications only found in partitioning clustering algorithm to predict the protein sequence clustering, it's not fully recognized.

Two different clustering algorithms such as K-Means and agglomerative from partitioning and hierarchical clustering methods are used to predict the protein sequences. Clustering is used to discover the relationship among proteins and evaluate the performance of above mentioned algorithms. Hierarchical clustering is better than K-Means clustering in terms of validity indices and time of execution [20]. The protein sequence predicted by the various clustering applications, such as Pro-Clust, Tribe-MCL, JACOP, Pro-K-means, Pro-LEADER, Pro-CLARA, Pro-CLARANS, and so on. These all are the partitioning clustering algorithm for predict the protein sequence.

Protein clusters according to frequency of their amino acids in UniprotKB/Swiss-protDB with the use of hierarchical cluster analysis. Protein clustering is a valuable for unknown metabolic associations and constrains [21]. Each Sequence represented by probability densities functions (PDF) similarity of protein measured by Hellinger distance. Two protein data sets are mixed between Influenza and Ebola virus. It compared by two hierarchical clustering algorithm similarity measure with sequence alignments (HCAWSA) and similarity measure without sequence algorithm. It gets the feasible and accurate solution [22].

Protein sequence motif evaluated by its motif and their structure. A greedy initialization used to improve the k-means clustering techniques. This improved k-means clustering algorithm discovers the weak and clever sequence motifs [23]. FCM is used to identify the protein sequence motif. This algorithm used to separate the whole data set into smaller granules. Granular and DBI measure introduce for this data set. The granular FCM and Improved k-means combination (FIK) to produces better motif sequence pattern than k-means [24]. The sequence motif identified from sequence segments. The sequence segment doesn’t give potential motif apply unsupervised segment selection techniques SVD combined to FCM Granular computing and finds the sequence motif pattern from protein families [25]. Biclustering is used to solve optimization problem in gene expression dataset which find the exact motifs sequence in future. Random cuckoo search gives better result than PSO in Biclustering [26].

Protein sequence motif of particular protein discovers the structure, functions and activities. Rough Granular computing is introduced, before that we reduce the segment size by SVD. Subsequently, apply RGC. The SVD with rough granular computing generates highly structured motif patterns. The motif predicted by the use of DBI, HSSP-BLOSUM62 [27]. Motif identified by new seed initialization techniques for sorted pair wise distance calculated. The seeds for k-means better than cluster for identify motif sequence pattern [28].
IV. SOFT COMPUTING IS A MILLSTONE ON PROTEIN SEQUENCE

Soft computing approaches include Genetic algorithm, Hidden Markov Model, Fast Fourier Transformation, Support Vector Machine, Dynamic programming and Artificial Neural Network. Hybrid soft computing approaches are delivering the better accuracy than individual soft computing approaches [29]. Soft computing differs from hard computing in terms of imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness and low solution cost for the emerging field of conceptual intelligence. Soft computing techniques such as fuzzy logic, neural networks, genetic algorithms, etc., are used to solve the issues allied to data mining. Not only individual soft computing methodologies but their hybridization also has proved [30].

The soft computing approaches are: Fuzzy Logic (FZ), Artificial Neural Networks (ANN), Evolutionary Algorithms (EAs) (including genetic algorithms (GAs), genetic programming (GP), Neural Networks (NN), evolutionary strategies (ES)), Support Vector Machines (SVM), Simulated Annealing (SA), Ant Colony Optimization (ACO) and Tabu Search (TS). The application of soft computing methods like fuzzy set, artificial neural network and genetic algorithm in bioinformatics have been briefly discussed [31].

Artificial neural networks is widely used in analyzing gene expression level data on the genomic scale, sequencing genes on the genomic scale, sequencing proteins and amino acids, etc., in terms of nonlinear mapping, high accuracy for learning, and good robustness. Review of some artificial neural networks and related pattern revealing techniques used in protein sequence analysis. Soft computing techniques used to solve the bioinformatics problems in lots of data that may have noise, missing values, uncertainties, etc. [32].

The Protein sequence is compared by the amino acid similarity matrices. The connection between the properties and similarity matrices of amino acids whished by equivalence classes [33]. Both plants and animals genomes have been sequenced. Soft computing techniques developed to predict the specific genome and accuracy level of gene [34]. Protein sequence is predicted by fuzzy logic, genetic algorithms, neuro-fuzzy, and neuro-genetic. The Markov model combined with NN is to propose the better results [35]. The soft computing based splice sites predictor can’t solve the canonical splice sites problem. It gives a false positive result in screenings test. The soft computing techniques such as fuzzy logic, genetic algorithms, neuro-fuzzy and neuro-genetic are suitable for gene/protein prediction [36].

A neural network method (SPINE-2D) is a sequence-based prediction of residue–residue contact maps. It predicts the secondary structure protein via huge training. Meanwhile, accurate method for predict the contact map [37]. These methodologies used in numerous focuses such as biology, medicine, computer science, engineering, chemistry, physics, and mathematics. Bioinformatics is a fast growing field in sequence analyses of DNA, RNA and Protein. The soft computing techniques used to solve the bioinformatics problem of uncertainty in huge space, and handling the imprecision. It perform the numerous tasks such as gene mapping, DNA, RNA and Protein alignment and comparison, Gene regulatory network, Structure prediction of Protein. Molecular design and docking are the various algorithm used in this field such as finding similarities in string, pattern prediction, similarities among structure (motifs). Fuzzy logic is improving the protein sequence motifs. GA is used for Protein structure prediction and clustering, gene regulatory network identification [38].

The k-means and fuzzy k-means groups similar patterns. GA and Particle Swarm Optimization (PSO) check the optimality of these algorithms. It demonstrates the number of cluster is more and less runtime by using fuzzy k means. Subsequently, it gives best result for finding protein or gene sequence than traditional algorithms [39]. The FCM, Adaptive Fuzzy Granular model with SVD
entropy, Fuzzy Granular with SVD Entropy, and Rough Granular model with SVD Entropy, K-means algorithms used to explore the highly similar protein sequence motif from huge data set [5]. A [40] metaheurisic approach with a Genetic Algorithm with Levy Flight (GA-LV) was applied to classification of cancer genes from microarrays. It gives better accuracy.

Protein sequence is a vital and emerging filed in bioinformatics. Sequence motif clustered with the help of soft computing techniques especially FCM, Adaptive FCM, K means, Rough K-means. In Distributed environment suitable for huge data set, segmenting, clustering the centroid, and reduce the space and time. Its shows the better screening tests such as True Positive and True Negative values.

V. DISTRIBUTED CLUSTERING ROLE IN PROTEIN SEQUENCE ENVIRONMENT

The most vital intent of distributed clustering algorithms is to cluster the distributed datasets without obligatory downloading all the data into single site. Clustered the data locally and extract the representatives of these clusters send to global site where the clusters based on local representatives [41]. In [42] Distribution based Fuzzy Estimate Spectral Clustering (DFESC) technique used to diagnosis the cancer from protein sequence. jaccard similarity measures the similarity between each protein sequences.

Parallel clustering reduces the processing time and efficient data retrieval for large data. PPFC (Parallel Probabilistic Fuzzy Clustering) approach used to retrieve the protein sequence in efficient and effectively [43].

![Figure 1. Distributed clustering](image-url)
Distributed clustering is widely used to deal with very large and heterogeneous data sets which can’t be gathered centrally. Here the global cluster is good as the best Centralized Clustering Algorithm (CCA), limited collection of local model into single site. Data are partitioned to a number of processor which find local cluster and send back to superior process which uses DBSCAN to combine local cluster to global. In distributed memory system, Parallel Markov Chain Monte Carlo (RPMC) algorithm used to find motifs by Gibbs sampling method. Parallel k-means algorithm used to reduce the map for disk based data which is run in distributed environment [44].

<table>
<thead>
<tr>
<th>Sequence Repeats</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tandem or Dispersed Repeats</td>
<td>Consecutive patterns or a pattern. Ex. TGCATAGCTCATCCTCATGGA is CAT repeat 3 time consecutely.</td>
</tr>
<tr>
<td>Pattern-based or profile-based</td>
<td>Pattern of the motif in the first approach and the locations of the motif is later</td>
</tr>
<tr>
<td>Single or multiple sequence search</td>
<td>Repeat sequence in a single sequence or conserved repeats</td>
</tr>
<tr>
<td>Exact or approximate repeats</td>
<td>An exact repeat or repeats that have at most a specified Hamming or edit distance between them</td>
</tr>
<tr>
<td>Simple or structured repeats</td>
<td>Simple motif or a structured DNA motif</td>
</tr>
<tr>
<td>Probabilistic or combinatorial methods</td>
<td>All repeats at the expense of increased computation time.</td>
</tr>
</tbody>
</table>

Nucleotide sequence data are being produced. Exact clustering algorithms, such as hierarchical clustering rarely suitable for in terms of run time and memory usage. HPC-CLUST, an optimized cluster large number of pre-aligned DNA sequences by running on distributed computing hardware. It allocates both memory and computing resources efficiently, and can process more than a million sequences in a few hours on a small cluster [45].

To predict the protein sequence by using the distributed clustering concept is adopted for huge data set, segment the large data, reduce the space and time, as well as improve the accuracy of result.

V. CONCLUSION

In this survey, Most of the work carried out for protein sequence by clustering with soft computing methodologies. We propose the distributed clustering algorithm has been solving the troubles linked for huge data set. Many researchers chooses soft computing techniques and still many researches that are being going on in this techniques. Especially fuzzy logic is becoming so suitable for this type of sequence prediction in protein. The protein has huge biological molecules that contain amino and acid group. We have illustrated our discussion with a number of potential applications from different domains. For heterogeneous way, the distributed clustering is very crucial. Further work will consider the distributed clustering with soft computing techniques to predict the protein sequence and structure of an unknown sequence. We demonstrate the screening test such as True positive, True Negative, False Positive, False Negative and noticed that the cluster of huge protein sequence data set using distributed techniques instead of alignment method to reduce the time of execution and improve the efficiency with.
ACKNOWLEDGMENTS
The authors would like to thank the Institution for Selvam College of Technology, Namakkal, and Kongu Engineering College, Perundurai, Tamil Nadu, India for much assistance to the publication of this paper.

REFERENCES
2. O. Sasson and M. Linial, Protein Clustering And Classification, The Hebrew University of Jerusalem, Israel.


