# Structure Based Protein Multiple Sequence Alignment Algorithm on a Parallel System

Muhammad Ishaq Afridi and Yin Gui Sheng

***Abstract*—To enhance the speed and efficiency of structure based algorithms for protein Multiple Sequence alignment we need parallel processing. Secondary and tertiary structure of proteins are more important and informative then its primary structure. We get more accurate information's about proteins when we conduct structure based matching.**

**In this paper we have to focus on how to use parallel computing in terms of structure based algorithms. In case of protein structure based multiple sequence alignment several searches and structure matching are involved which require a lot of time and processing speed. If we use a parallel computer cluster or Grid then we can reduce processing time and get an optimal result. It will predict the function of protein from its structure.**

***Index Terms*— APDB (A with PDB is used for ADP), BLAST, Cluster based on Mesh topology, DLT (Divisible load theory), FASTA, expresso, MPICH-G2, MPICH, PipeAlign, PRALINE, RMSD, Structure based Protein MSA, SGE, T-coffee, Taylor's method.**

## I. Introduction

Protein structure is very complex and we spent months and years for correct protein structure prediction through experimentation, so we use computational method for quick structure prediction [1]. Structure based alignment is an integral part of homology modeling and threading [2]. Homology mean with same sequences, either in the same species called as Paralogs or different species called as orthologs. Threading mean to align ("thread") a protein sequence to a known structural special locus ("motif"). In structure based multiple sequence alignment the number of structure comparison is very large and there is a need of efficient methods for this.

The comparison of 3D (three dimensional) protein structure involves complex computation and the use of parallel processing technique for quick and optimal alignment. Structure comparison algorithms are used to identify a set of residue equivalences between two proteins based on their three dimensional coordinates [3].

The extent to which two structures align is to measure the root mean square deviation (RMSD) [4]. To know the exact position of alpha carbons in amino acid chains of two protein sequences. Structure alignment is performed in two steps. First ordinary sequence alignment for example through BLAST (Basic local alignment search tool) or FASTA (Fast alignment) [5] and then the 3D structure comparison with various kinds of protein 3D databases [6]-[7]. In case of multiple sequence alignment this task need a lot of processing speed.

Parallel computing is the most efficient solution to handle this complex time consuming comparison. It can locally use a computer cluster based on Mesh topology [6]-[8] to perform stand alone structure based multiple sequence alignment or it can use hierarchical Grid structure when it perform web based alignment [9].

It can perform iteration of a single algorithm for more optimal results. Parallel computing is also helpful in iteration of structure based multiple sequence alignment algorithms. The use hashing method, to generate a hashing table to speed up comparison is more efficient.

## II. Structure Based Protein Multiple Sequence Alignment

Structure based alignment is more informative and give us more efficient information about Proteins. 3D structure is more conserved then simple amino acid sequence. Tertiary protein structure evolves more slowly than primary structure [10]. So it is possible to improve the accuracy of our alignment by including information's about the three dimensional structure of protein [11].

The famous K2 algorithm successfully hybridizes a fast vector-based SSE alignment technique with a lower, but reliable, GA (Genetic algorithms) that aligns the amino acid positions [12].

We can use PRALINE [13] (a kind of multiple sequence alignment algorithm), the T-coffee[1] module expresso [14] and Pipe Align[2] [15]. Using expresso at T-coffee website [14], usually we submit a series of sequences in FASTA format. Each sequence is automatically searched by BLAST against the Protein Data Bank (PDB) database, and matches are used to provide a template to guide the creation of the multiple sequence alignment [16].

From structural information we can assess the accuracy of multiple alignments after it has been made and also it can assess the quality of a protein multiple alignment [17]. We must know accession number of at least two proteins.

Accession number can be found by performing BLASTP[3] (search protein database using a protein query) at NCBI (National center for biotechnology information) [18],

Manuscript received October 27, 2011; revised November 28, 2011. This work was supported in part by the Higher Education Commission of Pakistan and China Scholarship Council.

Muhammad Ishaq Afridi and Yin Gui Sheng are with the College of Computer Science and Technology Harbin Engineering University, Harbin, Heilongjiang, China. (E-mail: ishaqafridipk@gmail.com; yinguisheng@hrbeu.edu.cn)

[1] An algorithm that Combine sequence and structure of Protein

[2] A new toolkit for protein family analysis

[3] Search protein databases at NCBI and is faster than BLASTN by applying novel sequence similarity method.

restricting the output to protein databank (PDB) [17]. After that we perform multiple sequence alignment and input the result of this alignment to the APDB server at the T-coffee website. A with PDB is used for adenosine di-phosphate (ADP) with accession number. In this case we use PDB accession number in place of the name.

The output provide an analysis of the quality of the alignment on the basis of all pair wise comparisons of those sequences having structure as well as average quality assessment for each protein[2].

Structural alignment allows the superposition of one protein structure on to the other after rigid rotation and/ or translation. The structure based alignment is also helpful in protein folding, to share the same fold or the arrangement of α-helices and/or β-sheets within a protein structure [19].

The main approach to find how well two structures align is to measure the root mean square deviation (RMSD). The RMSD is a measure of how closely the alpha carbons of two amino residues are positioned [4]. In Parallelization of our structure based multiple sequence alignment we have to adopt divide and conquer method to get fast and accurate alignment.

## III. Mesh Based Cluster for Parallelization of Structure Based Protein Multiple Sequence Alignment Algorithm

In bus topology we employ divisible load theory (DLT) [20] to implement our algorithm.

The use of Mesh topology is preferred to implement our structure based protein multiple sequence alignment [6]. Structure alignment involves complex computation. So we have to use mesh architecture for this purpose. The independent computation of matrix element is exploited in order to distribute computation among several processors in the mesh [21].

The physical topological arrangement of processors, offer natural advantage to handle multiple sequences in a concurrent fashion. We actually measure the root mean square deviation of different proteins to assess their similarity so Mesh topology is more suitable in this case although it has some demerits.

In our proposed parallel system there are many rows and every row of processors is allowed a set of sequences. We use clustering strategy for structure based multiple sequence alignment [4]. It is simple to implement. The alternative to clustering method is Taylor's method [22].

In Taylor method we obtain similarity scores through smith-waterman algorithm and order the sequences into a cluster by decreasing similarity scores. We use DLT [20] to partition the computational space among the processors in the mesh to enable simultaneous processing in order to achieve higher speed up.

Assume that the mesh architecture as a tightly coupled with no communication delays structure compromising NxM processing nodes as shown in the figure 1. Give name to each row as Ri, i=1, 2…., N, and the processors on each row as Pk, k=1….M.

Each row has a master node that coordinates the activities of the processors in that row [6]. Assume that a process that start an instance of MSA (multiple sequence alignment) retrieves a set of sequences to be aligned from a database and injects to the mesh following the below mentioned distribution strategy [23]-[24].

The main sequence pool "O" is divided equally among all the rows so that each row handles the same number of sequences, say, "Q". Thus, each row of the mesh will be processing a subset of sequences and arrive at an alignment that is best for that subset.

All master nodes from all rows collect and combine the result to give an optimal alignment. Consolidate all locally aligned subset of sequences by any heuristic approach like Taylor's or Improved Taylor's method.

Align all locally aligned sequences from each row to improve overall score and then store it for comparing the quality of the output from the next iteration [6].

The alignment of multiple sequences is based on the next highest score in protein database and it is not an optimal alignment. New sequence is aligned with previous aligned sequence set.

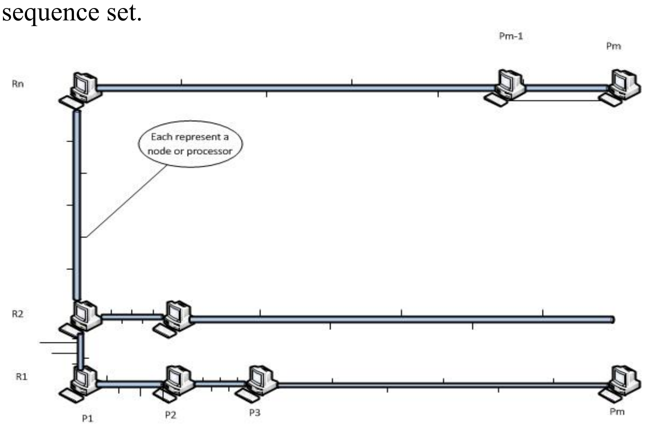

Fig. 1. NxM Mesh structure[12]

## IV. Application of Heterogeneous Hierarchical Grid

When we use expresso at T-coffee website [14]-[17] for structure based protein multiple sequence alignment then it can also utilize the facility of heterogeneous hierarchical Grid [25] computing by applying the same mesh strategy in each individual cluster within the Grid. Usually we prefer UNIX PC clusters.

Heterogonous mean that Grid resources belong to different administrative domains, run different software and have different access control strategies and the connecting network are also different performance and structure wise so this is termed as a hierarchical Grid[26]. The Grid resources are geographically distributed.

Each cluster in the hierarchical Grid has a mesh topology as explained earlier and connected by *Myrinet* (Intra cluster connection) [4] or Ethernet switch [8]. The intra cluster bandwidth is very high but inter cluster bandwidth is very low. The packet traffic between clusters and within cluster is controlled by different application like MPICH-G2 and MPICH.

[4] The intra-cluster connection speed is very high up to 250MB/s, but inter-cluster connection is comparatively slow up to 80MB/s.

The normal application bandwidth inside the cluster is very high while the inter cluster communication is varying depend on different conditions of connecting networks. We run an application to send or receive data packages between the clusters [6]-[24]. Sometime we can control the size and frequency of data packages.

The software architecture of hierarchical Grid can be divided into two layers. The upper layer is the MPICH-G2 layers that run on the control node of each cluster and allow inter cluster communication.

The lower layer is called MPICH(A grid enable implementation of the message passing interface) and is run on all nodes within a cluster and allows intra cluster communication [9]-[27].

Each cluster has a Sun Grid Engine (SGE) installed. SGE is Distributed resource management software. It can allocate parallel tasks from control nodes to the execution nodes inside a cluster and work as a specialized type of application or operating system.

Parallel processes can communicate via master node in a row when they are in the same row of processors or through MPICH when they are inside a single cluster or through MPICH-G2 when they are in different clusters.

In mesh topology some nodes are assigned for control such as master nodes in each row of mesh topology and the other nodes are called execution nodes [6]-[25].

A master node in each row is actually responsible for the collection of result or optimal single pair wise alignment and control and coordination at that row.

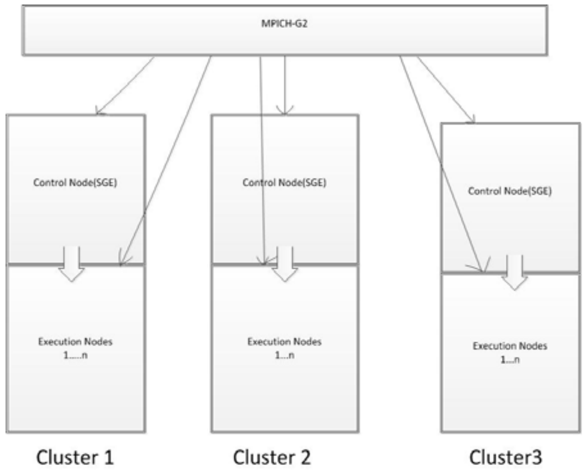

Fig. 2. The hierarchical parallel programming environment consisting of MPICH-G2, MPICH, and Sun Grid Engine [4].

## V. Implementation

Lack of resources and proper guidance prevents me from practical implementation. We don't have Bioinformatics laboratory or such kind of dedicated parallel Computer cluster.

I am a Self learned and self motivated scholar. We face very weak research and study environment. We are not entitled to any research facility or any research project. I am a fluent mandarin speaker, after detail discussion with our teaching staff; I understand that the background knowledge of our college professor is at a very low level. I cannot check its performance practically. Also there is no senior Research Professor or Scholar of Bioinformatics discipline in our University that can facilitate me. I try to implement this strategy and get practical results.

## VI. Discussion

The structure of Protein is very complex. So Protein structural alignment is more time consuming and complicated. We use various kinds of algorithms and programs for correct computational structural prediction and structural alignment. Structural alignment gives us more comprehensive and accurate information's as compared to simple word based alignment. The inference based on structural alignment will be more accurate and precise.

To overcome the computational burden in case of structural alignment we can use cluster or grid facility. To run our algorithm on computer cluster for fast results or it can also employee a Grid or Hierarchical Grid. Some Alignment algorithms (Programs) are available in downloadable version for standalone or independent use (BLAST) and also can be used as a web based version (expresso).

Usually we use a computer cluster in case of local or stand alone alignment program. Grid technology is more useful in case of web based structural alignment.

## Acknowledgment

This work and study is supported by Ministry of Education, Government of Pakistan and China Scholarship council. I am also thankful to my supervisor Professor Dr. Yin Gui Sheng for his kind support and help.

## References

[1] Krishnan S P, Liang S S, Veeravalli B. Towards high performance computing for molecular structure prediction using IBM Cell Broadband Engine--an implementation perspective [J]. BMC bioinformatics: 2010, 11 Suppl 1S36.

[2] Pevsner J. Bioinformatics and functional genomics [M]. Hoboken, N.J.: Wiley-Blackwell, 2009.

[3] Wohlers I, Domingues F S, Klau G W. Towards optimal alignment of protein structure distance matrices [J]. Bioinformatics (Oxford, England): 2010, 26(18): 2273-2280.

[4] Zomaya, A. Y. 2006. *John Wiley Series on Parallel and Distributed Computing* Database on-line. Available from Scopus.

[5] Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods in enzymology* 183: 63-98. Database on-line. Available from Scopus.

[6] Low, D. H. P., B. Veeravalli, and D. A. Bader. 2007. On the design of high-performance algorithms for aligning multiple protein sequences on mesh-based multiprocessor architectures. *Journal of Parallel and Distributed Computing* 67, no. 9: 1007-1017. Database on-line. Available from Scopus.

[7] Lin, X., Y. Lu, J. Deogun, and S. Goddard. 2007. Real-time divisible load scheduling with different processor available times. In Database on-line. Available from Scopus.

[8] Baldridgr, K., P. E. Bourne. 2003. The New Biology and the Grid, Grid Computing Making the Global Infrastructure a Reality Database on-line. Available from Scopus.

[9] Bertil Schmidt, Chen C, Liu W, Hierarchical grid computing for high performance Bioinformatics. Grid Computing for Bioinformatics and computational biology: a publication of the John Wiley: 2008,

[10] Fiser A. Template-based protein structure modeling [J]. Methods in molecular biology (Clifton, N.J.): 2010, 67373-94.

[11] Berger, M. P., P. J. Munson. 1991. A novel randomized iterative strategy for aligning multiple protein sequences. Computer Applications in the Biosciences 7, no. 4: 479-484. Database on-line. Available from Scopus.

[12] Fogel G, Corne D. Evolutionary computation in bioinformatics [M]. San Francisco, Calif.; Oxford: Morgan Kaufmann; Elsevier Science, 2003.

[13] IBIVU center for bioinformatics VU, PRALINE multiple sequence alignment algorithms.

[14] Expresso and 3D coffee (T-coffee) combining sequences and structures. Dr. Cedric Notredame, PhD. Group Leader Comparative Bioinformatics Group Bioinformatics and Genomics Programme Center for Genomic Regulation (CRG) Dr Aiguader, 88 08003 Barcelona Spain

[15] Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., et al. (2003). PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Research, 31(13), 3829-3832.

[16] Protein Data Bank PDB-101 an Information Portal to Biological Macromolecular Structures. PDB ftp archives at RSCB[5] PDB web link.

[17] Di Tommaso P, Orobitg M, Guirado F, Cloud-Coffee: implementation of a parallel consistency-based multiple alignment algorithm in the T-Coffee package and its benchmarking on the Amazon Elastic-Cloud [J]. Bioinformatics (Oxford, England): 2010, 26(15): 1903-1904.

[18] National Center for Biotechnology Information, U.S. National Library of Medicine8600 Rockville Pike, BethesdaMD, 20894USA

[19] Rozwarski D A, Gronenborn A M, Clore G M,Structural comparisons among the short-chain helical cytokines[J]. Structure (London, England: 1993): 1994, 2(3): 159-173.

[20] Yao, J. 2007. *Scheduling network applications based on divisible load theory (DLT):* 131. Database on-line. Available from Scopus.

[21] Robertazzi, T. G., D. Yu. 2006. Multi-Source Grid Scheduling for Divisible Loads. *IEEE 40th annual conference on information sciences and systems:* 188-191. Database on-line. Available from Scopus.

[22] Eidhammer, I., I. Jonassen, and W. R. Taylor. 2004. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis* Database on-line. Available from Scopus.

[23] Bharadwaj, V., D. Ghose, V. Mani, and T. G. Robertazzi. 1996. *Scheduling Divisible Loads in Parallel and Distributed Systems* Database on-line. Available from Scopus.

[24] Zhang, X., Y. Yan. 1995. Modeling and characterizing parallel computing performance on heterogeneous networks of workstations. *Proceedings of Seventh IEEE Symposium on Parallel and Distributed Processing:* 25-34. Database on-line. Available from Scopus.

[25] Singh A, Chen C, Liu W, A hybrid computational grid architecture for comparative genomics[J]. IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society: 2008, 12(2): 218-225.

[26] Lucchese, F. d. O., E. J. Huerta Yero, F. S. Sambatti, and M. A. A. Henriques. 2006. An adaptive scheduler for grids. *Journal of Grid Computing* 4, no. 1: 1-17. Database on-line. Available from Scopus.

[27] Bergeron B P. Bioinformatics computing [M]. Upper Saddle River, NJ: Prentice Hall, 2003.

**Muhammad Ishaq Afridi** was born in Bara Khyber Agency FATA Pakistan. He is currently with College of Computer Science and Technology Harbin Engineering University China as a PhD scholar since September 2008. He worked for two years as a Lecturer in Computer Science at Virtual University of Pakistan (2006-08). He got my M.Sc. IT degree from Kohat University of Science and Technology Pakistan in 2005.

[5] RSCB is managed by two members, Rutgers and UCSD, and is funded by NSF, NIGMS, DOE, NLM, NCI, NINDS, and NIDDK.