Grid Environment for Dengue Virus Protein Analysis

C. Murugananthi[#1], D. Ramyachitra[*2]

[#]*Department of Computer Science, Bharathiar University*
Coimbatore, Tamil Nadu, India
[1]murugananthiselvi3@gmail.com

[*]*Department of Computer Science, Bharathiar University*
Coimbatore, Tamil Nadu, India
[2]jaichitra1@yahoo.co.in

Abstract — **Bioinformatics applications such as protein analysis require a coordination of very large databases, tools and methods. This is because thousands and lakhs of proteins being deposited into databases by the researchers which results in wide range of database search when one wants to predict function, structure and protein sequence similarities. A Grid environment can be viewed as a virtual computing architecture that offers the facility to solve high throughput problems by taking advantage of many computers geographically dispersed and connected by a network. Bioinformatics applications gain benefit in such a distributed environment in terms of increased efficiency, reliability and availability of computational resources. Bioinformatics is mainly concerned with the analysis and processing of data related to the biology of humans and other species. Dengue virus infection is a serious health problem infecting billions of people worldwide. This paper gives an analysis of the dengue virus proteins using grid technology.**

Keywords— **Grid Computing, Bioinformatics, Protein Analysis, Dengue Virus Protein.**

I. Introduction

Grid Computing is a high performance computing infrastructure with large-scale pooling of resources that may be processing cycles, storage or data that allows sharing of various distributed resources across many administrative domains and used to solve large scale computational problems [1]. In a grid environment, users can access the resources without knowing where they are physically located [2]. Grid provides the facility to combine huge amounts of computing resources which are geographically dispersed to undertake large problems and workloads as if all the resources and servers are located in a single site. Grid environment solve the large scale technical or scientific problem that requires a large number of computer processing cycles or access to vast amounts of data [3]. Grids integrate computation, networking, information and communication to offer a virtual platform for computation and data management [4].

Study of the evolution of organisms or species is vital for various biological applications. Researchers have developed several models and techniques for the efficient analysis of biological data [5]. Various computational methods, techniques and tools have been developed for building phylogenetic or evolutionary trees for a set of sequences, predicting secondary structure and motif patterns of particular proteins [6]. These methods and tools compare the sequences with the large databases and perform high end computational calculations [7]. The great resource requirements of life science combined with the huge number of data-parallel applications in this field and the availability of high performance grid computing infrastructure lead to the openings for emerging grid-enabled life science applications [8]. The rest of the paper is organized as follows. Section 2 discusses the grid applications in bioinformatics. Section 3 presents protein analysis and protein databases. Section 4 discusses the dengue protein analysis. Section 5 describes grid infrastructure for dengue protein analysis. Finally, section 6 gives the conclusion.

II. Bioinformatics on Grid

Bioinformatics is a discipline of biological research involving the combination of computers, databases and software tools. It involves the development of algorithms to utilize and manage biological databases in knowledge-based analysis [9]. Study of the evolution of different species or organisms is important to biologists since it has practical applications including drug discovery, vaccine development and finding the function of the proteins [10].

Many applications in bioinformatics produce massive amount of data over very short periods of time, and this requires considerable computational and storage capabilities [11]. Scientific applications require months or years of data processing to produce results [12]. Some drug design problems may involve the job of screening 180,000 compounds that may need up to 540,000 hours or over years of execution time on a single desktop computer [13]. These factors re-enforce the need for grid technology because of the role that it can play in facilitating research in the life sciences. The range of problems in life sciences will also lead to the development of a wide range of grid-enabled technologies to play a variety of roles, for example computational grids [4] can provide number crunching capabilities, and data grid [14] delivers a secure infrastructure for all data management needs.

III. Protein Analysis

Proteins are large organic molecules composed of amino acids arranged in a linear chain and held together by peptide bonds. They constitute an essential part of organisms, participating in all processes within and between cells [15]. Functions of proteins are determined by its structure.

Protein structures are described at different levels. Primary structure is a linear sequence of amino acids. In secondary structure, segments of amino acids often fold into stable structures that include alpha helices and beta pleated sheets. Tertiary structure is the packing of alpha-helices, beta-sheets, turns and random coils. Quaternary structure describes the spatial organization of the chains [9].

Biological databanks cover nucleic acid and protein sequences, macromolecular structures and functions. SWISS-PROT and PIR are protein sequence databases. PROSITE, Pfam, PRINTS are secondary databases. Protein Databank (PDB) is a worldwide repository for the processing and distribution of 3D biological macromolecular structure data and it is maintained by the RCSB. SCOP and CATH are databases that offer hierarchical classification of entire PDB according to the folding patterns of the proteins [16].

Availability of three-dimensional structure of any biomolecule is vital to understand its function and structure at the molecular level and to undertake any molecular design tasks [17]. Motifs are components of a more fundamental unit of structure and function [18]. Primary structure data is basis for prediction of secondary and tertiary structures of proteins [9]. The secondary structure elements helix, sheet, turn and coil constitute the building blocks of the folding proteins.

The aim of secondary structure prediction is to provide information and location of helices, strands and random coil segments within a protein from its amino acid sequence data. Therefore, prediction of the secondary structure of a protein is used as the first step in an attempt toward predicting its tertiary structure [16].

IV. Dengue Virus Protein Analysis

Dengue virus is a global threat to health around the world. Most infected people experience dengue fever, with fever and rashes and very bad headaches that last a week. Researchers are studying the dengue virus to develop drugs to cure infection and vaccines to prevent infection before it starts [19].

Dengue fever is caused by four closely related viruses namely DEN-1, DEN-2, DEN-3 and DEN-4 [20]. Dengue virus is a single strand of RNA. It is denoted as positive sense RNA because it can be directly translated into proteins [21]. The viral genome encodes ten genes.

The genome is converted into a single, long polypeptide and then cut into ten proteins. Among these ten proteins, three of them are structural proteins that are in the virus surface and deliver the RNA to host cells and seven of them are nonstructural proteins that direct the creation of new viruses once the virus gets inside the cell [19]. Capsid Protein, Membrane Protein, Envelope Protein are three structural proteins and NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5 are seven nonstructural proteins [22].

A. *Methods and Materials*

Dengue virus proteins and their organisms, FASTA sequences were downloaded from the PDB [23] and UniProtKB [24] databases. The secondary structure of dengue proteins were retrieved using the structure prediction tool SOPMA [25].

Structure features were analyzed using PDB database. FASTA sequence given as input and secondary structure of proteins were found. Sequence motifs for each protein were identified using motif prediction tool GenomeNet [26]. An antigenic site for each protein was determined using Immunomedicine Group tool [27].

V. Dengue Virus Protein Analysis on Grid

We analyzed the dengue virus proteins for finding their protein types (ten protein types) and its sequences. Also, we analyzed the secondary structure and motif patterns of these proteins for predicting its tertiary structure and functions. This analysis needs extensive database search and large amount of computational power. So this analysis can be implemented in grid environment for fast execution and better performance.

Fig. 1 shows how protein analysis is implemented in the grid environment.

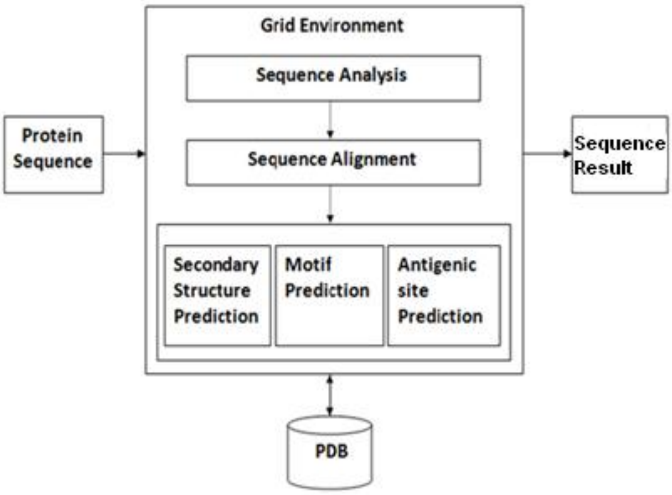

Fig. 1 Architectural framework for protein analysis on grid

Fig. 2 shows the grid infrastructure for dengue protein analysis. This infrastructure consists of the following components.

- Resource Broker
- Information Service
- Scheduler
- Computing Elements

A. *Resource Broker*

Resource broker manages the interactions between the components of the system. It accepts protein input sequence from the user and sends the results back to the user. It provides the user request to the scheduler and collects the result. Resource broker manages the interactions between the components of the system. It accepts protein input sequence from the user and sends the results back to the user. It provides the user request to the scheduler and collects the result.

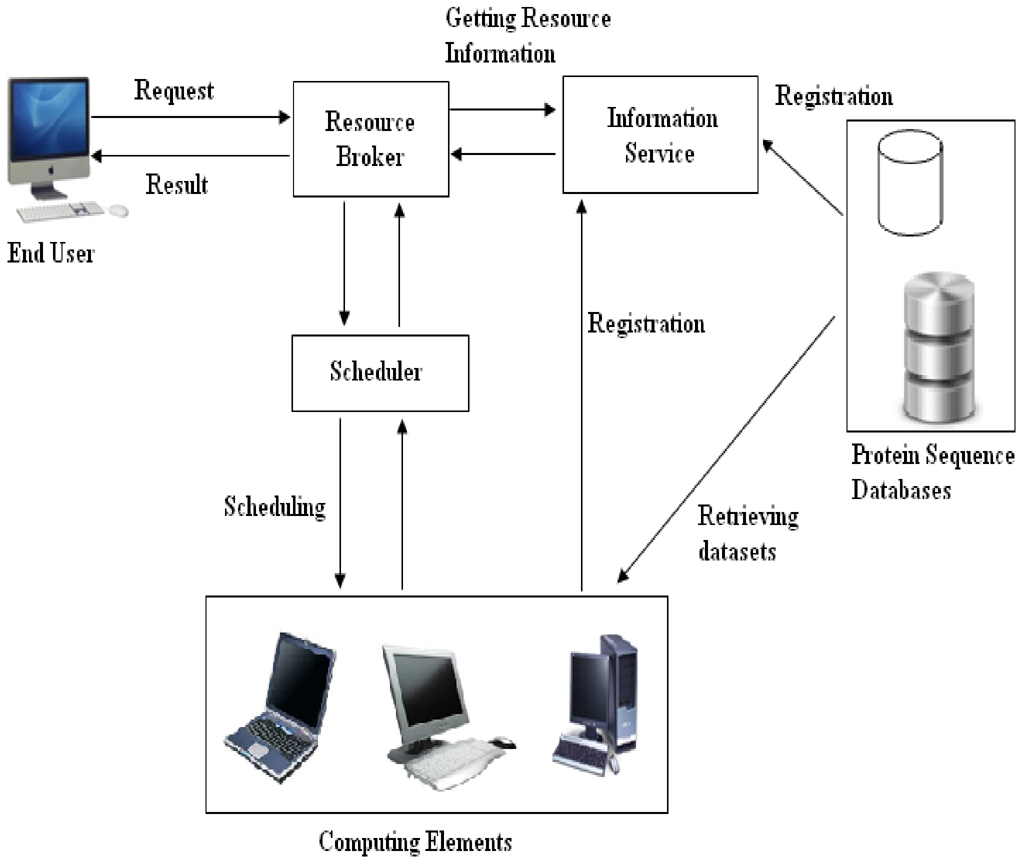

Fig. 2 Grid infrastructure for dengue protein analysis

B. *Information Service*

It collects information about all the resources including computing elements and protein databases. All computing elements register their information such as availability, capacity and processing speed to information service. Resource broker collect the resource information from this component.

C. *Scheduler*

Scheduler schedules the requests to the suitable processing element and monitors the progress of the job. This component is called by the broker and its main work is split the whole analysis work of one protein into sub problems and distributing or balancing the computational load to the Computational Elements. In this dengue protein analysis, the user provides input protein sequence in the form of request to the broker and each analysis such as secondary structure prediction; motif identification and antigen site determination are scheduled to independent systems based on its availability and capacity.

D. *Computational Elements*

This component performs the execution of the job and produces the result. It downloads the required protein data set from the databases and processes the jobs. Computational elements wrap the Bioinformatics tools and algorithms as Web Services. Here it downloads the secondary structure databases and motif databases for performing the protein analysis.

Dengue protein analysis may need up to weeks or months of execution time on single computer system. But using grid technology, protein analysis can be completed in a short time. End users can run the applications using grid infrastructure by just registering at the grid portal. They no need to have in depth Grid knowledge and can benefit from a wide range of computing resources that undertake larger problem sizes.

VI. CONCLUSION

Protein analysis involves wide range of database searches. The grid infrastructure can be used for executing concurrent protein analysis in a distributed system. Grid technology can help to reduce the execution time of the dengue protein analysis. Based on increment of the grid resources, running time of the application can also be reduced. Future research involves the improvement of the grid infrastructure for the diagnosis and development of the vaccine for the dengue virus. In the proposed system, scheduler split the problems and distributes the workloads. In future, separate computational component will be created and policies will be generated for splitting larger size problems.

REFERENCES

[1] Ian Foster, Carl Kesselman (eds.),"The Grid: Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, 2004.

[2] I. Foster, C. Kesselman, and S. Tuecke, The anatomy of the grid: enabling scalable virtual organizations, Int. J. Supercomput. Appl., 15, 200-222 (2001).

[3] M. Hemamalini, "Review on Grid Task Scheduling in Distributed Heterogeneous Environment", International Journal of Computer Applications (0975 – 8887) Volume 40– No.2, February 2012.

[4] Fran Berman, Geoffrey Fox, Tony Hey, "The Grid: Past, Present, future", Grid Computing-Making the Global Infrastructure a Reality, Wiley series in Communications Networking and Distributed systems, 2004.

[5] Arun Krishnan, "A survey of life science applications on the grid", New Generation Computing 22(2004)111-126, Ohmaha, Ltd, and springer-Verlag.

[6] EL-Ghazali Jalbi, Albert Y. Zomaya, "Grid Computing for Bioinformatics and Computational Biology", wiley publications.

[7] Yadnyesh Joshi, Sathish Vadhiyar, "Analysis of DNA sequence transformations on grids", J. Parallel Distrib. Comput. 69 (2009) 80_90.

[8] Dr.G. Raju, Manjula. K.A., "A Study on Applications of Grid Computing in Bioinformatics", IJCA Special Issue on "Computer aided soft Computing Techniques for Imaging and Biomedical Applications" CASCT, 2010.

[9] P. Narayanan, "Bioinformatics A Primer", New Age International (P) Ltd., Publishers, 2006.

[10] Understanding Evolution, http://evolution.berkeley.edu

[11] L. Ferreira, V. Berstis, J. Armstrong, M. Kendzierski, A. Neukoetter, M. Takagi, R. Bing-Wo, A. Amir, R. Murakawa, O. Hernandez, J. Magowan, and N. Bieberstein, "Introduction to Grid Computing with Globus", IBM Red Book, 2003.

[12] Y. Sun, S. Zhao, H. Yu, G. Gao, and J.Luo, "ABCGrid: Application for Bioinformatics Computing Grid", Bioinformatics, 23(9), pp 1175-1177 (2007).

[13] R. Buyya, K. Branson, J. Giddy, and D. Absramson, "The Virtual Laboratory: A toolset to enable distributed molecular modeling for drug design on the world-wide grid, J. Concurr. Comput. Prac. Exp., 15, 1-25 (2003).

[14] A. Chervenak, I. Foster, C. Kesselman, C. Salisbury, and s. tuecke, The data grid: towards an architecture for the distributed management and analysis of large scientific datasets, J. Network Comput. App., 33, 187-200 (2002).

[15] Fotis E. Psomopoulos, Pericles A. Mitkas, "Bioinformatics algorithm development for Grid environments", The Journal of Systems and Software 83 (2010) 1249–1257.

[16] P. Shanmugavel, "Trends in Bioinformatics", Pointer Publishers, 2006.

[17] J. C. Setubal and J. Meidanis, "Introduction to Computational Molecular Biology", Brooks/Cole Publishing Company, 1997.

[18] S.B. Needleman and C.D. Wunsh, "A general method applicable to the search of similarities of amino acid sequences of two proteins", J. Mol. Biol., 48(2), 443-453 (1998).

[19] RCSB, Protein Data Bank, Availabe: http://www.rcsb.org/pdb/101/motm.do?momID=103.

[20] Jacqueline L Deen, Eva Harris, Bridget Wills, Angel Balmaseda, Samantha Nadia Hammond, Crisanta Rocha, Nguyen Minh Dung, Nguyen Thanh Hung, Tran Tinh Hien, Jeremy J Farrar, "The WHO dengue classification and case definitions: time for a reassessment", International Vaccine Institute, Lancet 2006; 368: 170–73.

[21] Justin G Julander, Stuart T Perry, Sujan Shresta, "Important advances in the field of anti-dengue virus research", Antiviral Chemistry & Chemotherapy 2011; 1:105-116 (doi: 10.3851/IMP1690).

[22] Nivedita Gupta, Sakshi Srivastava, Amita Jain, Umesh C. Chaturvedi, "Dengue in India", Indian J Med Res 136, September 2012, pp 373-390.

[23] http://www.rcsb.org/pdb/results/results.do?qrid=73FCB85C&tabtoshow=Current

[24] http://www.uniprot.org/uniprot/?query=dengue+virus+proteins&offset=25&sort=score

[25] http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html

[26] http://www.genome.jp/tools/motif/

[27] http://imed.med.ucm.es/Tools/antigenic.pl