A Study on Applications of Grid Computing in Bioinformatics

Manjula.K.A
Department of Information Technology
Kannur University
Kerala, India

Dr.G.Raju
Department of Information Technology
Kannur University
Kerala, India

ABSTRACT
Huge volume of biological data, which is heterogeneous, autonomous and dynamic in nature, is being produced at a rapid pace throughout the community. Along with the increase in biological data, many tools are being designed to analyze them by different research groups. Integration of these biological data and tools is becoming one of the major topics in the bioinformatics community. Grid computing is emerging as a key infrastructure for a wide range of disciplines for resource sharing and collaboration over wide and open networks. Provided with fundamental mechanisms of access to distributed databases and tools, the grid is an ideal candidate for the integration of bioinformatics applications.

Keywords
Bioinformatics, Grid Computing, BioGrid.

1. INTRODUCTION
Research in the area of bioinformatics has grown with each passing day in recent years as demands for more computing power increased. The solutions to these demands usually involve using parallelism techniques. Although Cluster environments can reduce the execution time and increase alignment efficiency, they may not be a good solution when aligning very large genomic databases, which are of distributed nature. This is where Grid computing can play a significant role. Grid Computing can coordinate the resources of distributed virtual organizations and satisfy a great many computational demands [1]. In addition to integrating distributed resources, Grid Computing can reduce server idle times via management of integrated computing resources. Considering the capabilities of Grid Computing, this paper looks at the benefits of using Grid Computing in bioinformatics and how such a combination will enhance the applications of bioinformatics. In the following sections this paper briefly discusses about Grid Computing and bioinformatics, followed by the benefits of such combinations and finally looks at some of the initial works in this area.

2. GRID COMPUTING
The term, grid computing, has become one of the latest buzzwords in the IT industry. Grid is a system that coordinates resources that are not subject to centralized control using standard, open, general-purpose protocols and interfaces to deliver nontrivial qualities of service [2]. Grid computing can be thought of as distributed and large scale cluster computing and as a form of network distributed parallel processing. This innovative approach of computing leverages on existing IT infrastructure to optimize resources and manage data as well as computing workloads [3].

Grid is an infrastructure that involves the integrated and collaborative use of computers, networks, databases and scientific instruments owned and managed by multiple organizations. Grid computing is applying the resources of many computers in a network to a single problem at the same time - usually to a scientific or technical problem that requires a great number of computer processing cycles or access to large amounts of data (Figure 1). Grid computing requires the use of software that can divide and farm out pieces of a program to as many as several thousand computers. Grid middleware provides users with seamless computing ability and uniform access to resources in the heterogeneous grid environment [4].

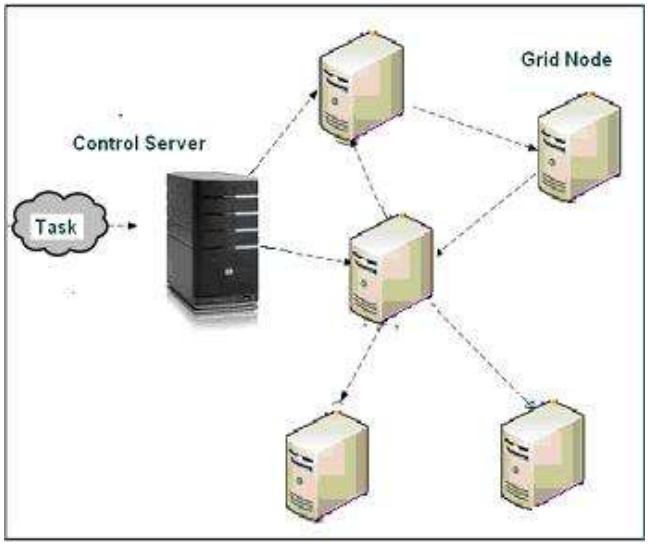

Figure 1. Grid Computing Structure

Grid computing appears to be a promising trend for three reasons [5]:

- Its ability to make more cost effective use of a given amount of computer resources.
- As a way to solve problems that can't be approached without an enormous amount of computing power.

- This suggests that the resources of many computers can be cooperatively and perhaps synergistically harnessed and managed as a collaboration toward a common objective.

Grid computing is becoming a critical component of science, business, and industry [6]. Grids could allow the analysis of huge investment portfolios in minutes instead of hours, significantly accelerate drug development, and reduce design times and defects. With computing cycles plentiful and inexpensive, practical grid computing would open the door to new models for compute utilities, a service similar to an electric utility in which a user buys computing time on-demand from a provider.

Larger bodies of scientific and engineering applications stands to benefit from grid computing, including molecular biology, weather forecasting, financial and mechanical modeling, immunology, circuit simulation, aircraft design, fluid mechanics, biophysics, biochemistry, biology, scientific instrumentation, drug design , tomography, high energy physics, data mining, financial analysis, nuclear simulations, material science, chemical engineering, environmental studies, climate modeling, neuroscience/brain activity analysis, structural analysis, mechanical CAD/CAM, and astrophysics.

3. BIOINFORMATICS

Bioinformatics is the combination of biology and information technology. This discipline encompasses computational tools and methods used to manage, analyze and manipulate large sets of biological data. Bioinformatics is essential for achieving so many complex tasks such as use of genomic information in understanding human diseases, identification of new molecular targets for drug discovery and in unraveling human evolution mysteries. Essentially, bioinformatics has three components [7]:

- The creation of databases, allowing the storage and management of large biological data sets.
- The development of algorithms and statistics, to determine relationships among members of large data sets.
- The use of these tools for the analysis and interpretation of various types of biological data, including DNA, RNA and protein sequences, protein structures, gene expression profiles, and biochemical pathways.

The combination of IT with biology proves to be of immense use mainly because of two reasons-

- First, many bioinformatics problems require the same task to be repeated millions of times. For example, comparing a new sequence to every other sequence stored in a database or comparing a group of sequences systematically to determine evolutionary relationships.
- Second, computers are required for their problem-solving power. Typical problems that might be addressed using bioinformatics could include solving the folding pathways of protein given its amino acid sequence, or deducing a biochemical pathway given a collection of RNA expression profiles.

There are useful bioinformatics tools which can be of great help to biologists. One such tool is BLAST (Basic Local Alignment Search Tool), which finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families [8]. The Figure 2 represents a sample BLAST result page scrolling through which provides information such as unique request ID (RID), query-database information, a link to taxonomy reports, a graphical display showing alignments to the query sequence, descriptions of sequences producing significant alignments, and pair wise alignments between the query sequence and each BLAST hit sequence.

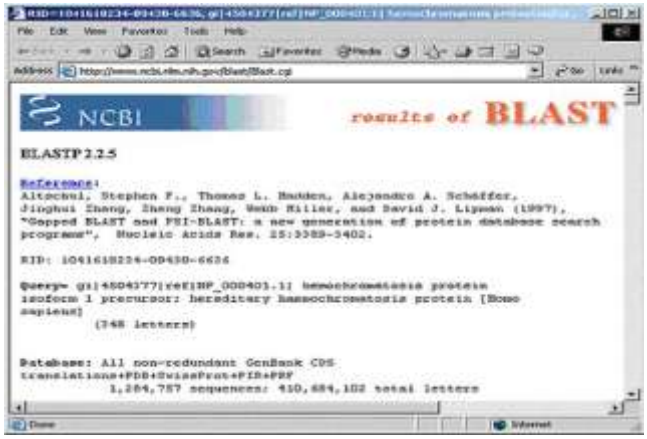

Figure 2. Execution of BLAST

Biological data are being produced at a phenomenal rate. Most common analyses have evolved to a larger scale, for example from the study of a single gene/protein to a whole genome/proteome, from a single metabolic pathway to Systems Biology. Hence, now Bioinformatics is requiring research infrastructures able to store very large biological data sets, complex and heterogeneous, and to make these data available for intensive scientific computing.

The future of bioinformatics is integration [9]. Facing to this growing need, we should take advantage of the new developments in computing such as distributed computing.

4. NEED FOR GRID COMPUTING IN BIOINFORMATICS

The bioinformatics research area is now faced with a mountain of ever-increasing and distributed information. For example, finding a single gene of the *Oryza sativa* (rice) genome one must spend weeks, if not months, wandering through approximately 40 million base pairs. These data are scattered in many data repositories. Thus, not only do we need an efficient tool to visualize and analyze DNA data, but the integration and exchange of information on a particular gene or coding regions from different international collaborative databases needs to be done in a careful, but robust manner as well. Along with the increase in biological data as seen in Table 1, many tools [Table2] are being designed to analyze them by different research groups [10]. Integration of these biological data and tools is becoming one of the major topics in the bioinformatics

community. One of the goals of the integration is to enable users to access bioinformatics tools and up-to-date biological data across multiple heterogeneous databases.

Table 1. Size of Some Biological Databases

Name	Nature	Rev.	Entries	Size(MB)
GenBank	Gene Sequence	153	56,620,500	224,000
EMBL	Gene Sequence	86	69,783,593	~100,000
Swiss-Prot	Protein Sequence	49.5	216,380	824
TrEMBL	Protein Sequence	32.5	2,807,081	6,347
PROSITE	Protein Signature	19.25	1,411	14
pFAM-A	Protein Signature	19.0	8,183	2,104
PDB	Protein Structure	Apr. 2006	36,121	23,316

Table 2. Example of BioInformatics Programs

Name	Algorithm	Input data
BLAST	Similarity	Gene/Protein Sequence
FASTA	Similarity	Gene/Protein Sequence
SSearch	Similarity	Gene/Protein Sequence
ClustalW	MSA	Protein Sequence
Multalin	MSA	Protein Sequence
PattInProt	Pattern/Profile	Sequence, Pattern, profile
GOR4	PSSP	Protein Sequence
SIMPA96	PSSP	Protein Sequence
SOPMA	PSSP	Protein Sequence

PSSP: Protein secondary structure prediction;

MSA: Multiple Sequence Alignment

Grid computing is emerging as a key infrastructure for a wide range of disciplines for resource sharing and collaboration over wide and open networks. Provided with fundamental mechanisms of access to distributed databases and tools, the grid is an ideal candidate for the integration of bioinformatics applications.

Massive computing resources are often required for large-scale bioinformatics analyses. Many bioinformatics tools such as BLAST and HMMER require powerful computers for better performance. However, small or mid-scale biological laboratories cannot afford powerful servers or dedicated clusters. Instead, they usually have dozens of personal computers and workstations which are often relatively idle. Grid technology provides an alternative approach for these laboratories to utilize distributed idle resources to meet the needs of computational capability.

Location transparency is an important feature of grid technology [11]. Users can access applications without being aware of where these packages are installed. The parallel version bioinformatics tools when used in grid environment can help us reduce the waiting time of alignment and improve performance of complex tasks such as sequence alignment.

The emerging grid computing technologies enable bioinformatics scientists to conduct their researches in a virtual laboratory, in which they share public databases, computational tools as well as their analysis workflows [12]. The implementation of existing bioinformatics applications on Grids represent a cost-effective alternative for addressing highly resource-demanding and data-intensive bioinformatics tasks [13].

5. BIOLOGICAL GRID PROJECTS AROUND THE WORLD

Given the possibilities of grid computing, it is no surprise that there is a huge interest in grid computing technology around the world. Many genomes have been sequenced and their annotation requires larger and larger databases. The storage and the exploitation of these genomes and of the huge flux of data coming from post-genomics put a quickly growing pressure on the computing tools and resources in the laboratories. The integration of a platform dedicated to biology into GRID opens up new perspectives in terms of computing resources and data storage and there are BioGrid projects and works proceeding around the world. EuroGrid BioGrid [14], Asia Pacific BioGrid [15], UK BioGrid [16], North Carolina (NC) BioGrid [17], and Singapore BioGrid [18] are some of them.

The EuroGrid BIO–GRID aims to develop intuitive user interfaces for selected packages from different areas of biomolecular research and compatibility interfaces with their databases. The result will be an integrated biomolecular toolkit that allows streamlined work processes, and a job execution component that makes all systems in the BIO–GRID available for simulation runs with a uniform and intuitive user interface.

The UK BioGrid is called MyGrid. MyGrid is an e-Science Grid project that aims to help biologists and bioinformaticians to perform workflow-based in silico experiments, and help to automate the management of such workflows through personalization, notification of change and publication of experiments.

The NC Grid is providing the computing data storage, and networking capabilities to support the genomics revolution. Members of the North Carolina Genomics and Bioinformatics Consortium are working with computer and networking companies to create the North Carolina Bioinformatics Grid [17].

An information network is being built at Indiana University for large data and computationally intensive applications in several sciences, using advanced data grid technologies. With national and international collaborations in physics, bioinformatics,

geology, and computer science, this will provide scientists access to local and globally distributed computing resources. The Singapore BioGrid is applying Clustal-G on Grid system to implement sequence alignment [18].

These projects around the world are all using grid technology for bioinformatics. The BioGrid technology is turning out to be the most popular and effective in solving biology problems.

6. CONCLUSION

Grid computing can be a good solution for the challenges faced in bioinformatics field. Since bioinformatics demands more computing power, integration of distributed, huge and complex data as well as applications of heterogeneous networks, Grid computing environments can be a right choice. The integration of a platform dedicated to biology into GRID opens up new avenues in terms of computing resources and data storage. There are BioGrid projects like EuroGrid BioGrid , Asia Pacific BioGrid, UK BioGrid , North Carolina BioGrid which are aimed to contribute to the field of biological computing. The BioGrid technology is opening up new perspectives for bioinformatics.

7. REFERENCES

[1] C. T. Yang, T. F. Han and H. C. Kan, "G-BLAST: a Grid-based solution for mpiBLAST on computational Grids", Concurrency and Computation: Practice and Experience, vol. 21, no. 2, pp. 225-255, 2009.

[2] I. Foster, "What is the Grid? A Three Point Checklist", July 20, 2002. Available: http://www.mcs.anl.gov/~itf/Articles/ WhatIsTheGrid.pdf.

[3] M. Rabb and C. Doninger, "Grid computing and SAS", SAS Institute Inc.US, 2004, Avaliable: http://support.sas.com/rnd/ scalability/papers/101948_1204.pdf.

[4] P. Asadzadeh, R. Buyya1, C. L. Kei, D. Nayar, and S. Venugopal, "Global Grids and Software Toolkits: A Study of Four Grid Middleware Technologies", Available: http://www.buyya.com/ papers/gmchapter.pdf.

[5] B. Fran, F. Geoffrey and H. Anthony, "Grid Computing : Making the Global Infrastructure a Reality", Chichester : Wiley, 2003.

[6] J. H. Kaufman, T. J. Lehman, and J. Thomas, "Grid computing made simple", The Industrial Physicist, pp 32-33, Aug- Sept 2003.

[7] R. Gupta, "Bio-Informatics,Bonding genes with IT", in INDIACOM 2010, Jaipur, Feb 2010.

[8] BLAST, Available: http://blast.ncbi.nlm.nih.gov.

[9] J. Fox, "What is bioinformatics?", The Science Creative Quarterly, Issue4, 2009.

[10] C.Blanchet, R. Mollon, D. Thain and G. Deleage, "Grid Deployment of Legacy Bioinformatics Applications with Transparent Data Access", in 7th IEEE/ACM International Conference, Barcelona, pp 120-127, 28-29 Sept. 2006.

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

[12] X. Gong, K. Nakamura, K. Yura and N. Go, "Toward Building Grid Applications in Bioinformatics", in Fourth Australasian Symposium on Grid Computing and e-Research (AusGrid 2006), Hobart, Australia, 2006.

[13] J. Andrade, M. Andersen, A. Sillén, C.Graff and J.Odeberg, "The use of grid computing to drive data-intensive genetic research", European Journal of Human Genetics, 15, 694–702, March 2007.

[14] Eurogrid, www.eurogrid.org

[15] APBiogrid, http:// www.apbionet.org/apbiogrid

[16] Mygrid, www.MyGrid.org.uk

[17] Ncbiogrid, www.ncbiogrid.org

[18] Singapore BioGrid, www.bic.nus.edu.sg/biogrid