Life Science Applications in Grid Environment

C. Murugananthi
Research Scholar
Department of Computer Science
Bharathiar University, Coimbatore

D. Ramyachitra
Assistant Professor
Department of Computer Science
Bharathiar University, Coimbatore

ABSTRACT

Grid Computing is a high performance computing infrastructure with large-scale pooling of resources that may be processing cycles, data or storage that allows sharing of various distributed resources across many administrative domains and used to solve large scale computational problems. In the grid environment, users can access the resources without knowing where they are physically located. Grid Computing has its applications in various areas such as science, finance, business, health care, government etc. This study provides the various disciplines within the life science area that uses grid computing to solve complex problems.

Keywords

Grid Computing, Life Science, Bioinformatics, HealthCare, Proteomics, Genomics.

1. INTRODUCTION

Grid provides the capability to combine a vast amount of computing resources which are geographically scattered to undertake large problems and workloads. Grid environment solve the large scale technical or scientific problem that needs a great number of computer processing power cycles or access to mass of data. Life science is one of the fastest growing application areas in grid computing. There are an increasing number of life science applications that utilize grid technologies.

Grid and distributed computing have been a core component of the IT infrastructure of many of the large- and medium-sized life sciences companies. Grid computing is emerging as a promising approach to speed up analyses for database searches, data mining and sequencing. It divides tasks into parallel programs that execute independently, using unused computing cycles to minimize cost. In high-performance computing, development of grid technologies and improvements in parallel computing make the power needed for these complex tasks available to every scientist and research engineer [1].

Scientists and researchers depend on databases to access the mass of data that they produce. Researcher needs to search lot of different databases to find all the information related to their research. Maintaining up-to-date and fully functioning of all these databases, tools and techniques to search them are a huge and complex task. Hence there is a need of access to up-to-date information related to their research. The major challenge for data analysis in life sciences is to offer an integrated and up-to-date view of rapidly growing volume of data in a number of formats. The aim is to provide the most up-to-date version of all databases for a task in a grid environment. So there is a necessity for a service that will bring up-to-date each site storing the databases through the grid with their last modifications [2].

Computational biology, genomics, bioinformatics, computational neuroscience areas make use of grid technology as way to access, gather and mine data, to bring out large-scale simulation and analysis. Grids are used in various areas of life sciences such as bioinformatics, computational chemistry, molecular modeling, genomics, proteomics and health care. The Grid environment provides a common infrastructure for data access and simultaneously offers safe and reliable data access methods while processing the data [3].

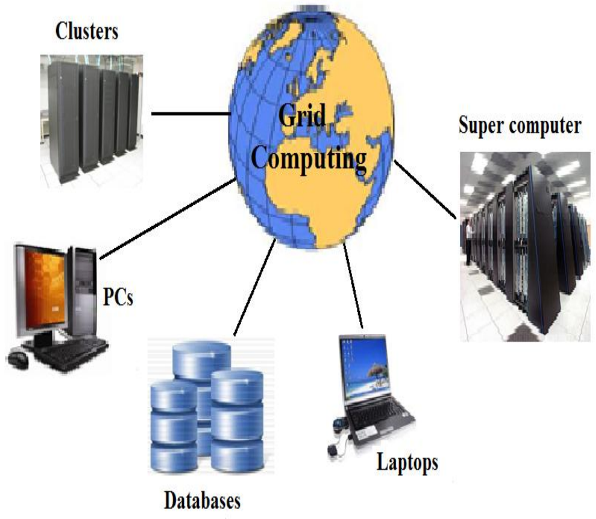

Fig. 1: Grid Computing Structure [29]

Life sciences make use of the Grid Computing to implement sequence comparison algorithms and facilitate molecular modeling using the collected secured data. A vast amount of data is being created from biochip, protein structure analysis, DNA sequencing and molecular modeling. The data obtained with these technologies might be useful when they are processed and interpreted. The need for processing this mass and rapidly increasing data in a short period of time has resulted in the development of high throughput techniques. 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 [4].

The rest of the paper is organized as follows. Section 2 discusses the grid applications in bioinformatics. Section 3 presents grid enabled applications in health care. Section 4 discusses the ongoing life science projects that use grid technology. Finally, section 5 gives the conclusion.

2. BIOINFORMATICS

Bioinformatics is a discipline of biological research involving the combination of computers, databases and software tools. Bioinformatics is necessary to accomplish lots of complex tasks such as the use of genomic information in understanding human diseases, detection of new molecular targets for drug discovery and in resolving human evolution mysteries [5].

Grid computing plays a vital role in bioinformatics, which focuses on the search for information and comparison of new information to that which is already established. It involves the comparison of an undiscovered sequence to the sequences in a database to find out the similarities among the sequences or making predictions about the sequence based on current knowledge of similar sequences. Protein sequence databases, gene expression databases, gene sequence databases and related analysis tools help researchers to find out whether and how a particular molecule is involved in a disease process [6].

Drug discovery and biotech research companies need powerful, high-performance solutions to search through and analyze huge amounts of data. High-speed, high-performance computing power and industrial-strength databases perform a wide range of data-intensive computing functions: mapping proteomic and genetic information, data mining to identify patterns and similarities and text mining using huge libraries of information [7]. These activities require high-performance computer infrastructures with access to huge databases of information.

Bioinformatics requires research infrastructures that are capable to store very large, complex and heterogeneous biological data sets and to make these data available for intensive scientific computing. Mass of computing resources is often required for large-scale bioinformatics analysis. Most of the bioinformatics tools such as HMMER and BLAST needs powerful computer resources for better performance [5]. Location transparency is an essential feature of grid computing [8]. Users gain access to applications without being aware of where these packages are installed. Scientists do their researches in a virtual laboratory, in which they share computational tools, common databases and their analysis workflows using powerful, high performance grid computing technologies.

2.1 Genomics and Proteomics

Modern approaches to biology such as genomics which looks for individual genes in the variety of human DNA or proteomics which tries to describe the most complicated molecules in the body that require sifting through huge amounts of data. Scientists deal with basic elements, genes and proteins and work to understand their structures [9]. Using grid computing, scientists can compare structures to those already defined such as in the human genome project and simulate new ones with computational biology and chemistry.

Grid computing is emerging as one way to make database search happen fast, using idle computing resources in the company or around the world via the Internet to run these sequences. Each computer in the grid is given a small subset of the main task, such as comparing a particular protein structure to part of a large database and carries it out independent of any other system in the grid [3].

The nature of genomic and proteomic data analysis is that it can be easily adapted to the grid computing environment. Genomics applications are data driven and have long execution times because it requires the integration of different biological tools and databases [6]. Because of this reason the development of a grid workflow to distribute the computational job on a remote computer via the grid platform will enable the possibility to achieve complex workflow of analysis on vast amount of data from a variety of organisms' genome. Complex and huge amount of data are produced in this field. So high performance computing is needed to achieve the intensive data processing and analysis.

The huge amount of data produced through sequencing of the human genome has increased the demands for cost-effective and flexible alternatives in proteomics, genomics and drug discovery. Grid environment is useful for solving high performance computing tasks in genetics and proteomics using algorithms [9]. The data in genomic sequence analysis involves sequences of either nucleotides (DNA or RNA) or amino acids. Genomic analysis basically involves either comparing pairs of sequences, looking for patterns in individual sequences or working with small families of sequences to expose patterns that can be compared to other sequence [21].

The sequence analysis in biology involves sequence alignment, repeated sequence searches, comparing a peptide or DNA sequence with the database of sequences or other bioinformatics mechanisms on a computer. A sequence alignment is a method of arranging the sequences of RNA, DNA or protein for detecting structural, functional and evolutionary relationship in biological sequences [10]. It also facilitates the annotation of new sequences, modeling of protein structures, design and analysis of gene sequence experiments [6].

One of the most commonly used bioinformatics tool is BLAST (Basic Local Alignment Search Tool). It is used to find the parts of local similarity between the biological sequences. A BLAST gets a query sequence and searches it against the database chosen by the user and arranges the query sequence against every sequence in the database and calculates the statistical importance of matches [22]. There are different types of BLAST tools that are available according to the type of query sequence and database. BLAST can also be used to identify the functional and evolutionary relationships between the sequences and finds the members of gene families.

Users of the genomics grid wants to find, differentiate among and select the most suitable services and may need to be reported while resources are changed or updated. When a great number of bioinformatics resources are needed, it is essential to coordinate them in a workflow on the grid platform. An analysis of genome is the development of data processing scripts to divide and assemble genome data and results [11]. Based on the analysis, dividing and assembling operations can be managed in a generic manner. The access to and analysis of many large databases in proteomics and genomics forms the base of research in computational biology. Currently, several gene and protein sequence databases have been categorized and distributed on the grid environment [12].

The formalized knowledge representations (ontology) will play a vital role in any genomics and proteomics application. The grids have to establish a shared ontology for life science applications. These developments will be most important to build an open, interoperable software tools and services for supporting discovery operations that are able to combine clinical and genomic data in grid. The grid computing proved to be an up-to-date tool essential in genomics to support and enable the collaboration of people using resources through highly capable computation and data management systems [8]. It also enables the projects in genomics to share computational resources, to perform computer simulations and large-scale data analysis and high-throughput sequence analysis.

Protein structure databases employ several methods and varying levels of biological data on well described proteins to obtain protein signatures. An analysis on protein field can be performed with the help of searching tools and a set of databases [12]. Based on the methods used, this software can take long time to finish since they have to match the domain models against the query sequences producing a large number of combinations.

There are two types of structural methods. First method includes library of complex structural domains and second involves local characteristic of proteins. Working on local characteristics of the molecule is a difficult task, but constructing patterns of the protein local topology, the analysis can be very sensible in identifying the geometry of a functional site [13]. These methods can be time consuming due to the vast number of configuration to explore while verifying similarities. Because of this reason, a grid based approach can be very useful to tackle these new challenges, offering a scalable solution for problems that involve a high computational load.

2.2 Molecular Modeling

Molecular modeling is a methodology for modeling and simulation of molecules behavior. It combines the computer graphics and computational techniques and used in drug design, computational chemistry, computational biology. Drug design using molecular modeling techniques includes screening a high number of molecules of compounds or ligand records in a chemical database to recognize those that are potential drugs. This method is called as molecular docking. It aids scientists to forecast how small molecules bind to biocatalyst or receptor protein. Docking each molecule in the target chemical database is both a large computational and data intensive task. Grid technologies offer inexpensive and effective solutions for running molecular docking tasks on large-scale, wide area distributed and parallel systems [14].

Drug design is a computational and data challenge problem because it includes screening billions of compounds in chemical databases. Screening each and every compound based on structural complex needs a few minutes to hours on a normal PC, that is screening all compounds in a single database may require years. The same problem can be solved with a large scale grid of thousands of supercomputers within a day. If using a massive network of peer-to peer mode grid computing infrastructure such as SETI@Home means, then the drug discovery problem could be solved within a few hours [15].

Grid technologies reduce drug compounds from billions to thousands or even hundreds, detecting the most promising candidates and hurry up the discovery process. Grid environment is more suitable to drug discovery process since computing the probability for one ligand to dock or fit, to one protein can be possible on each computer node in the grid, providing enormous parallelism. Using the high-speed powerful computation and massive data managing abilities of the grid, possible drug components can be studied and screened very rapidly [7].

The variety and range of computing uses in biotechnology and drug development provides a number of openings for both grid and high-performance computing. The huge amounts of data that need to be analyzed and the complexities of molecular modeling both points to parallel computing as a means to shorten the time to result.

3. HEALTH CARE

Several medical applications can use various functionalities offered by the grid computing. Grid provides an access to metadata databases and large medical images located in different sites. This is helpful for lots of data intensive medical image processing applications for which large datasets are needed. Doctor had to access a grid for medical image files, administrative databases and specialized instruments such as cardio angiography devices, MRI machines and CAT scanners. This could improve the analysis of complex medical images, diagnosis methods and enable life-critical applications such as remote cardiac monitoring and telerobotic surgery [16]. Grid computing is also useful in several situations like creating interactive medical simulations such as analyzing and managing medical images and in heart simulation. Grids are used in supporting virtual collaboration in e- Hospitals (a virtual network of hospitals that serves medical analysis services, medical training, e-surgery) [17].

3.1 Simulation-based Applications

The simulation of a treatment planning is mostly computing intensive. Treatment planning is the typical simulation based applications. They are based on a Monte Carlo simulation engine, with the ability to define the body geometry at the required resolution [18]. Grid is used in the simulation application in such a way that the simulation of several particles can be treated as a parallel problem and divided into many processes. These processes will be executed in parallel on CPUs offered by one or more computer centers that are typically remote with respect to the user that issues the request.

3.2 Medical Imaging Analysis

Medical Imaging Analysis is a class of application that can make use of the grid services. Medical imaging analysis needs both high level data management services and massive computing resources [17]. If many hospitals belong to the same virtual organization, they are able to share a common analysis algorithm, which is made available by a grid service provider acting as a server. Computer Assisted Detection algorithms can be triggered on demand from a remote repository. Using a local Graphic User Interface, an authorized user is able to upload a new image to the local storage, register it to the data catalogue and then, depending on the properties of the image, request for the execution of the analysis algorithm.

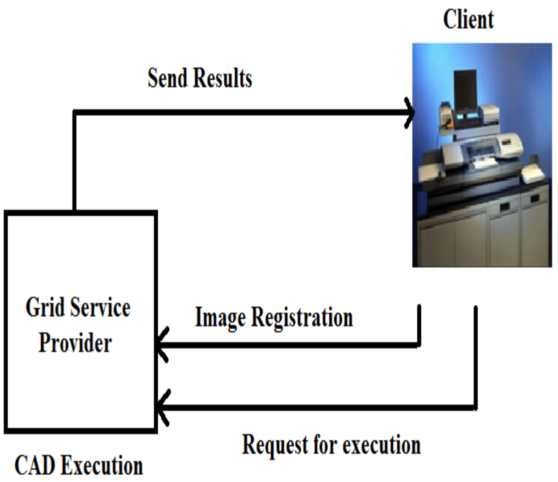

Fig. 2: Image Analysis using CAD algorithm in Grid Service Provider [19]

Based on license issues, the analysis can take place locally or on the grid provider site. The result of the analysis is exposed on the client GUI. The analysis algorithm is executed either as an interactive process or as a batch job, based on the implementation and the system configuration. The image analysis requires computing resources, storage and high-level GUI on the client side. The provider side requires computing resources and metadata management services. File transfer functionality between the server and the client is also required. A prototype implementation is based on Computer Assisted Detection algorithms. The advantage of using grid technology in image analysis is the possibility for sites to utilize proprietary algorithms that are too expensive for a small hospital on a pay per image basis: the license-related costs are sustained by the grid service provider [19].

3.3 Tele-radiology and Epidemiology

Tele-radiology is useful when radiologists want to share with expert colleague about the diagnosis or to show something about diagnosis to young radiologists during a training session. Assuming a grid virtual organization is deployed in N sites, each of them owns a distributed database of images, which are all registered with their metadata in the VO data catalogue, hosted by a GRID Service Provider. Any authorized user from any site in the system is able to query the data and metadata catalogue and retrieve a list of images that meet a selected set of requirements [16].

The epidemiology is a field, that collects statistical distributions of metadata over huge population of patients and the field of pathology needs to collect and assemble the data from a large number of sites. Hence grid environment is suitable to handle computational result from examining all databases [20].

4. LIFE SCIENCE PROJECTS

World Community Grid (WCG) creates the largest public computing grid to undertake research projects that gains the advantage for humanity. World Community Grid runs on BOINC (Berkeley Open Infrastructure for Network Computing) software. Using the unused processing power of computers around the globe, World Community Grid's research projects have studied the features of cancer, HIV, human genome, dengue, influenza, muscular dystrophy, clean energy, and rice crop yields[24].

Folding@home is a distributed computing project undergone with research about protein folding, aggregation, misfolding and related diseases. Volunteers those who are interested to help this project have to install the Folding@home software on their computers. It uses the idle processing resources of the volunteer's computer. It automatically uploads the results to server each time it finishes a job unit and downloads a new job at that time. The purpose of this project is to understand the mechanisms of protein folding and examines the reasons for protein misfolding and determines the diseases (alzheimer, Huntington and various cancers) caused by the protein misfolding [24].

BioinfoGRID combines the bioinformatics and its applications for molecular biology users with the grid infrastructure created by the EGEE Project. The BioinfoGRID project analyzes the applications in the fields of proteomics, genomics, drug discovery and transcriptomics. It minimizes the data computation times by allocating the computation on thousands of computers with the grid infrastructure created by the EGEE Project. The project supports the research on applications for distributed database access, microarray technology and distributed laboratory management systems, analysis of DNA data, gene expression studies, protein functional analysis, gene data mining, phylogenetics analysis and molecular dynamics simulations in GRID [21].

FightAIDS@Home project aim a single disease. Each computer processes one drug molecule and tests how well it docks with HIV protease, acting as a protease inhibitor. The software uses idle computer's processing cycles to support fundamental research in determining new drugs, building on growing knowledge of the structural biology of AIDS. In addition, this research assists in analyzing the mechanisms of multi-drug-resistance that the super bugs of HIV use to escape from the current anti-AIDS drugs. This research also helps to build, asses, refine, and share the protocols and tools that other labs use in their research against various diseases [26].

Rosetta@home is a distributed computing project for predicting protein structure. This project runs on BOINC (Berkeley Open Infrastructure for Network Computing) platform. The software of this project uses the idle processing resources of the volunteer's computer. Rosetta@home aims to predict and design the protein structure and protein-protein interactions with the use of several active volunteered computers. This project determines and designs the 3-dimensional structure of proteins in research that may lead to finding cures for various major human diseases such as Malaria, Alzheimer's, HIV, and Cancer [27].

Drug Discovery Grid project intends to design a platform for drug discovery using the peer-to-peer and grid computing technology. This project aims to solve data intensive and large-scale scientific computation applications in the fields of molecular biology and medicine chemistry [23].

WISDOM (World-wide In Silica Docking on Malaria) aims at developing new drugs for emerging and neglected diseases with a particular focus on malaria. Its main goal is to enhance the research and development on neglected diseases by using the open source grid technology for drug discovery [28].

5. CONCLUSION

Grid computing combines the massive computing resources together, producing a large system with enormous computational power that far exceeds the power of supercomputers. In grid environment the work is divided into small units that can be processed concurrently, so research time is minimized from years to months. Grid technology is also more cost-efficient, facilitating better use of critical funds. This paper has given a survey on the applications and projects in life sciences that uses grid technology. Life science applications demand more integration of distributed, huge and complex data, high computing power, as well as applications of heterogeneous networks. Thus grid technologies are sufficient to solve high-throughput life science applications.

6. REFERENCES

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

[2] EL-Ghazali Jalbi, Albert Y. Zomaya, "Grid Computing for Bioinformatics and Computational Biology", 2008, Wiley publications.

[3] Information Resources Management Association, USA, "Grid and Cloud computing", April 30, 2012, IGI Global publication.

[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] 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.

[6] Ahmar Abbas, "Grid Computing: A Practical Guide to Technology and Applications", 2009, FireWall Media, New Delhi.

[7] Robert Stevens, Robin McEntire, Carole Goble, Mark Greenwood, Jun Zhao, Anil Wipat and Peter Li, "myGrid and the drug discovery process", DDT: BIOSILICO Vol. 2, No. 4, July 2004.

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

[9] "Distributed, High-Performance and Grid Computing in Computational Biology", International Workshop, GCCB 2006, Eilat, Israel, January 21, 2007. Available: http://www.springer.com/computer/bioinformatics/book/978-3-540-69841-8.

[10] Jorge Andrade, "Grid and High-performance computing for Applied Bioinformatics", 2007, Available: kth.diva-portal.org/smash/get/diva2:12929/FULLTEXT01.

[11] Jorge Andrade, Malin Andersen, Lisa Berglund, Jacob Odeberg, "Applications of grid computing in genetics and proteomics", 2006, PARA'06 Proceedings of the 8th international on Applied parallel computing: State of the art in scientific computing, Pages 791-798.

[12] Wolfgang Eppler, Volker Hartmann, Wolfgang Wenzel. Forschungszentrum Karlsruhe, Karlsruhe, Germany, "Grid Computing for Proteomics". Available: fuzzy.fzk.de/eppler/pdf/proteinGridWenzel.pdf

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

[14] Mark L.Green, Russ Miller, "Molecular Structure Determination on a computational and data grid", 2004, pages 1011-1017, Parallel Computing. Available: www.elsevier.com/locate/parco.

[15] Rajkumar Buyya, Kim Branson, Jon Giddy and David Abramson, "The Virtual Laboratory: a toolset to enable distributed molecular modelling for drug design on the World-Wide Grid", Concurrency and Computation: Practice and Experience: Exper.2003; 15:1-25 (DOI: 10.1002/cpe.704).

[16] Jason C. Care, Forrest W. Crawford, Sarah J. Nelson, "Grid enabled magnetic resonance scanners for near real-time medical image processing", J.C. Crane et al. / Journal of Parallel and Distributed Computing 66 (2006) pages 1524 – 1533.

[17] Vincent Breton, Christophe Blanchet, Yannick Legre, Lydia Maigne, and Johan Montagnat, "Grid Technology for Biomedical Applications", 2005, pp. 204-218 Available: http://link.springer.com/ chapter/ 10.1007/11403937_17.

[18] A. Ferrari et al., "Update on the status of the FLUKA Moute Carlo Transport Code", 2006, CHEP'06 proceedings.

[19] P. Cerello, INFN, Sezione di Torino, Italy, "Grid Computing in Medical Applications", Available: http://indico.cern.ch/getFile.py/access?contribId=451.

[20] Mario Cannataro, Rodrigo Weber dos Santos, Joakim Sundnes, Pierangelo Veltri, "Advanced computing solutions for health care and medicine", Journal of Computational Science, Volume 3, Issue 5, September 2012, Pages 250-253.