Java-Grid Environment for Bioinformatics Applied Tools (JEBAT)

Noor Maizura Mohamad Noor¹, Ahmad Faiz Ghazali¹, Md Yazid Mohamad Saman¹, Zafarina Zainuddin²,

¹ Computer Science Department, Faculty Science and Technology, University Malaysia Terengganu (UMT), Malaysia
² School of Health Science, Health Campus, University Science Malaysia (USM), Malaysia

Abstract. The advancement in bioinformatics and medical fields continue to enhance together with the technology of various computing. Bioinformatics is the combination of life science and computer science or biology and information technology. Bioinformatics applications and expert systems sometimes requires high performance machine in order to process multiple data simultaneously. Grid computing can be applied where the needs is crucial, especially in critical fields such as bioinformatics and medical. Middleware, a form of management system that access and makes available network resources, control and schedule computations, and optimize the flow of data in a network is required. Java-Grid Environment for Bioinformatics Applied Tools (JEBAT) is proposed in order to design a grid platform which utilize middleware tool that can enables bioinformatics researchers to create software with Java-based technology and the idea of grid-based sharing resources.

Keywords: Bioinformatics, Grid computing, Java, node.

1 Introduction

Bioinformatics is a great field to be explored. Everything that happens around us can possibly be computed therefore that we can measure and employ it for human exploration in future. The increase of data size and algorithms to analyze complex data in bioinformatics shows the needs for high performance solution including grid. There are various ways to exploit the advantages of grid computing. For applications with large databases such as bioinformatics tool, preinstalled on particular nodes can significantly reduce processing time. Common data grid activities include analysis of massive amount of data in remote databases that is required in bioinformatics field, as an example is in genomic research.

Computational time is very crucial for certain advanced analysis. System development and re-codification to change and fix this might require months of work; therefore the results can be compared at the end. An example of another way to do it is to divide time-consuming tasks into small jobs that are to be executed in a form of parallel processing. Existing algorithm and strategies need to be modified so that it enables parallel execution of jobs. This will ease computational calculation processes.
in bioinformatics tasks and analysis of smaller separate jobs. Sharing in wide resources is available in the concept of grid computing environment.

In another words, grid computing can also be defined as a service whereby sharing of computing power is enabled via segregation of jobs in parallel from data storage capacity. Application can be executed on the grid computing platform where jobs are scheduled by modular engines designed. Grid architectures can also contains several nodes including local machine or desktop node. The common concept of execution is that each engines have own id and know its tasks including what kind of output is expected. Scheduling rules need to be defined for the application software in order to deploy it using grid computing. A sequence analysis system can then be designed, run, proposed and evaluate its performance.

Job from client-side explicitly passed the data via parameters to an available grid node, and they will receive back the result of completed processes. This jobs flow and processes are surely hidden from user (scientist and expertise) where only visual and end result being displayed in the client-side tool. The management and transfer between local and grid resources are enabled. Integration of these mechanisms in reference engine based on comparison of hypothesis through Bayesian network concept inside this expert system is another special feature here.

Definitions and introduction of grid computing in bioinformatics has been discussed in section 1. Section 2 briefly explains on the history of grid computing including its current advancement in Malaysia. Subsequently on the next section, several existing bioinformatics applied tools using grid computing is discussed. The proposed system framework using java-grid environment is stated in Section 4, and the whole review is then concluded in Section 5.

2 Short Overview of Grid Computing in Malaysia

Malaysian Institute of Microelectronics Systems (MIMOS) through its KnowledgeGrid Malaysia gained an increase of computing power from this connection to Enabling Grids for E-SciencE (EGEE) that consists of clusters in Germany, Britain, Austria, Turkey, the Netherlands, Italy and other nations. Grid computing in Malaysia had being started to evolve. A few thousands CPUs at MIMOS Berhad’s Technology Park Malaysia headquarters, which are Malaysia’s grid computing system has been successfully connected to a European Union-supported grid, EGEE. This initiative will ease Malaysian researchers to make a better collaboration outside the country [1]. Besides in Malaysia, Multimedia Super Corridor (MSC) just launched this program a few months back.

In Malaysia there are only very a few users involved in the grid computing because many practitioner and experts including businesses and industrial people still do not fully understand how the KnowledgeGrid can assist their daily operations. “Industries would be able to keep the bulk of their IT budgets for other aspects of product development because KnowledgeGrid would meet most of their computing needs. MIMOS Grid Computing targeted to change IT capabilities for the country to the next level. MIMOS spearheads the KnowledgeGRID Malaysia, a National Grid Computing Initiative under the Ninth Malaysia Plan. [2]. We can see the increasing
needs for our own Grid sharing system because the increase of its future potential over years.

3 Several Bioinformatics Applied Tools using Grid Computing

Grid computing is being applied because it can provide an environment which enable easy and fast access to distributed data resources. One of it is BioGrid Application Toolkit which was created for accessing a grid network that can solve a computational problem for algorithms implemented in parallel and distributed computing [6]. Java is a good choice for grid computing environments because java is a modern, object-oriented programming language that makes software engineering of large scale distributed systems much easier. Furthermore, there are also other suitable programming languages for grid computing implementation, such as JSF (Java Server Faces). This kind of technology can be used to upgrade a group of single user computers like a supercomputer with a high computational power and storage capacity. It can solve complex calculations efficiently. Execution time will be shorter when it is done in parallel processing instead of conventional linear processing.

There is a large amount of genetic sequence information in DNA databases nowadays. Biologists and computer scientist can perform initial analysis, but since the execution time for many bioinformatics applications is currently not practical and size of the database began to increase, algorithms for sequence comparisons can be coded so that parallel processing execution can be done which will improve execution speed. Sequence matching is one of complicated and sophisticated but most important in computational biology.

Sequence databases from DNA can also be implemented on grid computing, where one of the most widely used bioinformatics tools is method from the BLAST program [7]. GridSphere [8] that is also one type of middleware to implement grid computing allows researchers and developers to adjust the requirements via XML based file which automatically generates a JSP file for web page front-end. Grid software packages and current development of grid technologies have various principal. The most popular and common used middleware software technologies for implementing a grid are Globus, Legion and UNICORE.

Another way to handle database is by using replication method if the nodes that working on these processes has many CPUs sharing the same storage and at the same time accessing a different set of databases. This is the reason that remote input or output overweighs the replication, but anyway it depends on the particular system algorithms itself. Quality of results can improve tremendously by scheduling a dynamic architecture that is most suitable with operations to be executed. Many research centers share their computing assets in grids that will automatically increase the storage for application resources and then number of processing. This is because grids allow secure sharing from various computational resources efficiently in administrative domains.

Scientists has always employed and created many approaches to analyze bioinformatics data. For each technique, there are might already be available efficient product or program to handle data in variety ways. The software tools made may have
certain advantages and disadvantages whether in terms of processing power or even its main functionalities. The concept of grid computing can also be explained as an integration of machines, databases and networks that aims to provide enhancement of tools and application within its scope of resources.

Today, there are already a big number of projects working on grid computing applied to bioinformatics. Biological databases and bioinformatics software tools such as for DNA sequence analysis need to be studied more from now on due to its demand. Another capability that usually exhibited from similar developed projects to run bioinformatics tool is utilization of multiple grid resources simultaneously.

The fast pace development in biology and medical fields requires software tools that able to compute operations with high performance even with massive amount of data to be managed. The software tools environment also needs to be able to be used easier and support requirements and needs of expertise. The user interface for the software tools should hide the structure of grid computing technologies whereby difficult to be understood by biologist. The design should concentrate more on the details for jobs to be executed and its analysis.

As internet bandwidth and communication technology are already increasing so much in this last few years, it is also expected that grid resources does not need to be in only one place in order compute scientific operations. Computational resource through the power of grid is also expected to be capable of these knowledge-sharing or even urgent collaborations where important decision and scientific operations need to be computed during emergency; such as handling DNA data of victims from natural disasters. Grid computing can be setup by connecting a number of computers and high performance networks.

Efficient and scalable algorithms need to be determined in order to solve large scientific and medical problems in many alternatives methods and views. Our approach is based on implementation using Java, which is already well-known as the best software solution nowadays includes for parallel processing and even grid computing has been applied in bioinformatics. Ideally, it takes 5 to 10 years for a new revolution to be accepted and recognized by specific community, including scientists. Hence, the planning and implementation should be started now. Grid computing has a close relation with peer-to-peer computing as well. Computational grid environment is also like a virtual organization [14].

Table 1. Several related researches on field bioinformatics using grid computing

<table>
<thead>
<tr>
<th>Title (Author, Year)</th>
<th>Description</th>
<th>Summary of contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grids in bioinformatics and computational biology (El-Ghazali Talbi, Albert Zomaya, 2006) [9]</td>
<td>Discussed a few papers related to Grid technology, parallel evolution strategy, and approach on computational Grid and large-scale multiple assignment.</td>
<td>Grids enable sharing of geographically dispersed resources either among machines, networks, data storage, visualization devices, or scientific instruments for large-scale distributed applications.</td>
</tr>
<tr>
<td>Title (Author, Year)</td>
<td>Description</td>
<td>Summary of contents</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>A Bioinformatics Grid Alignment Toolkit (Maria Mirto, Sandro Fiore, et. al, 2008) [6]Title (centered)</td>
<td>Each node of the Grid stored a part of databases. Different segment of databases are searched on each nodes, but results still collated in a single output file. They also proposed <strong>BioGAT – Bioinformatics Grid Alignment Toolkit</strong>.</td>
<td>For the alignment applied to a large set of data, load of work partitioned into different jobs. This is a good choice for each simulation.</td>
</tr>
<tr>
<td>Analysis of DNA sequence transformations on grids (Yadnyesh Joshi, Sathish Vadhiyar, 2009) [11]</td>
<td>They developed a grid system that performs parallel guided explorations of the rules on grid resources. They conducted experiments on the grid comprising machines in three countries.</td>
<td>The primary contributions of work are consists of: Modeling DNA sequence mutations using cellular automata, and resolved uncertainties related to phylogenetic tree.</td>
</tr>
<tr>
<td>Bioinformatics algorithm development for Grid environments (Fotis A. Psomopoulos, Pericles A. Mitkas, 2010) [12]</td>
<td>They present a novel theoretical workflow framework for data analysis in bioinformatics, which can operate entirely within the Grid.</td>
<td>A Grid environment works by taking advantage of many computers connected by network (even though geographically dispersed) and then provides the ability to get high throughput computation.</td>
</tr>
</tbody>
</table>

According to [4, 5] both of this system is expected to run on sharing condition. Besides, grid computing resource facilities can also be utilized. Forensic DNA Analysis System [4] will be the main reference engine since DNA sequences (in many forms) will need to be main database reference in bioinformatics and medical fields. Researchers [5] currently familiarized with Java Programming languages as well in order to implement this project. Researchers [5] mentioned “The expected outcome from this study is a new framework for interpreting and analyzing child sexual abuse crime cases that can be customized for other crime type analysis and application.”
Each node on the grid does not necessarily must be an application, it can also dynamically be presented either by a database or even a running task or any parallel processing computation to be executed. The applications or tools might be built based on concept of Decision Support System (DSS) or Artificial Intelligence (AI) if necessary. This will add the richness and variety of dynamic integration between components on JEBAT. Usage of XML embedded coding and files are expected to enable a lot of obstacles. This will act as a grid platform for scientific application consists of researchers projects on bioinformatics.

Bioinformatics computations always require more processing power than existing available power. These needs are being fulfilled by parallel processing such as in grid computing environments. The integration of computing resources will enable the sharing of data and processing power. Researchers [13] listed out the bioinformatics applications that are to be implemented on grid. Researchers [13] defined four layers architecture. They also showed applications user interfaces. Bio-STEER [13] is a very useful tool but they continue to upgrade it. They also list current and development planned objectives systematically, compared to other researchers.

4 Proposed System Framework using Java-Grid Environment

The aim of the JEBAT project is to design a grid platform which utilize middleware tool that can enables bioinformatics researchers to create software with Java-based technology with the idea of grid-based sharing resources. The research aims to close the gap between scientists and computer programmers in Universiti Malaysia Terengganu (UMT). A centralized network for bioinformatics researchers consists of scientists and IT experts through laboratories can be created. The scientific methodology and the development of software tools to handle massive amount of biological data is required to manage the project effectively where in the end a variety of different services can be provided to community.

The software tools are expected to have many special functions, including dynamic data retrieval and data analysis on grid computing environment. The functionality for these kinds of tools is very crucial. Other significant benefit from grid besides job distribution and computational power is security authentication. This project’s focus is to develop a framework to deploy a number of bioinformatics applications and databases on grid computing, starting with forensic DNA analysis tool. Nowadays where bandwidth increased while cost decreased, the needs for high performance computing keeps arise as well.

Massive data operation problems in bioinformatics can be solved using distributed computing platform concept. Since bioinformatics computation problems involve massive data operations, a good algorithm need to be designed and proposed in order to enhance its productivity in handling jobs. In demonstrating high performance computer starting from initial parallel and distributed computing until grid computing in future, development for bioinformatics applications nowadays focused more on software architecture or system framework rather than only hardware.

New bioinformatics applications requires large computational power especially grid which imitate supercomputing power. We need to concentrate more on job
services and significant functions from the system. The system framework of this grid platform are still requires to demonstrate data integrity and security authentication on sensitive data. Balance of network traffics is also very important in order to increase system performance. It is hoped that the algorithm and grid computing platform will be able to choose computing nodes automatically in order to be executed in parallel processing nodes.

Paper by [3] mentioned a lot regards grid computing security. They provide an extensive survey of the different methods to address security issues in grid computing environment. Classification is presented by [3] into System Solutions, Behavioral Solutions, Hybrid Solutions and Related Technologies. Each category presented in details. Intrusion detection system (IDS) is mentioned too because it can also be used in order to monitor networks from any intruders.

In JEBAT, visualization tool that will monitor the network flow is needed as well. This framework is also proposed in order to manage distributed databases systematically and dividing jobs into smaller portions so that any operation or calculation can be executed faster. The model consists of three primary layers: Web Layer, Service Layer and Resource Layer (Figure 1). The elements inside dotted box are referring to Web-based bioinformatics applied tools that can be included inside this grid platform; such as Forensic DNA Analysis System [4] and Forensic Child Sexual Abuse (CSA) System [5]. Physical network topology of JEBAT can be seen in Figure 2.

![Grid framework platform for postgraduates’ bioinformatics’ projects – JEBAT](image)

**Fig. 1.** Grid framework platform for postgraduates’ bioinformatics’ projects – JEBAT
5 Conclusion

Algorithms for high-throughput bioinformatics data analysis such as DNA sequence search and protein structure analysis requires CPU-intensive utilization, and great computing power that is more than a single computer or cluster can handle, besides long turnaround time. Researchers and biomedical institutes are beginning to develop software tools that can take advantage of available high-end computing resources in order to run computing tasks on a larger scale than usual. Bioinformatics problems can potentially being handled by grid computing that have implied power like a supercomputer.

The contributions of this research is the proposed framework as a grid platform that consists of bioinformatics applications with distributed databases and power of parallel processing on grid computing environment. This framework is expected to be able to enhance data integration from a number of bioinformatics applied tools too. This paper is also expected to survey current advancement for the implementation bioinformatics on grid computing environment. The features for system security should be investigated further in future work so that it does not consume too much power that can influence the stability of parallel processing tasks executed on Grid computing environment.
Grid technology has revolutionized bioinformatics by providing resources from different places and able to share computational power, database, storage, software applications as well as other related hardware requirements. There will be many benefits of grid computing environment in JEBAT, including: i) Resource sharing of DNA Database, ii) Easy to learn and use, iii) Friendly user interfaces and iv) Reduce time taken to operate bioinformatics computations and waiting for results.

In conclusion, grid computing is an emerged technology that must be implemented where it is necessary in order to enhance collaborations of work and speed up operations. Complex and time-consuming computational problems can be solved easier and faster than before. The potential of grid computing environment need to be explored furthermore due to its high demand in variety of fields, not only bioinformatics, but in business and financial transactions as well.

References