Pure Appl. Chem., Vol. 74, No. 6, pp. 891–897, 2002.

Grid computing and bioinformatics development. A case study on the *Oryza sativa* (rice) genome*

Wasinee Rungsarityotin, Noppadon Khiripet, Chularat Tanpraset, and Royol Chitradon‡

National Electronics and Computer Technology Center, National Science and Technology Development Agency, 112 Paholyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand

Abstract: 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. This research suggests a feasible means to overcome these problems by employing two main technologies. To support the exchange and communication between several sources of data, the grid database technology will be employed on the fast Internet2 backbone. Then, XML-based DNA data will be transported between collaborative sources for further analysis and representation. A preliminary version of our Web-based viewer for the XML data of the *Oryza sativa* genome is presented to illustrate the idea.

INTRODUCTION

Bioinformatics researchers around the world are now facing an obstacle related to a data integration problem. Not only that the molecular and genomic data required to represent the picture of the whole genomic system are dispersed across the world in different sources of data, but also each data source usually resides in its own format. Therefore, most algorithms and analysis tools only operate on locally available data and are not extensible to handle different data formats. The need to integrate data without disturbing local data accession and analysis tools requires a middleware layer and distributed data technologies, which have already proven to be feasible solutions to similar problems in business applications and the pharmaceutical industry [1–3].

XML

A middleware layer that can communicate between a diverse set of biological databases has to have a "multilingual" module that can understand the data that will come from different sources that have their own formats. XML (eXtensible Markup Language) [4] is a natural choice to enable biologists to create a common language for sharing information. The language is a specification developed by the World Wide Web Consortium (W3C), an international body of companies involved in defining standards for

*Plenary lecture presented at the International Conference on Bioinformatics 2002: North–South Networking, Bangkok, Thailand, 6–8 February 2002. Other presentations are presented in this issue, pp. 881–914.
‡Corresponding author: E-mail: royol@nectec.or.th. Other E-mail addresses: W. Rungsarityotin, wasinee@nectec.or.th; N. Khiripet, khirin@nectec.or.th; C. Tanpraset, chulak@nectec.or.th.

the Web, to express the structure of data without focusing on the presentation. Therefore, file formats can be easily explained through data type definitions (DTDs) or schema of an XML document [4].

XML has emerged as the dominant common exchange language to represent biological information, especially as the need to understand and correlate publicly available data increases. Consequently, XML will play a significant role in building an integrated environment that can evolve to accommodate the unpredictable growth of public repositories and diverse sources of data. This environment will enable researchers to ask and answer molecular and systems biology questions in a manner that makes optimal use of all the available information.

With many XML standards for data exchange, an advantage is that biologists can benefit from the development effort from the computer science community to provide a better query system that can relate information from different data sources. Relevant data for the rice genome, for example, are in different data formats and reside at several scattered sources across the world. The data are not readily available for common use or unified queries. XML will play an important role in representing different data structures, display of content, and format of the query language, in devising a unified framework for molecular and organismal biology.

Grid data system

Grid technologies enable sharing of bioinformatics data from different sites by creating a virtual organization of the data. The current grid-enabling software technology, such as Globus toolkits [5–7], allows the sharing of geographically distributed data. Therefore, the grid is able to help reduce the single point of failure inherited in a centralized database system. New research results can be stored on a local system and shared with the research community immediately. Users no longer need to know the location of their target information, but are able to access and retrieve in a transparent manner. This paradigm is extremely appropriate for large-scale genomic and proteomic activities.

Having different sources of data in different locations raises compatibility issues in the design of the grid computing infrastructure. The middleware layer therefore will interface with the grid services, such as security infrastructure and allocation manager, to transform the data from different sites to a standard format. In addition to providing a seamless access to the data repository, the grid infrastructure is ideal for parallel/distributed applications. Since the grid is a collection of computer-related resources, geographically distributed computing power can also be utilized in a similar manner to the aforementioned data-sharing concept. Computationally intensive tasks such as dynamics simulation or network modeling can be executed efficiently using all the computational resources available on the grid. The primary challenge is to create a software system that can manage the distributed computing components in a bioinformatics application so that they can access distributed data efficiently, i.e., minimizing the incurred communication costs by deciding if data are going to be sent to a system for computing or if a computation object (i.e., program) should actually be executed where the data are located.

The data grid infrastructure is shown in Fig. 1. The location(s) of the requested data will be supplied by the application's search engine running on the underlying grid data system. Once the locations are identified, data that allow the fastest access will be selected and processed according to the specified computation method. This process can be recursively applied in cases where the computation requires more input that cannot be found locally. Figure 1 portrays the data grid architecture assuming multi-institutional operation in a heterogeneous data and computing environment. The architecture presented shows the generic grid data services. The grid security system must be implemented as a gateway for every grid site providing uniform access to heterogeneous storage systems, which can be databases from different vendors, XML files, and even proprietary bioinformatics files. Furthermore, the grid security system will support the single signed-on capability so that the access authentication should be verified merely the first time. The storage system application programming interface (API) will provide the uniform accessing service to different (heterogeneous) storage systems. Users need not know how each storage system is operated. The meta-data catalog is implemented on top of the open standard

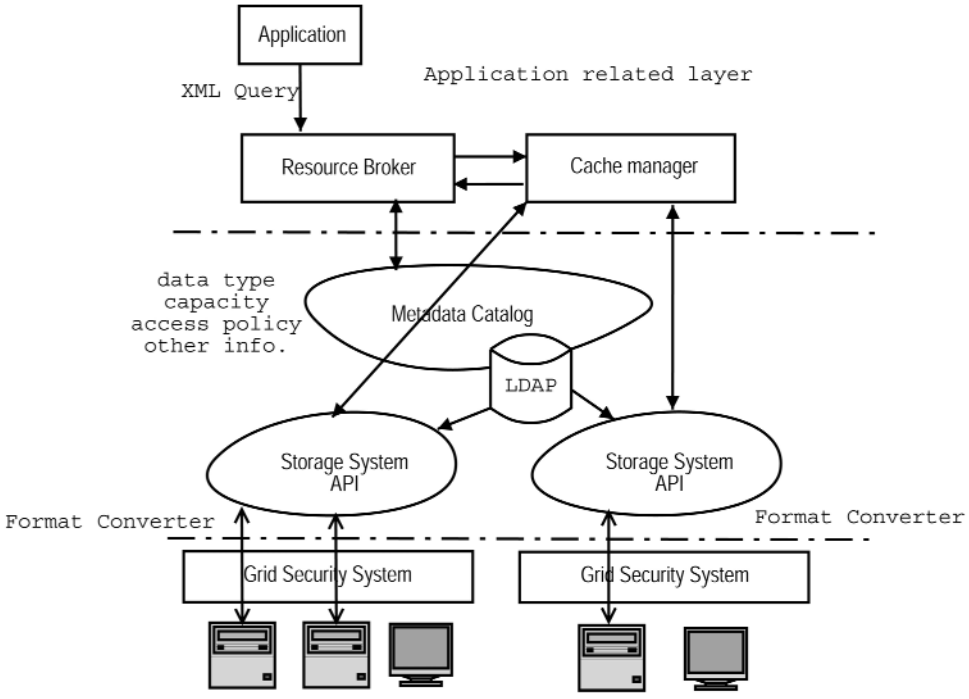

Fig. 1 Structure of our future grid data system. The architecture of the data grid system that allows accessing to large distributed heterogeneous data repositories and remote bioinformatics service providers.

protocol, called LDAP. This catalog stores storage system information such as information about file instances in the storage system. The top layer of this architecture presents the application-specific behaviors such as data access policy in cache management service. By doing so, we can encourage the reusability of basic grid data architecture while allowing users to tailor their applications to achieve high performance.

PRELIMINARY RESULT: A CASE STUDY ON THE RICE GENOME

The Rice Genome Research Program (RGP) has generated tremendous amounts of data in the past seven years, including data on DNA sequencing obtained from large-scale cDNA analysis; data on genetic mapping, including the locus position of DNA markers; and data on physical mapping of the rice genome [8,9]. These data are carefully and systematically recorded in computerized databases to facilitate analysis, storage, and retrieval. Data on this scale require extensive maintenance, continuous modifications and updates, interactive analysis, and easy access for both direct and indirect users. To arrive at a higher level of interpretation, the integration of rice DNA sequence data with molecular marker information is an essential infrastructure. The Thai Rice Genome Program, whose members include the investigators from Thailand listed in this proposal, have built an autonomous and loosely integrated resource to publish and maintain experimental data involved in the sequencing effort of rice in Thailand.

We have created a database that contains the most up-to-date collection of released sequences available publicly. In conjunction with the available sequence-ready map, bacteria artificial clone

(BAC)-end sequence (BES), and BAC fingerprint contigs (FPC), the sequence-based molecular markers selected from the integrated genetic map are anchored sequence-by-sequence onto each BAC. These molecular markers directly link genetic maps to the sequence-ready map. For any quantitative trait loci (QTL) of interest, tightly clustered markers directly link the corresponding critical regions onto the sequence-ready map of the Nipponbare strain of rice. BAC genomic sequences from Thai rice are aligned by sequence and marker contents with the corresponding sequence of the Nipponbare strain of rice. Express sequence tags (ESTs) of both Nipponbare and Thai rice located nearby can be treated as candidate genes. A complete map with molecular markers is in the process of being constructed for display on the Web as illustrated in Fig. 2.

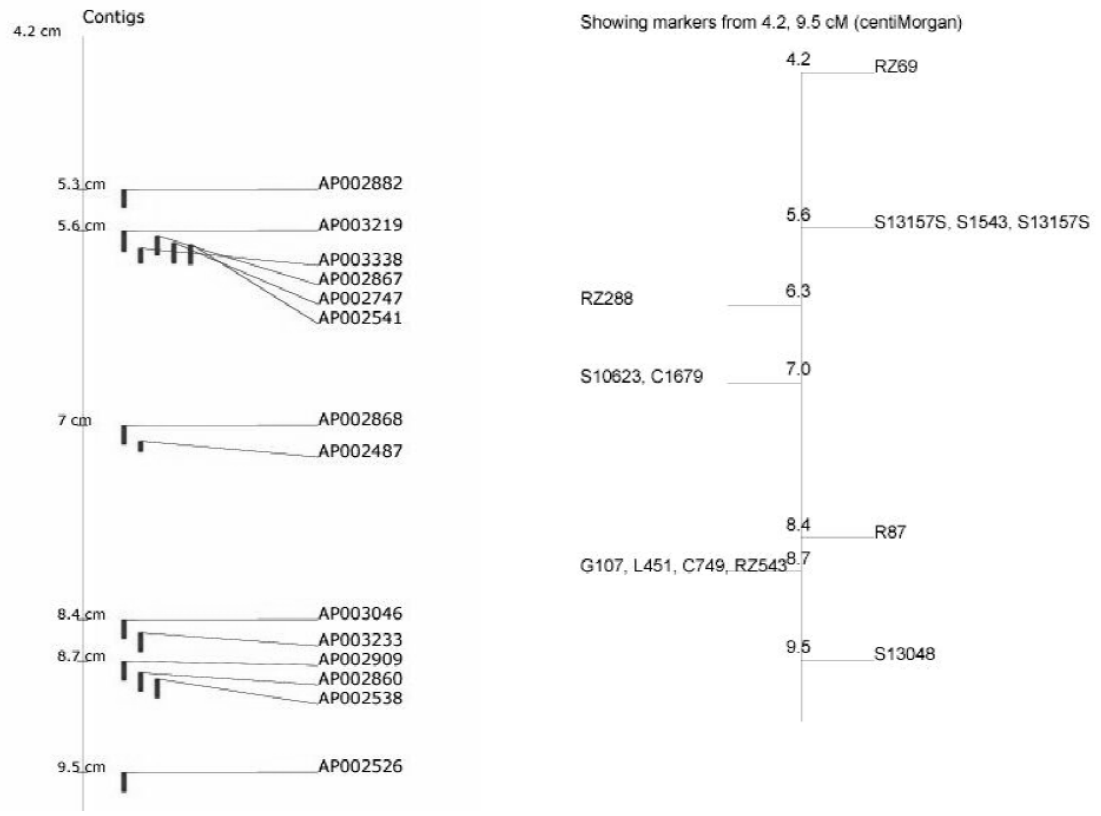

Fig. 2 Illustration of our Web-based interface showing markers from the rice genome. In the future, a physical map of BAC sequence from chromosome 1 of the rice genome will be integrated with a marker map on the right using XML technology. The user can eventually choose between textual or graphical output. With minimal programming involved, the graphics shown here was generated by simply transforming XML documents to scalable vector graphics (SVG), which can be readily displayed on a compatible Web browser.

DISCUSSION AND FUTURE WORK

The next phase of this research is to construct both the grid data structure and the query engine to respond to specific bioinformatics questions. Our system will deal with complex biological questions that can be broken down into several steps. Each step involves output elements that are produced by programs or services of the previous step, which are operated on various data sources. New information

from the products of each step will be added to our meta-database for later retrieval of related questions. Therefore, the final answer of the question will be only a part of the information stored in our meta-server. For example, a user specifies the type of information they wish to obtain by specifying various constraints, such as, "Display proteins in rice that have families with a size greater than 20 members with at least one known structure, whose corresponding gene expression is activated under dry conditions, and that are involved in interactions with at least 2 other proteins." [10].

The question can be posed in a query form using a simple interface such as one in Fig. 3, which will be translated to a corresponding XML message.

In principle, the translation of the query form (in Fig. 3) containing the query will result in an XML document. Assuming we have an imaginary set of XML tags, the translation can be defined as follow:

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE METASERVER_QUERY>
<SEARCH>
<key>protein</key>
<criteria="family size">20+</criteria>
<criteria="known structure">1+</criteria>
<criteria="gene expression condition">dry</criteria>
<!— This quality will be matched to a quantity equivalent to being dry —>
<criteria="protein interactions">2+</criteria>
</SEARCH>
```

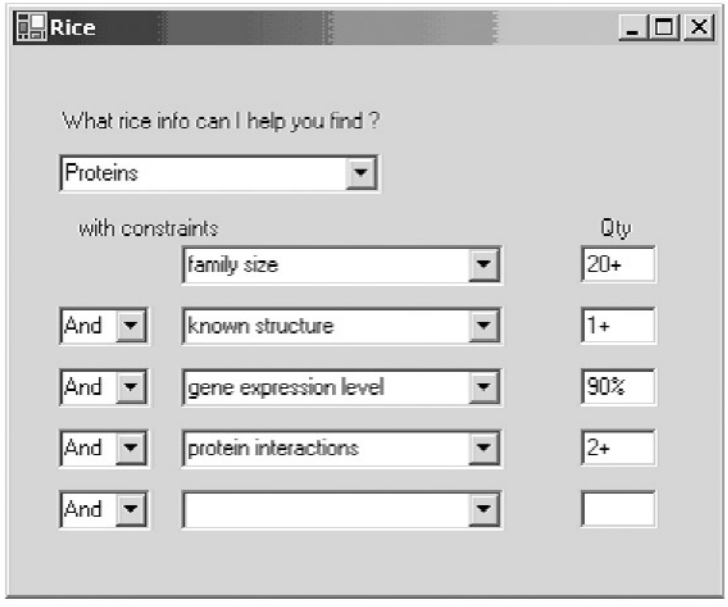

Fig. 3 Query form. We will provide a simple interface for posting queries. The user must provide which services they want, as well as constraints on the search. In this example, the service is finding proteins, using an intersection of these constraints: the size of the family, the matching to known structure in the protein database (PDB), the level of gene expression, and interactions with at least two or more proteins.

To answer the query, the computation may happen at local or remote sites depending on whether the local server has the tools needed to solve the problem. The public databases needed for these questions include protein sequence and structure databases, gene expression databases, two-hybrid databases, and functional annotation databases that will be presented in our grid data system. Then, the query steps can be broken down as follows (illustrated in Fig. 4):

1. Find genes in rice using gene-finding programs on the rice nucleotide sequence database.
2. Compare proteins encoded by the rice genes to all proteins using different sequence comparison methods and cluster into families.
3. Determine which families have experimentally determined or modeled structures (link to PDB [11] and model databases).
4. Link to gene expression data (exchange format based on GEML or MAML [12,13]).
5. Link to two-hybrid data (not yet specified).
6. Link to functional annotation databases (PROSITE [14]).

The XML-based query message will also encapsulate two important pieces of information, the data attributes and its computation methods. The attribute part is the specification of the requested data, which helps identify the location in the searching process. The method of computation is the algorithm, along with specific arguments used to process the data. Once the data locations are identified, it will be transferred back through the data grid system for processing at the local site. If those computations can be performed at the site with the acquired data, it will be processed at the remote site, depending on the application's data access policy. On the other hand, the algorithms can be converted to portable code such as Java that can then be sent to the original data location. The grid data system presented here has shown itself as a promising tool and infrastructure that can support the data acquisition, translation, and distributing computation tasks required by bioinformatics development.

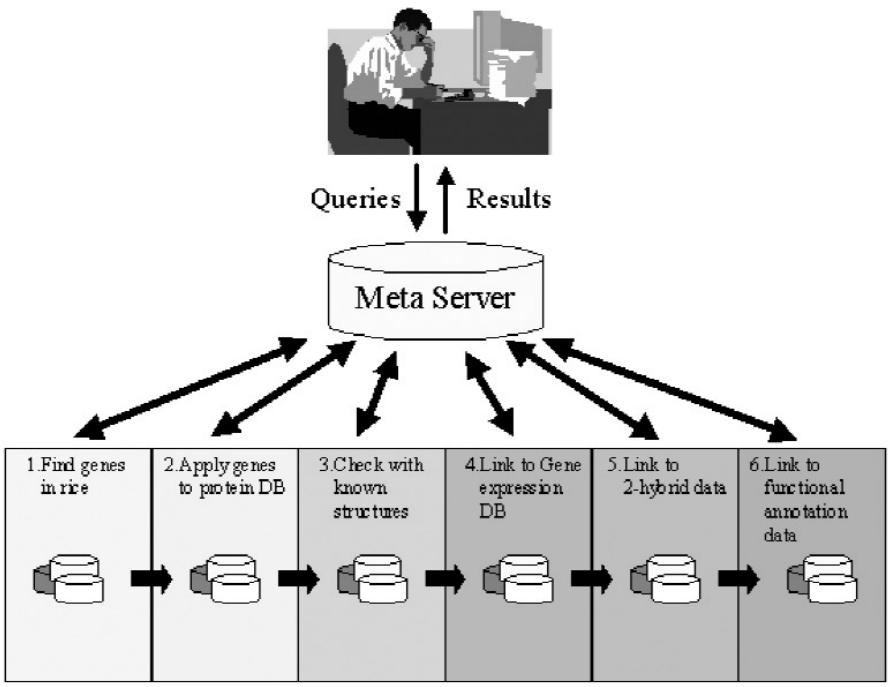

Fig. 4 Steps taken to process a complex query input using the interface depicted in Fig. 3.

ACKNOWLEDGMENTS

We would like to thank Dr. Apichart Vanavichit and his team (from DNA Technology Laboratory, Kasetsart University) who have helped obtain data of BAC clones and markers used in this work, including references about the International Rice Genome Program. We also thank Ram Samudrala (computational biology group, University of Washington, Seattle, Washington, USA) for useful discussion about future plans and Sissades Tongsima (high-performance computing division, NECTEC) for information and references about research on the grid data and computing.

REFERENCES

1. M. T. Roth and P. Schwarz. In *Proceedings of the 23rd VLDB Conference*, Athens, Greece (1997).
2. DiscoveryLink <http://www3.ibm.com/solutions/lifesciences/discoverylink.html>.
3. C. A. Goble et al. "Transparent access to multiple bioinformatics information sources", *IBM Sys. J.* (2001).
4. XML main Web page. W3C. eXtensible Markup Language (XML) <http://www.w3.org/XML/>.
5. I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. (1997).
6. I. Foster and C. Kesselman. *The Globus Project: A Status Report*, pp. 4–19, IEEE Computer Society Press (1998).
7. S. Vazhkudai, S. Tuecke, I. Foster. In *Proceedings of the First IEEE/ACM International Conference on Cluster Computing and the Grid (CCGRID 2001)*, pp. 106–113, IEEE Press, May (2001).
8. The Rice Genome Research Program Web site (IRGP). <http://rgp.dna.affrc.go.jp/>.
9. Monsanto Rice Genome Research Web site. <http://www.rice-research.org>.
10. Personal communication with Dr. Ram Samudrala, University of Washington, Seattle, Washington.
11. The RCSB Protein Data Bank (PDB). <http://www.rcsb.org>.
12. Rosetta Inpharmatics, Inc. Gene Expression Markup Language (GEML). <http://www.geml.org>.
13. Microarray Gene Expression Databases (MGED). Microarray Markup Language (MAML). <http://www.oasis-open.org/cover/maml.html>.
14. PROSITE Database of protein families and domains <http://www.expasy.ch/prosite/>.