行政院國家科學委員會專題研究計畫　成果報告

整合型生物資訊處理系統之建造

計畫類別：個別型計畫
計畫編號：NSC91-2218-E-009-015-
執行期間：91 年 11 月 01 日至 92 年 07 月 31 日
執行單位：國立交通大學資訊科學學系

計畫主持人：陳俊穎

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 10 月 31 日

計畫中文摘要

生物資訊系統的整合為現階段重要研究課題之一，而網路上已經包含大量的生物資訊服務。例如基因體序列資料庫（NCBI GenBank, EBI EMBL），蛋白質資料庫 (PDB, ExPASy SWISS-PROT），及相關之資料存取介面和各種資料分析的工具。雖然這些生物資訊服務可讓大眾從網路上讀取資料，但并不是彼此完全的整合在一起。其中一個主要的原因是因為這些網路資源並不是針對工具本身的溝通而設計，而是針對生物學家實際上的使用需要。

我們的目的是在網際網路上建造一個完全整合的生物資訊服務系統，以提供近期內生物資訊大量資料處理之需求。此系統應用網路服務(Web Services) 概念，可將現行生物資訊系統轉變成容易組合及更改的單元. 我們除了將繼續改進與為持此基層網路服務架構外，並將與生物學家合作擴增不同生物資訊服務. 我們相信本系統能大幅提升現行系統的價值.

關鍵詞 生物資訊 基因體 序列資料庫 網路服務 分散式系統 元件式系統發展

計畫英文摘要

Integration of bioinformatics systems has been one of the research topics receiving much attention in recent years. Currently, there are numerous publicly available bioinformatics databases and analysis tools that are widely used among biology researchers. However, most of these popular bioinformatics services are web-based and therefore not integrated enough for effective data sharing and high-throughput data analysis purposes. The problem can be attributed largely to that fact that these services are designed primarily for human access, rather than for incorporation with other applications.

To address this issue, we have been developing a suite of integrated, web-based bioinformatics services base on the web services concept - software applications running over Internet and exchanging data via standard XML-based protocols. In particular, we designed and implemented a distributed computing framework in which popular bioinformatics services can be turned into web services and their functionality transformed and combined much easier, thanks to the flexibility of XML and standardization of the communication channel among these services. As a next step, in addition to continuously improving our web services infrastructure, we will gradually incorporate different types of bioinformatics services into the framework by collaborating with other biology researchers, and build more intuitive and systematic interfaces for biological data analysis. We believe such a framework can generate much more value than the sum of those by individual, partially integrated services that is common today.

Keywords bioinformatics, microarray, genomics, sequence database, web services

Introduction

Integration of bioinformatics systems has been one of the research topics receiving much attention in recent years. Currently, there are numerous publicly available bioinformatics databases and analysis tools that are widely used among biology researchers [1-5]. Examples include genome sequence databases such as NCBI GenBank, protein databases such as PDB, as well as their associated web-based retrieval interfaces and data analysis packages. However, most of these popular bioinformatics services are not sufficiently integrated for effective data sharing and high-throughput data analysis purposes. The problem can be attributed largely to that fact that these services are designed primarily for human access (via web browsers), rather than for incorporation with other applications.

To address this issue, there are many integration attempts (for example, [6, 7]). We have been developing a suite of integrated, web-based bioinformatics services base on the web services concept - software applications running over Internet and exchanging data via standard XML-based protocols [8, 9]. In particular, we designed and implemented a distributed computing framework in which popular bioinformatics services can be turned into web services and their functionality transformed and combined much easier, thanks to the flexibility of XML and standardization of the communication channel among these services.

As a next step, in addition to continuously improving our web services infrastructure, we will gradually incorporate different types of bioinformatics services into the framework by collaborating with other biology researchers, and build more intuitive and systematic interfaces for biological data analysis. We believe such a framework can generate much more value than the sum of those by individual, partially integrated services that is common today.

Objective

Our ultimate goal is to develop a suite of fully integrated, Internet-based bioinformatics services in order to meet the emerging demands on high-throughput bioinformatics data analysis, with an aim to make biologists' work easier. On the technical side, we are working on a more suitable, Internet-based *infrastructure* for data sharing and research collaboration. In addition, through the design and implementation of the system, we hope to establish expertise in developing large-scale, reusable, *component-based* software.

The immediate goal for this "pilot" project is for the newly joining project participants, from computer science perspective, to investigate and be familiar with the three areas indicated above, namely, biology, distributed computing, and software engineering, so as to prepare for collaboration with biologists and construction of

larger-scale systems in the near future. We hope to identify key elements and issues in each of these areas, and to construct a *minimal prototype* to exercise and validate the conclusions obtained along the way. More specifically, we are planning to build the prototype such that

- For biologists, the system can retrieve some basic genomic data (e.g. nucleotide and protein sequences, genes, etc.), possibly from the Web, and store them into a local database. It will provide common operations on the stored data (e.g. BLAST) as well as some visualization tools (e.g. sequence browsing, gene information display). Both web-based interface and standalone GUI front-end will be developed for ease of use.
- For data sharing and collaboration, multiple instances of the system running on different machines will be able to access data and computational resources from one another. Standard data format and communication protocol will be defined, if necessary.
- For system architecture, we will strive for modularity and extensibility for the design so that new data types and computational tools can be added gradually.

To summarize, we want to apply modern distributed computing technology as well as component-based development practices to the domain of bioinformatics. Through such a practical path we expect to broaden our knowledge and advance the technologies in both biological research and software development. For biology community, our work will provide an enabling vehicle for convenient access to existing bioinformatics services, and stimulate further development in new ones.

Approach

To ensure a sound system architecture and design, several key requirements need to be met.

- With the vast quantities of biological data ready and growing world wide, the system should help biologists make sense of the data and to extract conclusions and hypotheses that are biologically meaningful. To support this, services for mass data storage and high-performance computation are needed so as to support efficient data mining solutions.
- Currently, data with different formats and varying semantics are scattered around among different software systems, possibly distributed across network. It is desirable to have systematic ways to unscramble these massive, heterogeneous, distributed, yet connected data. More importantly, the semantic differences among these biological data should be controlled properly in order to enable valid data exploration and summarization (see [7, 8]).
- Large-scale biological data analysis requires utilization of distributed computing resources. One consequence is that software systems and tools, each with

different target problem domain, need to be able to collaborate with one another seamlessly without human intervention. Besides the semantics issue above, simple and effective mechanisms for software collaboration are important.

In fulfilling the goal, however, we do not try to develop a full-blown system at first attempt. In this project, we concentrate on the last issue above, i.e. infrastructure for the coordination and collaboration of distributed bioinformatics services. Because current standard distributed computing platforms still remain complicated for application development, we developed a simple distributed computing framework based on the on the web services concept - software applications running over Internet and exchanging data via standard XML-based protocols.

Results and Discussions

Figure 1 shows the architecture of our framework schematically:

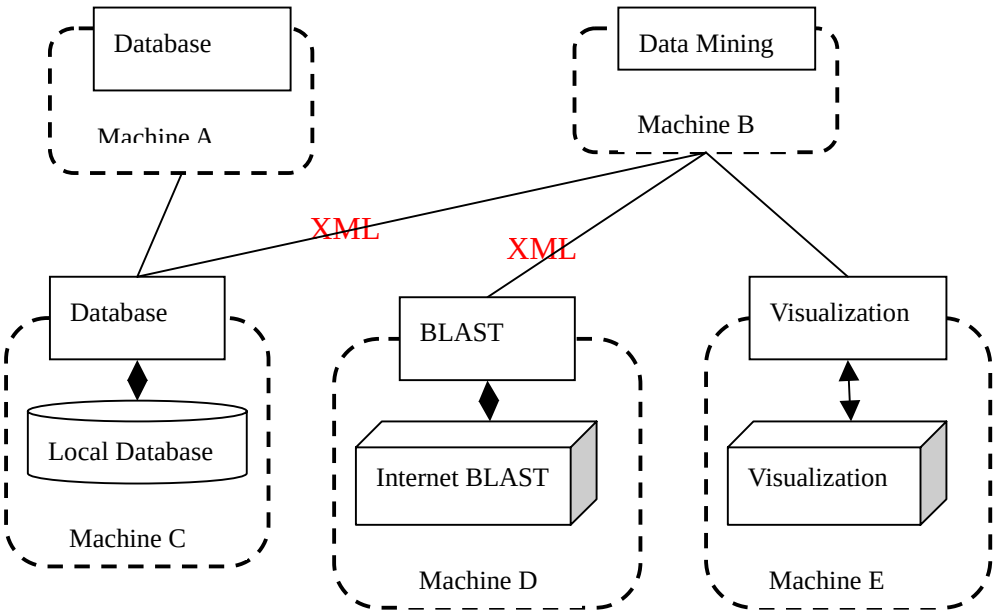

Figure 1. Schematic view of the system.

In the figure, different bioinformatics services running on different machines understand the syntax and semantics of the XML messages offered by other services. Some services, e.g. Visualization, may require supports from other services, while others may be self-contained. To make it clearer, consider some of the data analysis services below that have been implemented with the framework:

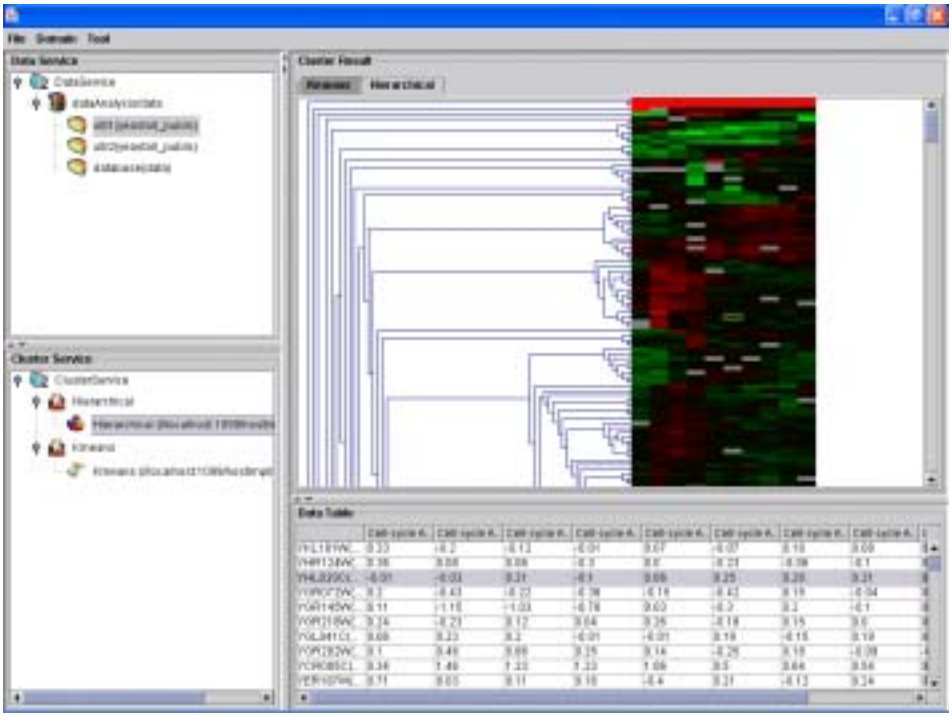

Figure 2. Distributed services that together support Hierarchical Clustering

The example configuration of the system above contains four distributed services.

1. Data storage service (upper left) primarily for storing numerical data.
2. Data analysis service (lower left) providing some analysis algorithms
3. Visualization service (right) providing visualization support given appropriate input data
4. Data mining service (the whole interface), a graphic console integrating and interacting with the other services. It is worth noted that this service is a thin client in that all essential resources provided by previous services are remote to the client, and all the interactions are done with XML messaging.

Although as mentioned before that we are concentrating on the distributed computing infrastructure in this project, we have also allocated resources on other requirements along the way. For mass data storage and high-performance computing, we are currently developing a service exploiting the extreme expressive power offered by relational databases, including SQL that offer flexible data querying and formatting, and store procedures that allow processing massive data without converting data between storage and memory.

On the other hand, for data semantics issue, we are also developing a simple data and

service description framework so that there are unambiguous definitions for different services and data types. In fact, although not explicitly stated, the implementation above has taken this into consideration. For example, each of the services has a type identifier that uniquely identifies what each service should do, which constrains the XML format it can accept and response. In addition, the relations among services are not fixed. Instead each service should declare what types of services it needs in order to fulfill its own functionality, and it is up to the application (the data mining console above, for example) to allocate appropriate services to meet the requirement.

References

[1] GenBank, National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, USA. http://www.ncbi.nih.gov/Genbank/index.html

[2] EMBL, European Bioinformatics Institute (EBI), Wellcome Thrust Genome Campus, UK. http://www.ebi.ac.uk/Databases/

[3] DNA Data Bank of Japan (DDBJ), Center for Information Biology, National Institute of Genetics, Japan. http://www.ddbj.nig.ac.jp/

[4] The Protein Data Bank (PDB), http://www.rcsb.org/pdb/

[5] ExPASy Molecular Biology Server, http://www.expasy.org/

[6] Siepel, *et al.* ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources, Bioinformatics, vol. 17, no. 1, pp 83-94.

[7] Lion bioscience SRS, http://www.lionbioscience.com/solutions/products/srs

[8] W3C Web Services Activities, http://www.w3.org/2002/ws/

[9] The UDDI Project, http://www.uddi.org/

[10] Gene Ontology (GO) Consortium, http://www.geneontology.org/

[11] Microarray Gene Expression Data (MGED) Society, http://www.mged.org/

Evaluation

The proposed items to accomplish in the project proposal are summarized below:

- To proper formulate system requirement and specifications, acquire knowledge in biology, investigate available Internet bioinformatics resources, and interact with researchers in biology and understand their needs.
- To devise sound architecture and design, study related fields including distributed computing platforms, software engineering, object-oriented methods, component concepts, database and data mining techniques.
- To be familiar with actual system development process, including the practices of software engineering concepts and methods, and the use of various commonly

used software development tools.

In short, this project is the first step toward a fully integrated, high-performance bioinformatics system, and the primary objective is to train the participants in each of the topics above. In this regard, the project objective is met in that students are much more experienced in design and programming (in Java) now for more challenge problems, although training in biology in general and bioinformatics in particular still have room for improvement.

In addition to the training objective, there are two other accomplishments that we think exceeds what was planned. Firstly, we have established a more systematic development process and standardized development tools used, including the Java programming language, CVS (Concurrent Version Systems), Eclipse IDE (Integrated Development Environment), TWiki (a web-based collaborative tool for concept exchange among developers). Secondly, we have stabilized the web service infrastructure (rather than a prototypical one) that support both traditional socket communication and Java RMI (Remote Method Invocation). This design not only simplifies future service development in our lab, but also enables independent service development using other programming languages.