WWW Issues For Conducting Sophisticated Bioinformatics Analysis

David Schibeci and Matthew Bellgard, and Kim Carter (Corresponding Author)
Centre for Bioinformatics and Biological Computing [HREF1]
Division of Business, Information Technology and Law
Murdoch University [HREF2], Murdoch WA 6150, Australia
Email: carter@arginine.murdoch.edu.au

Abstract

The rapid explosion in the amount of biological data being generated worldwide is outpacing efforts to manage the analysis of the data. Bioinformatics researchers wishing to perform analysis on molecular data have a plethora of local and Internet-based tools available to perform different data analysis, but as researchers often wish to analyze and re-analyze data in a manageable, intuitive way, the problem arises of how to integrate these tools and the different data formats they encode. In addition, there are now new information technologies like XML and CORBA for representing and distributing biological data. The major questions that arise are: i) is it possible to develop a web-based information system for downstream, multiple pass analysis with the emphasis on analysis rather than storage? ii) can this information system facilitate and manage hypothesis driven analysis? iii) can this information system be adaptable to changing information technologies as well as data and information from new bio-technologies? iv) can this information system be able to ensure the analyzed data remains up-to-date in light of new data as well as reporting new information as it becomes available? In an attempt to answer some of these questions, we have designed a web based system Information System for whole Genome Data (ISGD). We discuss relevant issues for conducting sophisticated Bioinformatics analysis, and in addition, we review and discuss the latest technologies, like XML and CORBA, that impact the ongoing development of ISGD.

Introduction

The massive increase in computer processing power and advancement of sequence analysis techniques over the past 5 years has seen an incredible increase in the amount of biological data required to be created, stored and analyzed. The EMBL [HREF3] nucleotide sequence database contains almost 11 billion bases in 9.9 million records (EMBL, December 2000). While EMBL has been operating since 1982, the last five years has seen a huge increase in the size of the EMBL database. Since 1994, the EMBL database size has nearly doubled every year, and has tripled over the last 12 months alone. The draft release of the human genome sequence (June 2000) takes up over 3 gigabytes of computer storage space, as part of the latest EMBL release (Release 65 HTG). However, with the ever increasing volume of bioinformatics data available on the World Wide Web, the platform for downstream analysis, re-analysis and storage, management and distribution of analyzed data becomes ever more critical.

A report to the US National Science and Technology Council (NSTC, 1998) identified that "sequencing data and related datasets are growing at an exponential rate, far outstripping efforts to manage and analyze these data". Boguski (1999) discusses how sequence analysis is approaching a "wall" in terms of its ability to reveal reliable and detailed inferences from sequence data. This wall is attributed to the accuracy and organization of the data, and the
reliability and consistency of annotation rather than the speed and sensitivity of alignment programs. These issues are also highlighted elsewhere, for example Anderson and Bansal (1999), Bellgard (1999a), Walker and Koonin (1997).

Whole genome comparative analysis can be considered to be hypothesis driven and thus a researcher requires the ability to easily ask "What if" questions to test theories on genome organization, structure and evolution. While "warehouses" of data (eg EMBL [HREF3], GenBank [HREF30]) as well as web sites containing bioinformatics tools (eg [HREF4]) are suitable for single pass analysis, such as comparing a new sequence against a public database or calculating the reverse complement of a sequence, what is required is the facility to take the results of one analysis as the basis for conducting further downstream analysis (eg identifying new genes - Bellgard, 1999b) in a manageable, efficient way. With the diverse range of file formats (eg. FastA, Staden, Pearson [HREF31]), different computer platforms, as well as representations of bioinformatics data (eg. GEML - [HREF5] and BioML - [HREF6] ), it has become an increasingly daunting task working with different analysis tools. Researchers wishing to perform multiple-pass analysis of data, by feeding results of one program to another, encounter the problem of changing data from one format to another. This can be time consuming, frustrating and, more than likely, result in errors. There arises the need to develop a standardized but flexible format for representing different types of biological data (eg. primary sequences, functional data, ESTs) (Robbins, 1996a, 1996b, 1995).

GEML, BioML and OpenBSA ([HREF7]) are formats that have been developed to describe biological data in a structured but flexible manner, using XML (extensible markup language [HREF8]) standard. XML is a meta-language for defining markup languages (Eg. HTML) to produce documents that convey content with semantic structure (Elenko & Reinertsen, 2000). A BioML document, for example, may contain the sequence and associated reference information for a particular virus. While XML is a way of structuring the vast amount of biological data available, it's more complex than a simple text format and problems arise such as how to convert existing non-XML data to the XML format (Elenko & Reinertsen, 2000). However, with the plethora of text formats available and the problem that simple text files do not represent scientific data adequately, to perform high throughput data analysis, we need a more structured, manageable data representation (Robinson, 2000).

With a plethora of bioinformatics tools available (Bellgard, 1999a), the difficulty of learning each individual tool and integrating various tools, due to differing data and communication formats, there arises the need for a standard communications architecture that will allow different bioinformatics tools to communicate through the same common interface. The Common Object Request Broker Architecture (CORBA) is an open object oriented architecture developed by a worldwide consortium of vendors, developers and users (The Object Management Group - [HREF9]) to standardize interoperability of systems in distributed heterogeneous environments (Parsons et al 2000; Diskin & Chubin, 1998; Hu et al, 1998; Vinoski, 1997). CORBA provides interoperability between databases and client applications (Robinson, 2000), allowing cross platform, cross application support through standard interfaces. CORBA is becoming a popular choice distributed bioinformatics systems and can be seen in projects like ArkDB (Hu et al, 1998), Goldie (Anderson & Bansal, 1999), JESAM (Parsons & Rodriguez-Tome, 2000) and in the CORBA EMBL and Radiation Hybridization databases at EBI ([HREF10]).

A number of distributed architectures other than CORBA are also available. DCOM ([HREF11]) and DCE ([HREF12]) are seen as direct competitors (Diskin & Chubin, 1998) to CORBA, while architectures like those developed by Buttner et al (1999), Stonebraker et al
(1996) and Homburg et al (1996) appear to be designed for specific purposes rather than
generic object distribution. However, studies like BioORB (Cooksey et al, 1997) and Kemp et
al (2000) discuss how CORBA can be used to provide efficient access to bioinformatics
resources. CORBA has a number of advantages over other distributed architectures including;
platform independence and broad programming language support, an open public
specification, support for legacy systems and support from a large consortium - OMG (Diskin &
Chubin, 1998). While CORBA does have problems with interoperability between vendor
implementations, it's an enabling technology towards distributed integration of resources
(Cooksey et al, 1997). The application of CORBA to bioinformatics is "starkly appropriate"
(Parsons & Rodriguez-Tome, 2000) for a standardized access to data and services when
bioinformatics researchers are investigating / analyzing across different data sources.

Robbins (1996a, 1996b, 1995) identified three groups of problems for distributing complex
and diverse biological data. These are i) Technological - integrating distributed heterogeneous
databases; ii) Conceptual - unifying different semantic views on the meaning/interpretation of
data; iii) Sociological - getting projects to agree on technological and semantic standards.
Bioinformatics lends itself to an open source architecture with standardised data formats,
integrated tools and resources, using agreed-on standards while allowing customization for
hypothesis driven research. Letondal(2001) highlights several drawbacks in commercial
systems, such as BioNavigator ([HREF13]), that attempt to address the identified
infrastructure issues. A number of freely available systems have been designed, however they
are still in prototype stage. For instance, ISYS ([HREF14]) is designed to allow navigation of
biological data using graphical interfaces and visualization tools, though the current release is
a demonstration version. Goldie (Anderson & Bansal, 1999) uses distributed processing
techniques to improve the performance of comparing and aligning gene sequences. SEALS
([HREF15]) is a system of small command-line programs designed for sequence analysis of
large amounts of data. The Roslin Institute (Edinburgh) has developed ArkDB ([HREF16]), a
genome database for mapping data for single species. It would appear that the challenge still
exists to develop a comprehensive open source architecture for bioinformatics, building on
the strengths of existing designs.

An ideal system for whole genome analysis would at least include an Internet based client
/ server architecture to allow remote and local access to the system. The ability to expand
the system, via simple addition of modules would allow the system to evolve as new
biotechnologies, such as micro-arrays, become available. An automated primary and
secondary database update and report system would enable the internal data stores to
remain consistent, accurate and reliable, with the ability to incorporate information flowing
from experimental validation, such as protein / protein interaction and pathways. An essential
feature would include a quality assurance process, to allow quick re-annotation of previous
results in light of new data. Clustering, by distributing tasks to multiple machines, would
allow the system to take advantage of available processing power to increase efficiency of the
system. Integration and / or development of visualization tools to allow multiple views of
data, annotation, comparison, and comprehension in a graphical environment is highly
desirable to allow researchers to get a better picture of their analysis and results.

In light of these desirable features, we have designed a prototype system, the Information
System for whole Genome Data (ISGD). ISGD has a single common data representation to
handle the diverse range of biological data formats. The ISGD architecture is designed to
allow different distributed systems to communicate through the same common ISGD
interface. The prototype ISGD is designed to allow easy access to distributed heterogeneous
biological resources, researchers, enabling researchers to perform efficient and effective
WWW issues for conducting sophisticated Bioinformatics analysis

**Information Systems Genome Database (ISGD)**

**System Overview**

For bioinformatics researchers to have access to an integrated genome database, analysis and visualization tools, an easy to use graphical user interface (GUI) environment is required. ISGD provides database access, analysis and visualization tools, all of which are accessed via a World Wide Web server. ISGD has automated tools to keep databases up-to-date, report new analysis and new databases, and inconsistencies and revisions to raw genome sequences. ISGD features i) interfaces to various analysis and query based languages; ii) interfaces to visualization tools for raw, annotated and analyzed sequences; iii) a library of input and output programs to ensure data integrity and integration for new analysis programs; iv) software agents to automatically perform database analysis (Bellgard, 1999a). ISGD is designed to be flexible, scalable, leverage clustering / supercomputing power, and to allow repeat analysis while providing automated quality assurance. The current release (Version 3.0) of ISGD is still in the developmental stage and consequently, ISGD currently handles only whole genome sequences from GenBank, though expressed sequence tag (EST) support is being added, and the intelligent agents, visualization tools and clustering components are in continual development. The basic ISGD architecture is shown in figure 1.

![ISGD architecture diagram](http://ausweb.scu.edu.au/aw01/papers/refereed/carter/paper.html)

Intelligent agents continually update the central database by obtaining new data from external data sources like EMBL and GenBank. The agents determine if updating is required in the central and secondary databases (not shown), and a report is generated with each update. Individual researchers can access the central and secondary (personal) database and store their work in the secondary database, which in turn may be "value-added" to the central database. Researchers access ISGD via a web interface, while standard, platform
independent Perl applications and modules are used to connect applications to the central database and external data sources. The database abstraction layer (DBAL) allows the database schema to be changed or updated without affecting the rest of the system.

**System Prototype**

GUI access to the ISGD system is desirable to aid researchers in visualizing the research process and results. The GUI environment designed for ISGD is based on the ORBIT (Bellgard, 1999c) project. ORBIT is an integrated GUI environment designed to provide customizable interfaces for bioinformatics programs. The client provides the GUI interface to the available bioinformatics tools, and allows the exchange of data between the client and server. ORBIT is designed to allow the interfaces for the tools to be customized for each researcher. The current ISGD prototype has a web (HTML) based interface, but a customizable GUI like ORBIT is planned for future releases.

To address the problem of data representation, ISGD data is stored in a standard internal representation. Data coming into the system is converted into SQL statements allowing the data to be compared and converted more easily, allowing researchers to perform the single and multiple-pass analysis of the required data. The ISGD prototype is implemented on Solaris (Unix) using MySQL databases. MySQL was chosen over other open source database systems as it's fast, efficient and supported the particular SQL statements we required. In the current ISGD prototype, incoming data is converted manually (using automated routines) to the internal representation. Future autonomous automatic conversion routines are currently being developed, so little or no human interaction is required for this process.

The architecture of ISGD allows local and remote analysis of the ISGD data through a common interface, while allowing different tools to be used by converting to and from the standard common data representation. This architecture allows us to connect to heterogeneous external systems and to pass results between different tools. This addresses the problem of interconnection heterogeneous data sources.

**Results**

When Bellgard (1999d) undertook a study to determine if evolutionary changes could be inferred by G+C differences, the initial study (without ISGD) took several months to complete. The study was later repeated, using ISGD, and was completed in a week. The study was further extended when Bellgard (et al, 2001e) used ISGD to perform a comparative analysis of two complete genomes to determine G+C differences in bacterial species. These positive results demonstrate the ISGD model is an appropriate platform for conducting sophisticated bioinformatics analysis.

**Discussion**

As ever more biological data is generated and analysis tools are developed, the ISGD architecture will be expanded to incorporate new technologies. The three integration problems identified by Robbins, data representation, communications architecture and standards adoption, will still need to be addressed in future versions, as no-one appears to have found the ideal solution, yet.
ISGD is designed for whole genome sequences, and could easily be modified to handle other biological data like ESTs, incomplete or fragments of gene sequences or micro array data. What is desired is a universal data format, along the lines that GEML and BioML describe, to represent biological data for ISGD. Both Netscape ([HREF18]) and Microsoft ([HREF19]) have added XML support to their web browsers, demonstrating that XML is recognized as being an important technology for describing and structuring data and how it's displayed in a browser. Future versions of ISGD will incorporate an XML format, based on existing formats, as the internal data representation for sequence and analysis data.

Communications Architecture

The current ISGD prototype is implemented on a single server. The databases and bioinformatics tools are run locally on a single server machine. Future expansion of the single server ISGD prototype to more distributed one would allow us to take advantage of clusters or dedicated machines. A service manager for each tool would provide a gateway to a number of remote or local servers with particular tools. For example, FastA [HREF32] searches are sent to the FastA service manager, and then distributed either whole or in parts amongst the available FastA servers. By distributing the tasks, the workload is relieved from the main ISGD server, and tasks can be performed more rapidly. CORBA is a communications architecture that would allow us to accomplish this task. CORBA support was added to the Netscape Communicator browser , illustrating the popular support for CORBA. Integrating CORBA into future releases of ISGD would allow platform independent remote access to the ISGD system, while allowing ISGD tasks to be distributed to local or remote servers.

Standards

To address the issues of cooperation and coordination of communication and data representation standards, we have decided to adopt CORBA and XML for future versions of ISGD. CORBA and XML are open standards, recognized worldwide and appear to be the next step for integration of systems. Even if they aren't, if some consensus could be reached on standards for interoperability and representation, the whole bioinformatics community could benefit from worldwide integration of biological resources.

Conclusion

The amount of biological data being generated worldwide is growing rapidly, as is the need for analyzing and managing this wealth of information. It appears that not enough is being done to address the technology infrastructure of bioinformatics research tools and networks. With the expansion of the Internet and bioinformatics research worldwide, down-stream analysis and re-analysis of biological data is being performed locally and remotely. We have designed a system to enable tools to be integrated more easily, and to handle the dynamic nature (continual updating or revisions) of biological data and down-stream analysis. By introducing a common exchange format and communications architecture, the ever increasing amount of bioinformatics data can be gathered, analyzed and distributed in a standard, integrated manner, allowing bioinformatics researchers to concentrate on gathering and analysis of data and relieving them of the burden of learning and utilizing individual stand alone tools.

References
Anderson V.S. and Bansal A.K., "A Distributed Scheme for efficient pair-wise comparison of complete genomes", *Proceedings of The International Conference of Information, Intelligence and Systems 1999*, 48-55


Bellgard M.I. and Gojobori T., 1999b, "Identification of a ribonuclease H gene in both Mycoplasma genitalium and Mycoplasma pneumoniae by a new method for exhaustive identification of ORFs in the complete genome sequences", *Federation of European Biochemical Societies, FEBS Letters* 445 1999 6-8


Bellgard M.I., Schibeci D., Trifonov E. and Gojobori T., 2001e (to appear), "Early Detection of G+C differences in bacterial species inferred from the comparative analysis of two completely sequenced Helicobacter pylori strains", *Journal of Molecular Evolution*


National Science and Technology Council (NSTC), 1998, *Bioinformatics in the 21st century*,

WWW issues for conducting sophisticated Bioinformatics analysis

Available Online [HREF27]


MySQL, 2000, MySQL, Available Online [HREF17]


**Hypertext References**

HREF1
http://wwwcbbc.murdoch.edu.au/

HREF2
http://www.murdoch.edu.au/

HREF3
http://www.ebi.ac.uk/embl/index.html

HREF4
http://www.sanbi.ac.za/~rmuller/SMS/

HREF5
http://www.geml.org/

HREF6
http://www.biomp.com/

HREF7
http://corba.ebi.ac.uk/openBSA/

HREF8
http://www.xml.org/

HREF9
http://www.omg.org/

HREF10
WWW issues for conducting sophisticated Bioinformatics analysis

Copyright

Kim Carter, David Schibeci and Matthew Bellgard © 2001. The authors assign to Southern Cross University and other educational and non-profit institutions a non-exclusive licence to use this document for personal use and in courses of instruction provided that the article is
WWW issues for conducting sophisticated Bioinformatics analysis

used in full and this copyright statement is reproduced. The authors also grant a non-exclusive licence to Southern Cross University to publish this document in full on the World Wide Web and on CD-ROM and in printed form with the conference papers and for the document to be published on mirrors on the World Wide Web.

[AusWeb01 Seventh Australian World Wide Web Conference, Southern Cross University, PO Box 157, Lismore NSW 2480, Australia. Email: "AusWeb01@scu.edu.au" ]