An Overview of Bioinformatics Research at Lawrence Livermore National Laboratory

Terence Critchlow[1], Ron Musick[1], Tom Slezak[2]
Center for Applied Scientific Computing[1]
Joint Genome Institute / Computing Application Organization[2]
Lawrence Livermore National Laboratory

Abstract

Depending on who you ask, bioinformatics can refer to almost any collaborative effort between biologists or geneticists and computer scientists – from database development, to simulating the chemical reaction between proteins, to automatically identifying tumors in MRI images. At Lawrence Livermore National Laboratory (LLNL), we have come to use a slightly more restrictive definition. We consider bioinformatics to refer to the development, application, and research of data management and data mining techniques and technology within the domains of genomics and molecular biology. This definition includes diverse tasks such as the creation of a database to contain protein sequence and structure data, the integration of existing genomics data sources into one database, the creation of databases to support high-throughput production genome sequencing, and the automatic construction of a model for interpreting micro-array results. This short paper provides an overview of the history of bioinformatics at LLNL, briefly describing the bioinformatics challenges we face, and outlining the ongoing efforts to meet them by our bioinformatics team and the DataFoundry research project.

1. Introduction

Twenty years ago, there was very little genomics data available. Biologists would spend months painstakingly performing experiments to sequence small pieces of DNA or proteins. Because of the small amount of data, and their intimate familiarity with it, scientists usually managed the data they generated – typically by creating flat files that encoded the experimental data in a format particular to their lab. However, as the Human Genome Project (HGP) ramped up, and technology advanced, the amount of data being generated by individual biologists and their related labs grew dramatically. This approach to data management was not scaleable. Within a lab it became increasingly difficult to find, access, and validate the data that was being collected, and sharing data between labs was difficult because each lab used their own format.

The current "data-scape" for bioinformatics has been made even more challenging by the research funding infrastructure and the rapid growth of the World Wide Web (WWW). Experimental biology still receives the bulk of the public research funds spent on biology and biotechnology. The bioinformatics groups that support these efforts are typically understaffed, under-funded, and nearly overwhelmed with the immediate task of managing research and production data as it is generated. When combined with funding requirements to disseminate results on the web, the result is what can best be described as a cottage-industry with hundreds to thousands of small, independent, heterogeneous data sources made available to the scientific community. Even though the data is publicly available, for all intents and purposes the difficulty of using it makes it inaccessible to the scientist. This situation must change. The need for a more structured approach to data management has lead to several ongoing research and development efforts in bioinformatics (e.g. [1] [2]).

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48. UCRL-JC-138042.

There is a growing expectation in the community that the relationship between experimental and computer-supported biological research will dramatically change. Traditionally, biological "wet lab" experiments have not only been the primary source of data, but they have been the focus of the science and the method by which the majority of interesting results have been generated. Modern, data-rich biology is moving to a new paradigm of discovery and verification, which is largely based on informatics-oriented capabilities such as new data integration and mining techniques. Data warehouses (including the web) will be mined to discover new correlations between biological components. New tools will help explore and characterize these clusters of related objects with the goal of isolating features or characteristics of interest. This will lead to the formulation of new hypotheses that need further testing. In some cases, it will be possible to (in)validate a hypothesis on the basis of data already available. In others, it will be necessary to schedule experimental wet lab work for that validation. In this scenario, the bulk of the science has shifted from the lab to the computer, and the role of bioinformatics has shifted from data collection and storage, to being the primary resource for scientific progress.

2. Bioinformatics challenges

The modern era of bioinformatics efforts at LLNL began over 10 years ago when a small team of computer scientists was hired into a division of approximately 60 biologists to help better manage their data. This team utilized relational database technology to implement a production data management system that recorded all of the experimental data being generated by the scientists. This team has grown over time and has successfully interacted with the genetics researchers for many years, learning more about the intricacies of the biological domain and earning the trust of the scientists. Among the more pervasive challenges that we have faced:

- ***Interacting with a growing number of users***. The bioinformatics team now supports more than 120 geneticists in 20 labs, not to mention hundreds of scientists from outside of LLNL, each needing a different type of access to distinct sets of data.
- ***Managing pull technologies in a push environment***. Though slowly improving, the job of convincing scientists to spend hard-won dollars on production-oriented bioinformatics tasks is difficult, and on research-oriented tasks is nearly impossible.
- ***Adapting to changing wet-lab technology***. As lab techniques change, the information that needs to be recorded changes as well. The database needs to provide access to both the old and the new data.
- ***Integrating new computer technology***. As new hardware and software becomes available, it needs to be utilized to its fullest advantage to address outstanding user requirements (e.g. web interfaces to data).
- ***Reflecting new views of the data***. As the biologists' understanding of the underlying science evolves, the data representation needs to be modified to conform to these new models.
- ***Developing new data acquisition techniques***. As humans become the bottleneck in the data acquisition process, better use of their time and energy is required (e.g. bar coding samples to improve tracking).
- ***Improving data reliability***. As more data is entered into the computer, the chance for errors increases correspondingly. Integrity constraints and data validation identify errors before they are propagated.

Instances of the scaling and cost-cutting challenges above seem to arise daily, and drive much of the focus of bioinformatics teams across the country. In the short run, quickly reacting to these needs allows the scientists to make the best use of their limited resources. Unfortunately, addressing these challenges has stripped away many of the resources needed to address three difficult, long-term issues:

- ***Difficulties in using dynamic, heterogeneous data***. Rapidly evolving domains incur oppressive data management demands. Applications that use this data must cope with *different* and *rapidly changing* data vocabularies, representations, models, interfaces, complexity, and levels of curation.
- ***Development of practical systems that provide effective access to external data***. Scientists have been increasingly distributing genomics data using the WWW. Local scientists need intuitive access to this important body of information.
- ***Allowing much richer interactions with data***. As the amount of available data grows, scientists want to combine more information in novel ways.

3. Bioinformatics research at LLNL

Bioinformatics research at LLNL began with application-specific research focused on reducing the cost of maintaining dynamic databases. More recently, the DataFoundry project has extended this focus to include data integration and access across multiple external, dynamic data sources.

Application-specific research

For nearly 20 years, LLNL has used a team of computer scientists from an applications division to support BioMedical research. This has enabled a long series of customer successes in physical mapping and genomic sequencing, using customized solutions that were freed from the constraint of needing to be publishable units in major journals [3].

Since 1995, when the Web explosion became undeniable, we have focused some attention on the problem of reducing the development and maintenance cost of bioinformatics systems. This was a necessity, given the explosion of data sources, huge annual increases in our data production volume, and only incremental increases in bioinformatics manpower. Since 1995 we have developed and evolved a meta-data system for automating many of the tasks (SQL creation, table-level backups, schema browsing, data browsing, automatic web form generation, etc.) associated with databases for genomic research endeavors. As an example of the cost-savings of such an approach, automatic web form generation from meta-data means that applications using that system require no code changes whenever the database endures schema evolution (which is nearly constantly in the environment of a leading-edge genomics research lab.) Prior to adopting these techniques we were maintaining over 60,000 lines of pre-web input scripts for a single project. Using these techniques has leveraged scarce programming talent to be able to handle many more projects, each with increasing volumes of data. The meta-data system is scheduled to be re-written in late 2000 and will subsequently be published.

DataFoundry

DataFoundry is an ongoing research project aimed at improving scientists' interaction with their data. DataFoundry began in October 1996 as an inter-disciplinary research effort between the Center for Applied Scientific Computing, the bioinformatics team, and a small group of structural biologists. DataFoundry's initial task was to develop an infrastructure that would allow the bioinformatics team to create and maintain a consistent view of several autonomous data sources.

To accomplish this, DataFoundry developed a meta-data based infrastructure to support a mediated data warehouse architecture. Data warehouses have been used in industry for several years and, as shown in Figure 1, are typically comprised of 5 layers. The data sources contain the data to be integrated into the warehouse through the wrappers (which parse the data) and the mediators (which translate the data into the warehouse representation). The warehouse itself is a large data repository, usually a relational database, presenting a consistent view of the data available from the sources. Users interact with the warehouse through a set of customized interfaces.

The challenge in creating a data warehouse for the genomics environment lies in developing an infrastructure flexible enough to handle the dynamic nature of the domain. Unlike commercial applications, scientific data sources are extremely dynamic (i.e. they change their data representations frequently). Whenever a source changes its data format, the wrapper and mediator must be updated to handle the new representation. For extreme changes, or when a new source is integrated, the warehouse and interfaces may also need to be updated. This makes it extremely challenging to keep a warehouse functional when a large number of dynamic sources are being integrated.

DataFoundry's meta-data infrastructure contains a mediator generator (MG) program that automatically generates a mediator using a collection of declarative meta-data (see [4] and [5] for a detailed description of the meta-data format, and [6] for a comparison between this approach and a more traditional approach). In addition, the MG defines a class library that can be used by the wrapper to represent data obtained from the

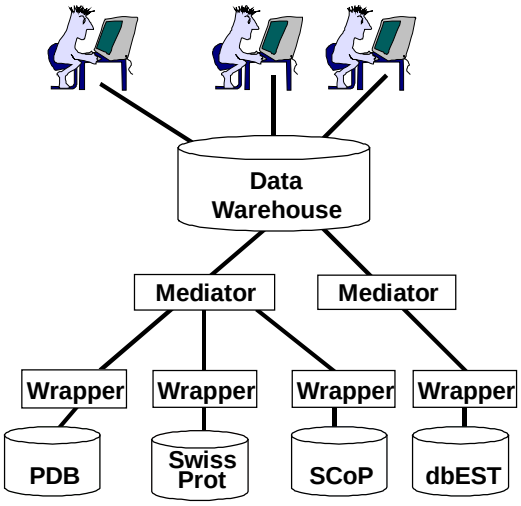

Figure 1 A Data Warehouse Architecture

source. This simplifies integrating new sources, since the administrator need only define the appropriate meta-data and write a wrapper using the resulting classes, instead of writing both the wrapper and the mediator from scratch. It also simplifies maintaining the warehouse, since the meta-data is significantly easier to update than the mediator. The MG has been used to integrate two new data sources (dbEST [7] and ScoP [8]) into an existing warehouse containing data from two other sources (PDB [9] and Swiss-Prot [10]). To interact with the data warehouse, the scientists use one of several interfaces depending on their query and level of expertise. Novices interact with the warehouse through either a forms based interface, which provides simple access for a small set of queries using html pages and cgi scripts, or a more functional applet/servlet based interface providing an intuitive graphical interface to the data. Intermediate users interact with the warehouse through a graphical query engine that allows them to form their own queries directly against the warehouse schema. Expert users directly query the warehouse either through an html form, or the programming language of their choice (perl, C/C++, etc.).

4. Future Work and Conclusions

The bioinformatics program at LLNL is an ongoing effort. The core bioinformatics team is focused on ensuring that the scientists they support are able to perform their research on a day-to-day basis, while the DataFoundry team attempts to address the long-term issues mentioned above. To date, DataFoundry has focused on providing access to fully integrated data from multiple sources. However, with the success of the WWW, more and more data sources are becoming available (more than 500 genomic sources were available at last count). It is impractical to expect all, or even most, of these sources to be integrated into a single, consistent view. Instead, DataFoundry is pursuing a hybrid strategy where critical data sources are fully integrated into the warehouse, but the interface is flexible enough to interact with additional sources at a primitive level. This will allow queries such as "find everything you can about this protein" to obtain significantly more information than is available from only the fully integrated sources. We believe that the capability to interact with as many data sources as possible, even at a very basic level, will become critical for geneticists to do their job as the human genome project enters its next phase.

References

1. Chen A, Markowitz V., "An Overview of the Object Protocol Model (OPM) and the OPM Data Management Tools". *Information Systems*, Vol. 20, No. 5, 1995.
2. Overton G C, Davidson S B, and Buneman P. "Database Transformations for Biological Applications". *DOE HGP Contractor-Grantee Workshop VI*, Santa Fe,NM, Nov 97.
3. Slezak T, Wagner M, and Yeh M. "A database system for constructing, integrating, and displaying physical maps of chromosome 19". Proceedings of the 28th Annual Hawaii International Conference

on Systems Sciences, L. Hunter and B. D. Shriver (eds.), IEEE Computer Society Press, Los Alamitos, CA, 5, pp.14-23,1995

4. Critchlow T, Ganesh M, and Musick R. “Automatic Generation of Warehouse Mediators Using an Ontology Engine”. Proceedings of the Fifth International Workshop on Knowledge Representation Meets Databases (KDRB'98) . Seattle. WA. May 1998.
5. Critchlow T, Ganesh M, Musick R. “Meta-Data Based Mediator Generation”. Proceedings of the Third IFCIS Conference on Cooperative Information Systems (CoopIS'98) New York, NY. August 1998.
6. Critchlow T, Musick R, and Slezak T. “Experiences Using a Meta-Data Based Integration Infrastructure”. Second International Workshop on Biomolecular Informatics. Feb 27 - Mar 3, 2000.
7. Boguski A M S, Lowe T M, and Tolstoshev C.M. “dbEST – database for expressed sequence tags”. *Nat. Genet*. 4(4):332-3. Aug 1993.
8. Murzin G, Brenner S E, Hubbard T, and Chothia C. “Scop: a structural classification of proteins database for the investigation of sequences and structures.” *Journal of Molecular Biology*. 247:536-540. 1995.
9. Callaway J, Cummings M, Deroski B, Esposito P, Forman A, Langdon P, Libeson M, McCarthy J, Sikora J, Xue D, Abola E, Bernstein F, Manning N, and Sussman J. “Protein Data Bank Contents Guide: Atomic Coordinate Entry Format Description Version 2.1”. Tech Report available from www.pdb.bnl.gov. October. 1996.
10. Bairoch A, Apweiler R. “The SWISS-PROT protein sequence database and its new supplement TrEMBL”. *Nucleic Acids Res.* 24:21-25(1996)