Going Back to our Database Roots for Managing Genomic Data

Joachim Hammer and Markus Schneider

Dept. of Computer & Information Science & Engineering
University of Florida
Gainesville, FL 32611-6120
{jhammer,mschneid}@cise.ufl.edu

1. Introduction

In the past decade, the rapid progress of genome projects has led to a revolution in the life sciences causing a large and exponentially increasing accumulation of information in molecular biology and an emergence of new and challenging applications. The flood of genomic data, their high variety and heterogeneity, their semi-structured nature as well as the increasing complexity of biological applications and methods mean that many and very important challenges in biology are now challenges in computing and here especially in databases. This statement is underpinned by the fact that millions of nucleic acid sequences with billions of bases have been deposited in the well-known persistent *genomic repositories* such as EMBL, GenBank, and DDBJ, etc. In addition, hundreds of specialized repositories have been derived from the above primary sequence repositories. Information from them can only be retrieved by computational means. Biologists are overwhelmed by this continuously growing data which is awaiting further refinement and analysis.

2. Data Management Problems in Bioinformatics

Discussions and cooperation with biologists have revealed the following main problems regarding the information available in the genomics repositories:

- The number and size of available data sources is continuously growing. As a result, there is much overlap and conflicting information and a proliferation of interfaces and portals.
- There is a constant worry that familiar sources sometimes disappear or get merged and that data records become obsolete.
- There is apprehension that essential information will be overlooked.
- There is little or no agreement on terminology.
- Query results are unmanageable unless organized into a customized, project-specific database.
- Database search functions are limited by the interface.
- Scientists are forced to understand low-level data management. For example, they are often required to learn and write SQL or code in some other programming language (e.g., Perl)

- Noisy data. For example, it is estimated that 30-60% of sequences in GenBank are erroneous.

In response to these *biology-centric problems*, the approach outlined in this position paper aims at overcoming the following fundamental *computer science challenges*: The deliberate independence, heterogeneity, and limited interoperability among multiple genomic repositories, the enforced low-level treatment of biological data imposed by the genomic repositories, the lack of expressiveness and limited functionality of current query languages and proprietary user interfaces, the different formats and the lack of structure of biological data representations, and the inability to incorporate one's own, self-generated data.

In comparison, current research is focused predominantly on integrating the heterogeneous and largely semi-structured repositories using federated or query-driven approaches. In addition, most of the analysis is performed outside of the repositories (e.g., sequence similarity search, visualization tools). Current efforts has resulted in complex middleware tiers between end-users and the data servers; these middleware solutions are both inefficient and require much involvement and input from the user (i.e., the human is the query processor).

3. Proposed Solution

In response to the challenges and problems described above, we advocate the increased and integrative employment of current database technology as well as appropriate innovations for the treatment of non-standard data to cope with the large amounts of genomic data. In a sense, we advocate a *"back to the roots"* strategy of database technology for bioinformatics. This means that general database functionality should remain inside the DBMS and not be shifted into the middleware.

Our integrating approach, which to our knowledge is new in bioinformatics and differs substantially from the integration approaches that can be found in the literature, rests on two fundamental pillars:

1. *Genomics Algebra*. This *extensible* algebra is based on the conceptual design, implementation, and database integration of a new, formal data model, query language, and software tool for representing, storing, retrieving, querying, and manipulating genomic information. It provides a set of high-level *genomic data types* (*GDT*s) (e.g., genome, gene, chromosome, protein, nucleotide) together with a comprehensive collection of appropriate *genomic operations* or *functions* (e.g., translate, transcribe, decode). Thus, it can be considered a resource for biological computation.
2. *Unifying Database*. Based on latest database technology, the construction of a unifying and integrating database allows us to manage the semi-structured or, in the best case, structured contents of genomic repositories and to transfer these data into high-level, structured, and object-based GDT values. These values then serve as arguments of Genomics Algebra operations. In its most advanced extension, the Unifying Database will develop into a global database comprising the most important

or, as a currently rather unrealistic vision, even all publicly available genomic repositories.

For a more detailed description of our approach, the reader is referred to [1].

4. Expected Impact

We believe our approach will cause a fundamental change in the way biologists analyze genomic data. No longer will biologists be forced to interact with hundreds of independent data repositories each with their own interface. Instead, biologists will work with a unified database through a single user interface specifically designed for biologists. Our high-level Genomics Algebra allows biologists to pose questions using biological terms, not SQL statements. Managing user data will also become much simpler for biologists, since his/her data can also be stored in the Unifying Database and no longer will s/he have to prepare a custom database for each data collection. Biologists should, and indeed want to invest their time being biologists, not computer scientists.

In addition, we believe the Genomics Algebra approach empowers biologists to perform complex analyses on large-scale collections of data without needing a computer scientist at their side. This is especially beneficial to biologists who work alone or in a small group since it "levels the playing field" permitting any biologist with a good idea to pursue it fully and not be discouraged by the lack of local infrastructure and support personnel. Finally, our approach fosters collaborations among biologists by providing both a convenient means for them to share data, and a powerful suite of functions to analyze their data. With our approach, users would choose to make their data accessible to other users (and specify which user(s) (if any) receive the privilege). Accessing custom user data in our unified database, and analyzing it, is indistinguishable from performing those tasks on the available public data. Currently, to use another biologist's data, especially a large quantity of data that's stored in a custom-designed database, requires one to link to the machine containing the database and write custom functions to perform the analyses.

References

[1] J. Hammer and M. Schneider, "Genomics Algebra: A New, Integrating Data Model, Language, and Tool for Processing and Querying Genomic Information." In *Proceedings of the First Biennial Conference on Innovative Data Systems Research*, Morgan Kaufman Publishers, Asilomar, CA, pages 176-187, January 5-8 2003.