Scientific databases

Biological data management

The term paper within the framework of the course
"Principles of Modern Database Systems"

by Aleksejs Kontijevskis

PhD student

The Linnaeus Centre for Bioinformatics and Dept. of Pharmaceutical Biosciences

Uppsala University

April, 2007

1. Introduction

The biological sciences encompass an enormous variety of information, from the environmental sciences, which give us a view of how species live and interact in a world filled with natural phenomena to cell biology, which provide knowledge about the inner structure and function of the cell and beyond. All this information requires classification, organization and management. Biological data exhibits many special characteristics that make management of biological information a particularly challenging problem. A multidisciplinary field called bioinformatics has emerged recently to address information management of genetic information with special emphasis on DNA and protein sequence analysis. However, bioinformatics harness all other types of biological information and the modeling, storage, retrieval, and management of that information. Moreover, applications of bioinformatics span new drug target validation and development of novel drugs, study of mutations and related diseases, anthropological investigations on migration patterns of tribes and therapeutic treatments.

2. Specific features of biological data

1. Biological data is highly complex when compared with most other domains or applications. Definitions of such data must thus be able to represent a complex substructure of data as well as relationships and to ensure that information is not lost during biological data modeling. Biological information systems must be able to represent any level of complexity in any data schema, relationship, or schema substructure. A good example of such system is MITOMAP database documenting the human mitochondrial genome (http://www.mitomap.org). The database information include data and their relationship about ca. 17,000 nucleotide bases of the mitochondrial DNA; 52 gene loci encoding mRNAs, rRNAs and tRNAs; over 1,500 known population variants and over 60 disease associations. MITOMAP includes links to over 3,000 literature references. Traditional RDBMS or ODBMS are unable to capture all aspects of the database data.

2. The amount and range of variability in biological data is high. Therefore, the systems handling biological data should be flexible in data types and values. Constraints on data types and values must be put with care, since the unexpected values (e.g. outliers), which are not uncommon in biological data could be excluded resulting in the lost of information.

3. Schemas in biological databases change rapidly. This requires improved information flow between various database releases, as well as schema evolution and data object migration support. In most relational and object database systems the ability to extend the schema is not supported. What happens now is that many biological/bioinformatics databases (such as GenBank, for example) release the entire database with new schemas once or twice a year rather than incrementally change the system as a change is needed.

4. Representations of the same data by different biologists will likely be different (even using the same system). Thus, it is necessary to have mechanisms, which could align different biological schemas.

5. Most users of biological data need read-only access only, whereas write access to the database is not required. Usually curators of the databases are the ones who need write access privileges. The vast majority of users generate a wide variety of read-access patterns into the database, but these patterns are not the same as those seen in traditional relational databases. User requested searches demand indexing of often unexpected combinations of data classes.

6. Most biologists don't have knowledge of the internal structure of the database or about its schema design. Biological database interfaces should display information to users in a manner that is applicable to the problem they are trying to address and that reflects the underlying data structure in an easily understandable manner. Biologists usually know what data they require, but they have no technical knowledge of the data structure or how a DBMS represents the data. Relational database schemas fail to provide intuitive information to the user regarding the meaning of their schema. Web interfaces, on the other hand, often provide preset search interfaces, which may limit access into the database.

7. The context of data gives added meaning for its use in biological applications. Therefore it is important that the context is maintained and conveyed to the user when appropriate. It is also advantageous to integrate as many context as possible to maximize the interpretation of the biological data. For instance, the sequence of DNA is not very useful without information describing its organization, function, etc.

8. Defining and representing complex queries is extremely important to the biologist. Hence, biological systems must support complex queries and provide tools for building such queries.

9. Users of biological information often require access to “old” values of the data – particularly when verifying previously reported results. Therefore, the changes in values must be supported through the archives to enable researchers to reconstruct previous work and reevaluate prior and current information.

All these specific characteristics of the biological data point to the fact that traditional DBMS’s do not fully satisfy the requirements put on complex biological data.

3. Existing Biological Databases

It has been estimated that are over major 1,000 public and commercial biological databases currently available to scientific community (by the end of 2006). These biological databases usually contain genomic and/or proteomics data. Some databases are also used in taxonomy. As already has been mentioned, the biological databases incorporate enormous ammount of various types of biological data including (but certainly not limited to) nucleotide sequences of genes, amino acids sequences of proteins, information about their function, structure, localization on chromosome, clinical effects of mutations, protein-ligand, gene-ligand interactions, as well as similarities of biological sequences can be found and so on. By far the most important resource for biological databases is a special yearly January issue of the journal Nucleic Acids Research. This issue categorizes all the publicly available online databases related to bioinformatics.

The most important biological databases can be roughly classified into the following groups:

- *Primary sequence databases* (include International Nucleotide Sequence Database (INSD) consisting of DDBJ [DNA Data Bank of Japan], EMBL Nucleotide DB [European Molecular Biology Laboratory] and GenBank [National Center for Biotechnology Information]). These databanks represent the current knowledge about the sequences of all organisms. They interchange the stored information and are the source for many other databases.

- *Meta-databases* (include MetaDB containing links and descriptions for over 1200 biological databases, Entrez [National Center for Biotechnology Information], euGenes [Indiana University], GeneCards [Weizmann Institute], SOURCE [Stanford University], Harvester [EMBL Heidelberg] and other). These meta-database can be considered a database of databases, rather than one integration project or technology. They collect information from different other sources and usually make them available in new and more convenient form.

- *Genome browsers* (e.g. Integrated Microbial Genomes system, The Ensembl Genome Browser [Sanger Institute and European Bioinformatics Institute] and many other). Genome Browsers enable researchers to visualize and browse entire genomes of organisms with annotated data including gene prediction and structure, proteins, expression, regulation, variation, comparative analysis, etc. Annotated data is coming usually from multiple diverse sources.

- *Specialized databases* (Human Genome Organization database, SHMPD The Singapore Human Mutation and Polymorphism Database and many other databases).

- *Pathway databases* (e.g. BioCyc Database, Reactome and other)

- *Protein sequence databases* (UniProt [UniProt Consortium: EBI, Expasy, PIR], Swiss-Prot Protein Knowledgebase [Swiss Institute of Bioinformatics] and many other)

- *Protein structure databases* (Protein Data Bank, CATH Protein Structure Classification, SCOP Structural Classification of Proteins etc.)

- *Microarray databases* (ArrayExpress [EBI], SMD [Stanford University], etc.)

- *Protein-Protein Interactions* (BioGRID [Samuel Lunenfeld Research Institute], STRING [EMBL])

4. The Key Biological Databases

4.1. GenBank

The world top DNA sequence database today is GenBank, maintained by the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM) (http://www.ncbi.nlm.nih.gov). Established in 1978 as a repository for DNA sequence data, it is now expanded to include sequence tag data, protein sequence data, three-dimensional protein structure, taxonomy, and links to the medical literature (MEDLINE). The database size in flat file format is over 100 GB uncompressed and has been doubling every 15 months. GenBank contains over 31 billion nucleotide bases of more than 24 million sequences from over 100,000 species with roughly 1,400 new organisms being added each month. Through international collaboration with the European Molecular Biology Laboratory (EMBL) in the U.K. and the DNA Data Bank of Japan (DDBJ) the data is exchanged on daily basis. The system is maintained as a combination of flat files, relational databases, and files containing Abstract Syntax Notation One (ASN.1). Although users of the database are not able to access the structure of the data directly for querying or other functions, complete snapshots of the database are available for export in a number of formats, including ASN.1. The query mechanism provided is via the Entrez application (or its web version), which allows keyword, sequence, and GenBank identifiers search through a static interface.

4.2. The Genome Database

The GDB database (http://www.gdb.org), created in 1989, is a catalog of human gene mapping data, a process that associates a piece of information with a particular location on the human genome. The GDB system is built around Sybase, a commercial relational DBMS, and its data are modeled using standard Entity-Relationship techniques. GDB distributes a Database Access Toolkit. As with GenBank, users are given only a very high-level view of the data at the time of searching and thus cannot make use of any knowledge gleaned from the structure of the GDB tables. Search methods are most useful when users are simply looking for an index into map or probe data. Exploratory ad hoc searching is not encouraged by present interfaces. Integration of the database structures of GDB and OMIM was never fully established.

4.3. Gene Ontology

Gene Ontology (GO) Consortium was formed in 1998 as a collaboration among three model organism databases: FlyBase, Mouse Genome Informatics (MGI) and Saccharomyces or yeast Genome Database (SGD). The goal is to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism.

The GO Consortium has developed three ontologies: molecular function, biological process, and cellular component, to describe attributes of genes, gene products, or gene product groups (http://www.geneontology.org/). Each ontology comprises a set of well-defined vocabularies of terms and relationships. The terms are organized in the form of directed acyclic graphs, in which a term node may have multiple parents and multiple children. A child term can be an instance or part of its parent. Latest release of GO database has over 13,000 terms and more than 18,000 relationships between terms. GO was implemented using MySQL, an open source relational database and a monthly database release is available in SQL and XML formats.

4.4. Swiss-Prot

The most important source of information on protein sequences are the Swiss-Prot + TrEMBL protein sequence databases (http://www.expasy.ch/sprot/). The Swiss-Prot protein knowledgebase is an annotated protein sequence database, maintained collaboratively by the Swiss Institute of Bioinformatics and the EBI. It strives to provide sequences from all species, combined with a high level of manual annotation, a minimal level of redundancy and a high level of integration with other biomolecular databases. To make new protein sequences available for the public as quickly as possible without relaxing the high annotation standards of Swiss-Prot, the EBI provides a complement to Swiss-Prot known as TrEMBL. TrEMBL consists of computer-annotated entries derived from the translation of all coding sequences in the DDBJ/EMBL/GenBank Nucleotide Sequence Database.

5. Database bioinformatics tools

While database systems provide facilities to manage large data volumes, many database systems only partially provide support for the numeric computations required to perform statistical assessment of scientific data and therefore require further development. This shortcoming limits the use of database systems by scientific users. The integration of numerical algebraic calculations enables to perform automatic optimization of entire computations, with the resulting benefits of query optimization, algorithm selection and data independence becoming available to computations on scientific databases. This removes the barrier between the database system and the computation allowing the database optimizer to manipulate a larger portion of the application.

Algebraic Optimization of Computations.

A pioneering work of Wolniewicz and Graefe extends the concept of database query and show on a case study how numeric computation over time series data can be implemented effectively in scientific databases. This frees the user from concerns about ordering of computation steps, algorithm selection, use of indices and other physical properties such as data distribution. The authors developed a scientific optimizer using the

“Volcano” optimizer generator, which could perform logical transformations and physical algorithm selection.

In the optimization of scientific computations, the identification of suitable transformation rules is of central importance. Once applicable transformation rules have been found and applied to generate equivalent logical expressions, the optimizer must find a set of physical algorithms that can implement or execute each expression. For instance, a join operator can be implemented as either a merge- or hash-based algorithm, while an interpolation can be implemented by any of variety of curve fitting algorithms. Other query optimizer issues include limiting the search space, detecting common sub-expressions and improving cost estimation.

Time series can be viewed as sets there each record is assumed to be tagged with a time value. Start and stop time, the average and maximum difference between samples, and whether or not the time differences are constant between adjacent records are all important for some operations. Spectra are treated in a manner very similar to time series, but with a frequency attribute attached to each record rather than a time value. The operators supported within the system are divided into two groups, logical and physical. The user’s original computation is composed of logical operators, while the final execution plan is expressed to the database in terms of physical operators. The authors demonstrate that besides standard relational operators as “select”, “project” and “join” other necessary operators can be included such as “a random sampling operator”, “digital filtering procedure”, “recomputation of the value of a single record based upon an averaging function applied to nearby records”, “interpolations” and “extrapolation” operators, operator for “merging two time series”, as well as “spectral filter” operator. The authors also include simple math function applications such as “correlation”, “convolution” and “deconvolution” of spectra. Additionally, physical operators implemented as iterators are included such as “pass a window over a sorted data set”, “fast Fourier Transform” and some other. It should be noted that some operators (e.g. Fourier transform of the time series and spectra) are “expensive” operators to perform and, therefore, the decision on when to move between normal and Fourier-space is important for optimization process. Logical transformations for scientific operators are also vital to the application of optimization. There are a number of transformations in specific scientific domains that are considered valid and thus needs to be implemented. Moreover, the user should be able to enable or disable certain transformations to control the accuracy of the results of the equation. The authors conclude that it is very beneficial to remove the barrier between the database system and scientific computations over the database. Thus, automatic optimization of integrated algebras is a crucial step in supporting scientific computations over database systems.

Sequence Retrieval System (SRS)

The high amount and diversity of data available in the large number of different databases to the scientific community is to the user's advantage but it also creates the problem of knowing exactly where to get specific information. Another problem is that different databases with different contents also use different formats adding further technical difficulties to the already complex task of accessing and exploiting the data. Even though many databases have developed linking systems to relate to data stored elsewhere, these links are difficult to use due to the differences between individual database implementations. With the expansion of databases, quick access to their contents has become an important issue. Different technical solutions are used or are under development at the major bioinformatics sites. The major contribution to these field was development of Sequence Retrieval System (SRS) at EBI. SRS has addressed many of the difficulties in database access, integration of databases and analysis tools. SRS is an integration system for both data retrieval and data analysis applications. It provides, under a uniform and simple to use interface, a high degree of linking between the databases in the system. This allows the user to perform simple and complex queries across different databases. Its original way of linking makes it possible to query across databases that even do not contain direct cross-references to each other. The user even is able to ask questions like "Give me all the proteins that share InterPro domains with my protein" or "Give me all the known 3D structures of this set of proteins". The SRS server at EBI (http://srs.ebi.ac.uk) contains more than 140 biological databases (including sequence and sequence related databases, bibliography, metabolic pathways, 3D structure and many other databases) and integrates many analysis tools. Results of such analyses are themselves indexed as SRS databases and can thus be linked to others by using pre-defined or user-defined SRS views.

Overall, the EBI and other major world bioinformatics centers provide a whole range of analysis tools for databases. There are currently more than 60 distinct services available at the EBI such as a range of sequence homology and similarity algorithms like (FASTA, BLAST and Smith-Waterman), sequence analysis tools (many European

Molecular Biology Open Software Suite applications), gene and structural prediction methods.

6. Scientific workflows

A scientific workflow is an executable description of a scientific process, which records each inline process required to take input data and produce a meaningful output product. Scientific workflows operate on large, complex, and heterogeneous data, are computationally intensive, and produce complex data products. Furthermore, scientific workflows are generally data-flow driven. Scientific workflow systems have traditionally been stand-alone applications designed for a specific domain for creating and executing scientific workflows, e.g. Kepler. Kepler provides scientists with a number of benefits, such as integrated environment in which scientists can design, communicate, and execute their analytical processes, data query, retrieval, and archiving tools. Kepler allows workflows to serve as a form of metadata and provide an opportunity for workflows and data to be reused to form novel and extended analyses. Instead of creating complex interfaces and tools for each domain, it is much more advantageous to provide domain specific customization.

In Kepler scientific workflows bring together data and services created by different groups and various individuals. The workflow applications written in Kepler encompass a wide variety of scientific domains, sub-domains, and specialties. Within Kepler, scientific workflows are authored in a graphical, drag-and-drop manner. Services contain typed ports that can be connected to other services or data sources. Ports can have simple atomic types such as “integer” and “string” as well as more complex structures, including arbitrarily nested array and record types. As a workflow is executing, data passes between ports via tokens that can be readily manipulated to meet the differing syntactic needs of other services. Data produced by a scientific workflow can be displayed graphically at run time, or written to disk for later use.

The main difficulties with scientific workflows are:

- The support of high-level conceptual models. It is necessary effectively capture the high-level aspects of a workflow, meanwhile preserving (but often hiding) the underlying technical details.

- Basic contextual metadata. A general lack of contextual metadata with respect to data and services is problematic for users (for example, for those who are trying to find new and relevant datasets and services).
- Schema and service-type semantics. Scientific data integration can be a complex and time-consuming process. Scientific data is highly heterogeneous, laden with structural, schematic, and semantic differences.

Kepler has adopted the OWL web ontology language as the primary language for capturing domain-specific terminologies used in semantic metadata. Kepler is designed to provide users with the ability to semantically register their dataset schemas and services using semantic annotations, which define a relationship between a service or dataset and terms in ontology. Kepler also leverages semantic annotations to provide customizable access to datasets and services, allows users to select and configure the classification ontology, view the hierarchically arranged ontology and see services classified according to the concept hierarchy. All these functionalities enable users to customize easily Kepler service presentation and provide ontology-based browsing of data and services. In addition, Kepler statically checks whether two connected services are "semantically compatible" and notifies the user when a connection is not considered semantically well typed. As large repositories of workflow components become available, finding relevant resources becomes more difficult. However, given a workflow service on the Kepler canvas, a user can search for all "semantically compatible" resources. Berkley et al. demonstrate the extensions to Kepler, including the advances in automating data-integration and service-composition tasks, the use of semantic annotations to verify that workflows are semantically meaningful, and the ability to search for contextually relevant components and data sets in situ.

Bibliography

1. R. Elmasri, S. B. Navathe. Emerging Database Technologies and Applications. p. 1022-1034. In: *Fundamentals of database systems. Addison Wesley, 5th edition*, 2006.

2. R. Wolniewicz, G. Graefe. Algebraic Optimization of Computations over Scientific Databases. *VLDB* 1999.

3. R. Apweiler, M. J. Martin, C. O'Donovan, M. Pruess. Managing core resources for genomics and proteomics. *Pharmacogenomics*. pp. 343-350, 2003.

4. C. Berkley, S. Bowers, M. B. Jones, B. Ludäscher, M. Schildhauer, J. Tao. Incorporating Semantics in Scientific Workflow Authoring. *SSDBM* 2005.

5. A. M. Lesk. Introduction to Bioinmformatics. *Oxford University Press*, 2002.

6. Various WWW resources.