OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Challenges in Data Management for Functional Genomics

MICHAEL GRIBSKOV

ABSTRACT

Biological databases face challenges in four main areas: (1) integration, interoperation and federation; (2) ontologies and definitions of semantics; (3) community annotation; and (4) integration of data analysis tools with databases. Each of these areas provides interesting targets for research and development.

INTRODUCTION

FUNCTIONAL GENOMICS can be thought of as the second stage of any genome sequencing project. Once the genomic sequence has been determined, the task of functional genomics is to use a combination of computational and experimental approaches to determine the function of the genes that comprise the genome. This requires a complex mixture of linking between databases, acquiring information from the biological literature, integration of genome-wide experimental results, and development of user-friendly interfaces and analytical tools.

A functional genomics database, for example the PlantsP (http://plantsp.sdsc.edu) or PlantsT (http://plantst.sdsc.edu) database, uses the genomic sequence as a scaffold on which other information is organized. This other information includes gene models, experimentally determined mRNA sequences, description of mutants and their phenotypes, gene expression experiments, protein–protein interaction experiments, various kinds of tagging experiments, and potentially many others. This integration creates problems on many levels.

INTEGRATION/INTEROPERATION/FEDERATION

Significant portions of the data incorporated in such a database represent information that is copied from other sources. Generally these sources are sequence databases such as Genbank, EMBL, Swiss-Prot, or PDB. This information often would not need to be copied if efficient, lightweight protocols and sufficiently powerful servers allowed this information to be queried *in situ* at the source. A sufficiently efficient protocol would also allow queries to be broken down and distributed across the relevant databases. This would create an ideal situation where queries always refer to the original source—presumably the most up-to-date version of the data.

Whether or not the query process can be made to run in real time, additional complexity arises due to the lack of standardized formatting for data and services. Although standards have existed for many years

San Diego Supercomputer Center and Department of Biology, University of California, San Diego, La Jolla, California.

for certain database formats, many databases remain incompletely specified and difficult to parse. The most rigorous specifications, mmCIF for proteins structural data and ASN.1 for sequence data have not been widely used, due possibly to their complexity (or possibly to ongoing changes in the NCBI ASN.1 specification). For many applications, simpler definitions of sharable objects such as sequences, structures, and services (such as BLAST-type sequence comparisons) would be more useful. Implementation of such systems, for example, DAS and MOBY, are underway.

ONTOLOGIES/LIMITED VOCABULARIES

The use of ontologies and limited vocabularies across many databases is an invaluable aid to semantic integration. Interest in this area has increased greatly over the last several years, but much work remains to be done. In the long run, it seems that no single hierarchy will be expressive enough to reflect the differences in viewpoint of geneticists, biochemists, crystallographers, and molecular biologists (a partial list). It seems to me that ultimately there must be many ontologies and that some framework of equivalences between individual terms in individual ontologies will have to be defined. In the short-term therefore semantic issues may get simpler, but in the longer term they will again become complex.

Another issue with ontologies concerns their complexity. There is a tension between groups who want and need a very detailed ontology, and those who need a less complex one. For many of our databases, a very deep and complex ontology is nearly as bad as none at all because annotators do not have the patience to search for the precisely correct term and will therefore use an incorrect one. The data is then well annotated in a standard form but nevertheless incorrect. Better methods of dynamically adjusting the level of detail are needed to address this problem.

COMMUNITY ANNOTATION/TEXT ANALYSIS

Electronic resources are the only way that we can deal with the depth of complexity represented by living cells. The number of molecules and the experimental data describing them are simply too large for the unassisted human brain. In the long run, it seems essential that the information in the database must derive from the scientific experts of the field rather than from a group of non-expert database developers. Significant social barriers make such an obvious step problematic; experts do not typically receive citation credit for participating in database annotation, nor is database curation/annotation generally considered in tenure review or other important social decisions. A number of resources are developing community-based annotation processes, although so far success has been limited. A related issue is the question of peer review. The scientific literature is enhanced by the peer-review process which filters out less significant or incorrect conclusions (to a limited extent). Most databases are completely lacking in a review process that validates their annotation—not surprisingly these annotations are often of low quality.

The alternative to expert community-based annotation is to have a database core provide curated information from the scientific literature. This is time-consuming and requires highly educated and trained annotators. As the social impediments to expert curation are likely to continue for the foreseeable future, there is a dire need for improved methods for computer assisted annotation based on the scientific literature. Classical text-retrieval and analysis approaches have had some success but much better systems are needed.

DATA ANALYSIS

Functional genomics integrates textual information with large experimental datasets, a few examples are given below. Microarray gene expression data and protein–protein interaction data are cases in point. It is often the case that these data require special analytical facilities to allow the user to use them to full advantage. In the case of microarray data this may mean providing a variety of normalization and clustering procedures. Protein–protein interaction data represents a network of interactions and therefore are not easily represented or queried using a relational database. Cases where images are important would benefit from

application of automated analysis at varying levels. The similarity is that in each case complex analytical functions must be integrated with a relational database, and that often, a user workspace and journaling capability is logically required. The complexity is increased by the fact that experimental approaches and techniques are under constant development leading to ongoing changes in the data and analytical procedures.

DISCUSSION

Construction of functional genomics databases is highly challenging; the data is complex, dispersed over worldwide sites, and subject to continual change. The social context, while improving, makes long-term funding and short-term cooperation of domain experts problematic. However, these resources are critical elements in the development of the scientific infrastructure needed to completely understand living organisms. This is increasingly so as the focus of biology shifts from reductionist analysis of parts to description and modeling of complex systems.

Address reprint requests to:
Dr. Michael Gribskov
San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0537

E-mail: gribskov@sdsc.edu