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

Complexities of Managing Biomedical Information

RUSS B. ALTMAN

THE PROBLEMS WE FACE with representing and manipulating biological knowledge are not unique to biology, but biology is particularly difficult because it has a confluence of many technical challenges that often appear isolated in other domains. One key point is that we are rarely talking about something being impossible in a relational model, but instead are often referring to the relative difficulty (relative to ad hoc solutions, object oriented solutions, text-based solutions, and others). The key features of biological data that make them difficult to represent and manipulate efficiently include the following.

DUAL NATURE OF SEQUENCE INFORMATION

We care about sequences as entities—genes, exons, transcripts—but we also care about individual elements of sequences, such as particular bases (for SNPs) or amino acids (for protein mutations, functional analyses). It is very difficult to find efficient representations that allow for this dual requirement of accessing entire blocks while also wanting to "explode" them. Furthermore, there are ways in which we parse the blocks into pre-splice, post-splice mRNA where we need to talk about subsegments.

IMPORTANCE OF HIERARCHICAL DATA REPRESENTATIONS

Biological structure is organized hierarchically from atoms up to organisms (and even ecosystems). There is therefore both an elaborate "part-of" reality to biology that is constantly re-invented, as well as a fundamental "is-a" reality in terms of classification (especially with regards to genes and species). Rapid reasoning up and down these two intermingled trees is fundamental to many biological analyses, and yet is sometimes cumbersome to support in existing systems.

GREAT DATA TYPE VARIETY, ESPECIALLY FOR PHENOTYPE DATA

As we enter the post-genomic period, it becomes clear that a major challenge is relating genotype data to phenotype data. With few exceptions, most phenotype data is collected in non-standard ways and is represented in databases in ad hoc manner. This great heterogeneity is a major challenge to the post-genomic computational analysis of data, because it is much more difficult to aggregate data from different groups, and because there is an infinity of experimental procedures that can be used to collected related, but non-identical data. The great heterogeneity of techniques for collecting data is a relative strength of biology, because models can be tested in a virtual infinity of manners, but it is a major representational challenge to the computational community, because we can not easily identify representational schemes that both capture the information in sufficient detail, and also allow aggregation and analysis with general purpose algorithms.

Departments of Genetics & Medicine, Stanford University School of Medicine, Stanford, California.

DISTRIBUTED NATURE OF BIOLOGICAL DATA

It is generally agreed that the best databases in biology are those that are carefully curated by specialists with domain expertise. A consequence is that most databases are very limited in their scope, with high quality primary data and a drop-off in ancillary data as you move away from the primary data. This is often addressed as a set of URL links between resources, which is fine for biologist humans, but less useful for computational engines. Thus, the assumption of relatively narrow data collections in a distributed environment needs to be supported in a more robust manner.

TENSION BETWEEN HUMAN AND COMPUTER CLIENTS

There is an unfortunate tension between computational representations and formats that are human-usable and those that are set up well for computational analysis. Data locality is very important for human processing (e.g., "what's all the available information about gene X?") but often antagonizes computational goals such as normal form and efficient query. In addition, complex computational representations that enable powerful algorithms often require "views" of the data that are not intuitive to biologists. Sometime these views can be deconvoluted in order to present information to human users, but this is typically a difficult and expensive activity. The dual requirement of "easy to understand" and "powerful for computation" is difficult to meet with current systems, and should only become more difficult as our understanding of biological complexity increases. Biology is too complicated to model graphically as ellipsoids and their qualitative interactions. These issues lead to three observations about current biological information systems:

1. Lots of Redundant Effort Spent Creating Middleware

Almost every project I am aware of has a huge effort to create "biology-friendly" middleware on top of either a relational database (recently) or on top of an elaborate file system of text files (in the past). This should be of concern to funding agencies, since precious research dollars are being used in many cases to invent the same kinds of interfaces. Of course, each is slightly different in order to accommodate either the biological situation or the biases/interests of the architects.

2. Poor Web Performance Without Extra Engineering

Another distressing trend is that towards optimization of web resources in order to support real-time queries on the web or by computational engines. Many investigators are surprised to learn that relatively reasonable relational database implementations of information resources can be difficult to tune for real-time query performance. This performance is really a result of all the complicating issues mentioned above. However, it also leads to substantial amounts of time being spent worrying about how to deliver information that is perfectly well represented and available, but just takes too long to retrieve with current DB + middle layer models. This leads to strategies for compiling content down for performance which can make maintenance issues more difficult.

3. Inadequate Query and User Interface

At tension with performance and need for expressive power is the need to reduce these data into forms that can be queried and understood by users. The query model for any information system will reflect its underlying organization (as well as the organization of any middle layer). The user interaction, can therefore be as complicated and intricate as the underying information system, and this may be a major barrier to acceptance. Thus, a major challenge to building information systems is the creation of robust interfaces

that supply access to the power and richness, but which also can be used by opinion-leader biologists who are impatient to "just get the answer." There are currently few general purpose tools for creating such environments, and thus much effort and funding is spent on special purpose, domain-specific interfaces that are expensive and single-use.

TWO OTHER CONSIDERATIONS

Finally, there are two other issues that complicate the management of biomedical data that should be recognized. First, the most trusted biological databases are manually curated by practitioners who lovingly curate and add their best judgment to the annotation of the data. These curators produce databases that are more trusted than un-curated databases, but also introduce challenges. They make understandable human mistakes and they also do not rigidly adhere to formats and semantics, in an effort to capture subtleties that are real, but which complicate the utility of databases for automated analysis. Thus, methods for helping these curators do their work while also maintaining formal correctness of the data are urgently needed. Second, there are significant privacy concerns about biomedical data, and the risks of re-identification of human subjects who allow their genetic and phenotypic information to be distributed for research purposes. The emergence of the HIPAA regulations on the exchange of protected health information is a major social and technical barrier to science that must be addressed directly in order to both facilitate scientific progress and protect the confidentiality and safety of human subjects.

Address reprint requests to:
Dr. Russ B. Altman
Departments of Genetics & Medicine
Stanford University School of Medicine
251 Campus Drive
MSOB X-215
Stanford, CA 94305-5479

E-mail: russ.altman@stanford.edu