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

Data Management Systems: Science versus Technology?

MICHAEL N. LIEBMAN

GAPS EXIST AND CONTINUE TO BE PROPAGATED between most data management systems in the biological domain, their potential users and the ultimate goal of extracting critical knowledge in support of research in molecular and cell biology. This is, in part, because the focus of these systems has been on the integration, annotation and rapid extraction of data from large, non-homogeneous data, which in itself is not a simple task. The fault is not actually with the data management system developers but rather in the inability for the research scientists to effectively recognize the complexity of the questions that need to be addressed and to communicate theses rather than questions that can be solved through the implementation of new technology. Thus, it has become somewhat apparent that the development of new technology frequently drives the science rather than the science driving the technology development.

There is no question that advances in computational algorithms (and grid computing) to support string comparisons and therefore drive solutions to problems such as sequence annotation have greatly improved the ability to qualify and classify the large data sets coming from the genomic sequences of humans and other organisms. Analogous development of hardware and software support the ability to solve large families of differential equations to approach the simulation of cell behavior and pathway. Both types of solutions, however, beg the scientific questions: (1) While sequence homology is used to imply structural and functional homologies, hence its use for annotation, has it been substantiated mechanistically so that it can be used successfully in all the instances, that is, homology ranges, in which it is applied? (2) While simulation of the sets of differential equations requires a degree of accuracy and dependability in the data, as well as completeness and uniformity of conditions, not typically available in experimental results, what is the sensitivity of the simulation to real data issues and constraints? The underlying biological hypotheses upon which the technology is being applied sits on a slippery slope that the technology, by itself, cannot address.

There are many scientists who study non-human organisms, but when we talk of bioinformatics applications in the context of cell and molecular biology, we are frequently using it as a surrogate. This extension towards the human organism significantly increases the complexity of the research problem and confounds it because of the more limited ability to develop adequate experimental results. Clinical data, including biochemical microarray and genetic analysis, tends to be less reliable, overall, than most cell and molecular biology data, particularly when disease diagnosis and clinical outcome are included. Thus, a major challenge to the translational researcher, that is, the research physician, is the dependency on the integration of data that ranges widely in both quality and quantity for use in research and evaluation. Such data ranges from quantitative laboratory results, to histological images, to written or transcribed comments entered into the patient's chart. All of this requires an additional level of technical complexity in the area of security because of the need to comply with HIPAA regulations as well as patient confidentiality, both of which typically cause academic medical centers to maintain separate networks for the clinical systems and university systems, for example, cell and molecular biology.

Computational Biology and Biomedical Informatics, Abramson Family Cancer Research Institute, University of Pennsylvania Cancer Center, Philadelphia, Pennsylvania.

PROBLEMS

1. There is an increasing need to capture temporal data in a manner that will enable the time domain to be readily accessible for analysis. Biological processes, including normal development and diseases, all take place over time.
2. High dimensional data needs to be evaluated for reduction of dimensionality as well as assessment of completeness for addressing the specific question under study.
3. Data needs to be "reduced" at the quality control level; this seems to be more common in the physical sciences than the biological sciences. Quality assessment of individual experiments as well as collections of experiments needs to be incorporated at the time of data entry into the database before further analysis can take place. This will enable better evaluation of threshold observation values, for example.
4. Non-parametric analysis needs to be further incorporated to reduce the bias that can occur in the (over) application of statistical methods. Statistics tend to be based on assumptions of data distribution or behaviors that are not always consistent with the mechanistic understanding of the biology. It is acceptable to use such statistical approaches when the resulting constraints on the analytical output can be established and conveyed to the scientist using the methods.
5. Patients (i.e. clinical data) do not present to physicians at equivalent time points in the same disease such that they are synchronized in their behavior and response to therapy—this is critical for correctly determining disease etiologies and their subsequent relation to underlying cell and molecular biology.
6. Disease diagnoses are significantly biased, in the United States, by issues of insurance reimbursement rather than for use in research applications and analysis. Thus this field type should be included as an annotation rather than a data type.
7. Pathways are probably dynamic constructs that are based upon topologies defined at the genome level and which actively respond through inclusion/exclusion of components segments under specific physiological conditions. They should therefore be incorporated in a dynamic graphical representation rather than a set of fixed relationships.
8. Graph data analysis needs to expand to incorporate higher dimensional representations of family stories or pedigrees, particularly to incorporate information theory analysis of incomplete family histories.
9. Epigenetic events, e.g. changes in standards of care, epidemics of infectious disease, advances in diagnostic technologies, need to be captured because of their potential for impact on clinical observations/decisions/diagnoses.
10. Text datamining needs to be addressed beyond the level of evaluation of abstracts of original research. This can be utilized to recognize concepts and relationships that may be outside conventional thought but are present in experimental observations and discussion.
11. Similarity needs to address multiple dimensions beyond size, shape and other morphological features, to include physical parameters and properties. Dissimilarity is an important concept not be to lost in its specification and analysis.
12. Ontology development needs to reflect biological process concepts, not only hierarchical alignment of data. Temporal data and alignment to enable capture and analysis of complex, parallel processes as actually occurring in nature is essential.
13. Evaluation of abnormal behaviors, for example, abnormal development, in disease, require better descriptions of normal behaviors, for example, normal development, in the healthy state. Characterization of normal is currently poor at best.
14. Phenotype definitions need to access temporal data to reflect biological process differences not just state differences.
15. Data normalization requires a multi-step approach with a potential meta-layer resulting in the development of a process-based ontology for describing the underlying biology/physiology.

This list is not all-inclusive but presents some of the issues that require cooperative and recursive interaction between the data management systems designers and implementers in support of the scientific users. It is readily known that the optimal use of existing technology in the data management area, by biological

and clinical scientists, is far behind the standard of practice, let alone the state of the art. The opportunities for research and new advances in data management system design will only result from a better understanding of where existing technologies do not address the ability to answer the complex questions of biology, and that requires the biologist learning how to ask those questions independent of the existing technology.

Address reprint requests to:
Dr. Michael N. Liebman
Computational Biology and Biomedical Informatics
Abramson Family Cancer Research Institute
511 BRB II/III
421 Curie Boulevard
University of Pennsylvania Cancer Center
Philadelphia, PA 19104

E-mail: liebmanm@mail.med.upenn.edu