TECHVIEW: COMPUTERS AND BIOLOGY

Bioinformatics in the Information Age

Sylvia J. Spengler

There is a well-known story about the blind man examining the elephant: the part of the elephant examined determines his perception of the whole beast. Perhaps bioinformatics—the shotgun marriage between biology and mathematics, computer science, and engineering—is like an elephant that occupies a large chair in the scientific living room. Given the demand for and shortage of researchers with the computer skills to handle large volumes of biological data, where exactly does the bioinformatics elephant sit? There are probably many biologists who feel that a major product of this bioinformatics elephant is large piles of waste material. If you have tried to plow through Web sites and software packages in search of a specific tool for analyzing and collating large amounts of research data, you may well feel the same way. But there has been progress with major initiatives to develop more computing power, educate biologists about computers, increase funding, and set standards.

For our purposes, bioinformatics is not simply a biologically inclined rehash of information theory (*1*) nor is it a hodgepodge of computer science techniques for building, updating, and accessing biological data. Rather bioinformatics incorporates both of these capabilities into a broad interdisciplinary science that involves both conceptual and practical tools for the understanding, generation, processing, and propagation of biological information. As such, bioinformatics is the sine qua non of 21st-century biology.

Analyzing gene expression using cDNA microarrays immobilized on slides or other solid supports (gene chips) is set to revolutionize biology and medicine and, in so doing, generate vast quantities of data that have to be accurately interpreted (Fig. 1). As discussed at a meeting a few months ago (Microarray Algorithms and Statistical Analysis: Methods and Standards; Tahoe City, California; 9–12 November 1999), experiments with cDNA arrays must be subjected to quality control. Variables as simple as temperature and illumination differences across a microarray slide can alter readings. Between slides, additional variables add to the difficulty of comparison. For example, John Quackenbush (The Institute for Genomic Research) described the complexities associated with assuring quality control between microarray slides in a presentation both humorous and disquieting in which he demonstrated how air conditioning can affect sample readouts. Manfred Zorn (Lawrence Berkeley National Laboratory, LBNL), chair of a working group on standards, launched a preliminary effort to lay down definitions and standards for microarray analysis with particular emphasis on experimental design, measurement, and analysis documentation.

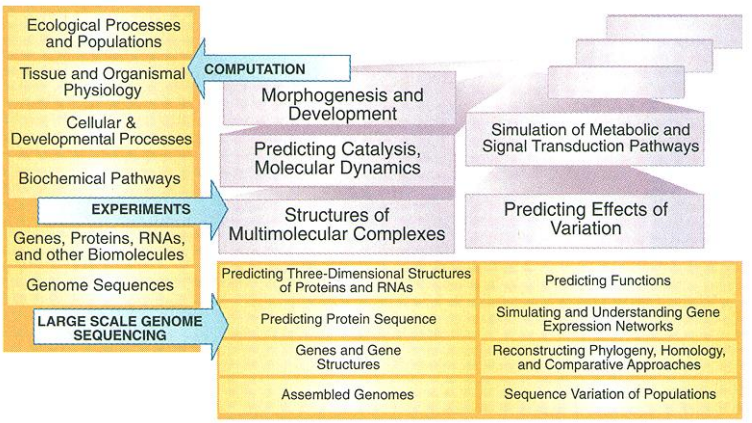

Fig. 1. The increasing need for centralized simulation and modeling to understand biology. The biology community requires extensive, integrated computational facilities to handle the wealth of data generated by, for example, cDNA microarray analysis.

New techniques for looking at the function of proteins through data mining of completed genomes continue to be developed. David Eisenberg's group (University of California at Los Angeles) has combined two techniques to facilitate the identification of protein function. The first, called the phylogenetic profile method, looks at the correlation of protein inheritance across different species. Each protein is given a phylogenetic profile denoting the presence or absence of that protein in various genomes. The result is that function can be assigned to uncharacterized proteins if they have a phylogenetic profile similar to the model profile (*2*). The second approach, the Rosetta Stone method, is a way of looking at the correlation of protein domains across species. Some proteins have homologs that are fused in other species, yielding clues as to the proteins with which they might interact. In addition, proteins that have been identified in particular complexes and pathways hint at the location and function of their homologs in other species. Together, these methods are leading to genome-wide predictions of protein function. Since the sequencing of the entire yeast genome 3 years ago, more than half of the then-uncharacterized yeast proteins have now been assigned a general function. Similarly, over 40% of the previously uncharacterized proteins in *Mycobacterium tuberculosis* have been assigned a function (*3*).

Comparing the genomes of different species also offers insight into the function of conserved noncoding regions of DNA sequence. At LBNL, Kelly Frazer and others are looking for conserved noncoding elements with greater than 70% identity over more than 100 base pairs in the mouse and human genome (*4*). Focusing on the sequences between genes, they selected a subset of sequences conserved in the human and mouse genome and searched for these sequences in additional species. Seventy percent of the elements examined were conserved in mammals, but 30% were unique to the human genome (*4*). This presents an opportunity to improve comparative sequence analysis methods and to identify distant elements conserved across species. Inna Dubchak at the Center for Bioinformatics and Computational Genomics (LBNL) has developed a new method for global alignment of sequences, their comparison, and the display of their identity. The algorithm (called Moving Average Point analysis, MAP) has identified experimentally determined conserved sequences in mouse, dog, and human genomes and will be the basis for a Conserved Sequence Index, funded by the U.S. Department of Energy and currently under development.

Organisms are, however, greater than the sum of their genomic parts. Acknowledgment of this complexity can be found even at the level of a single bacterium. Lucy Shapiro (Stanford University) has sought to define the genetic network that coordinates

The author is at the Center for Bioinformatics and Computational Genomics, LBNL 84-171, 1 Cyclotron Road, Berkeley, CA 94720, USA. E-mail: sjspengler@lbl.gov

the initiation of DNA replication under both temporal and spatial constraints (*5*). This complex network expands on earlier work that looked at DNA replication in the lambda phage and discovered the lysis-lysogeny switch. Adam Arkin (University of California at Berkeley) is building a computational toolkit for computer-aided simulation and analysis of developmental switches such as the lysis-lysogeny switch. His program—called BIO/SPICE (Simulation Program for Integrated Circuit Emulation) after the general-purpose circuit simulation program with built-in models for semiconductors—permits the experimenter to simulate chemical kinetic systems and parts of cellular pathways.

What are some of the problems facing biologists wishing to analyze and interpret vast amounts of data generated by genome sequencing, gene chip technology, and genome data mining? One of the most pressing problems is the limitation of computers with respect to analysis time, storage capacity, movement of data around the network, and data complexity. Many researchers resort to running their analyses for weeks or months at a time on home-assembled workstations. There is a better way, but usually there is a big gaping hole regarding personnel and financial support for expertise and hardware. A Beowulf assembly of many CPUs can handle parallel problems for a very reasonable price (less than the cost of a DNA sequencing machine), but such a setup requires a full-time person to run the assembly. Clearly, it is shortsighted to gather large amounts of data that cannot be analyzed in a timely manner because the computational power is not available. Although there is disk space to store everything researchers write, say, perform, or photograph, there is not enough computing power to analyze the data. There is also not a quick and easy way to transfer vast amounts of data from one institute to another, one computer to another, or one network to another. This will be the challenge of the next generation of the Internet. A further challenge will be to adequately search, analyze, and assess the quality of the data. Isolated ad hoc systems may be sufficient for individual databases, but they do not support the flexibility or integration that is required for modern research. When a valuable software product is developed, the lure of commercialization and the subsequent loss of skilled people from academia to industry is to be expected.

One of the biggest problems that demands international consideration is the question of standards and intellectual property rights. Time alone will not solve the problem, but rather will only make it more intractable. BioCat (http://bioinformer.ebi.ac.uk:80/newsletter/archives/4/biocat.html) listed 571 software programs in 1998. An update would double this number. In order to deal with the problem, the European Bioinformatics Institute (EBI) has developed two great ideas. The first is "copylefting," a concept from the computer software industry (http://www.gnu.org/copyleft/gpl.html) that supports open source and availability and prevents inclusion in proprietary programs. This does not, however, prevent further commercialization—the LINUX kernel has a GPL (General Public License). Heikki Lehavshaiho of the EBI suggests extending the concept to biological databases. The EBI Biostandards Project—originally funded through the EU, EBI, and participating biotechnology companies—has been a leader in CORBA (the Common Object Request Broker Architecture) developments in the Life Science Group. In the future, however, the project will be funded as an Associates' Program, which will decrease access to the standards.

Grant funding agencies in the United States such as the National Institutes of Health (NIH) and the National Science Foundation (NSF) have taken steps to address the shortage of computing hardware, software, and skilled bioinformatics personnel. The Working Group on Biomedical Computing of the NIH acknowledges the pivotal role of bioinformatics in biomedical research. In their June 1999 report on the Biomedical Information Science and Technology Initiative (BISTI), they made four recommendations (*6*). Key features of these recommendations include establishing

(1) Programs of Excellence in specific areas of biomedical research (part of this initiative would involve creating up to 20 centers at U.S. universities and independent research institutes with a special focus on bioinformatics).

(2) A program for information storage, curation, analysis, and retrieval (ISCAR).

(3) Adequate resources for computational support at the R01 (single principal investigator grant) level.

(4) A national computer infrastructure that can be scaled up as the amount of data grows.

The Working Group also recognized the shortage of biologists with appropriate computing expertise and called for strong NIH support of cross-disciplinary education and training within the Programs of Excellence. David Botstein, a geneticist at Stanford University and co-chair of the Working Group, pointed out the need for "people as professional in computing as in biology, just as we once needed people as competent in chemistry as biology." The pharmaceutical industry, agrobusiness, and biotechnology companies often raid academia and each other for people with the appropriate interdisciplinary skills.

Since the BISTI report was published, a number of institutes in the United States have developed RFAs (grant applications in a specific area) that include requests for funding for bioinformatics tools and databases. It is not clear, however, that the spirit of BISTI will permeate the NIH study sections, which hand out federal money in the form of grants to biomedical researchers.

At the NSF, director Rita Colwell has set information technology as the highest priority in the NSF budget for 2001. The proposed budget request was $230 million for information technology, an increase of 43%. NSF and NIH have been exploring a partnership in computational biology, particularly in the area of supercomputing.

Bioinformatics groups and institutes in Europe, Japan, and Australia are exploring the same areas scientifically and are facing some of the same funding problems. EBI and other European-funded life science centers ran into a funding buzz saw last fall. The crunch was over funding infrastructure, the databases and tools that make the EBI headquarters in Hinxton, U.K., a world-class facility. Although the crisis is temporarily stayed, there has been no clear resolution of funding for EBI databases and other infrastructure in the next few years. This shortsightedness is not universal. Indeed, member nations of the EU individually are taking an aggressive stance. Research and training programs, with funding, have been announced in France, Germany, Israel, and Switzerland. In the Pacific Rim, both Japan and Australia have new programs that will support bioinformatics networks.

As we enter the postgenomic era, it is becoming clear that the really interesting aspects of biology go far beyond assembling genomes and finding genes, and even beyond predicting the function of endless DNA sequences. Bioinformatics and data generation must continue to develop hand in hand to enable us to understand the complexities of cells, development, physiology, and populations. It will be exciting to watch the cooperation between bioinformatics and biology in the coming years.

References

1. C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, 1968).
2. M. Pellegrini *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 4285 (1999).
3. E. M. Marcotte *et al.*, *Nature* **402**, 83 (1999).
4. G. G. Loots *et al.*, personal communication.
5. I. J. Domian *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6648 (1999).
6. D. Malakoff, *Science* **284**, 1742 (1999). Information on BISTI report available at www.nih.gov/welcome/director/060399.htm