BIOINFORMATICS EDUCATION

# Hedgehogs, foxes, and a new science

*Charles DeLisi*, Charles Cantor, and Zhiping Weng*

C.P. Snow's classic, *The Two Cultures and the Scientific Revolution,* laments the lack of communication between the sciences and the humanities. The problem, however, is general, and occurs even within the sciences themselves. The training of information scientists and biologists is a case in point. Computers have revolutionized virtually every aspect of life from business and entertainment to the hard sciences, but their diffusion into the biological sciences has been slow.

The first wake-up call occurred 15 years ago, when a routine search of a protein sequence database revealed a startling homology between the transforming protein of the simian sarcoma virus and human platelet-derived growth factor. A few years later, when the US Department of Energy (Washington, DC) initiated the Human Genome Project, it became obvious that the need to manage and analyze unprecedented quantities of data it would generate would bring sophisticated information technologies full force into the biomedical sciences. Bioinformatics, a word previously known to only a few specialists, entered the lexicon, and is now used commonly, if inappropriately, to encompass the full range of computational activities in the biological sciences, including the management, mining, and analysis of molecular, cellular, and systemic (e.g., ecological) databases.

Although the Genome Project was motivated in part by a desire to stimulate the biotechnology industry and to quicken the pace of discovery by bringing advanced technology into the generation and analysis of biological data, no one—certainly none of us—fully anticipated the magnitude of the changes that would occur. Conspicuously absent from the plans of the federal agencies was the all-important area of bioinformatics personnel.

Today, genomics—which includes the identification and functional characterization of genes, and uses such methods as positional cloning, high-speed screening, and bioinformatics—is generally regarded as the major road to accelerated drug target discovery and to the development of new and more sensitive diagnostics for the detection and prevention of disease.

Included in the information now in public databases are several hundred thousand sequences representing approximately three quarters of expressed human genes, 10 complete microbial genomes, and a high-resolution map of the mouse genome. The information doubling time of approximately three years is staggeringly short. As striking as these facts are, they trivialize the complexity of the information explosion, and they convey neither the difficulties of using the data nor its potential for revolutionizing medicine.

A hint at the future is perhaps best conveyed by considering the advances in the experimental methodology used to obtain genomic data. Among the most significant is the development of DNA microarray technology—hundreds of thousands of DNA samples displayed on solid-phase supports, read by hybridization with flourescent probes, and directly linked to microprocessors. One result is high-throughput measurement—with unparalleled specificity, accuracy, and reliability—of the expression of every gene in a given cell and, more importantly, the ability to rapidly compare levels of gene expression in normal and pathological tissue. This will make it possible to compare normal and diseased heart tissue; to compare normal and cancerous cells in various stages of transformation; and to stratify complex disorders into homogeneous disease categories. Such technologies will clearly generate information at a rate unimagined, even at the beginning of the decade, and will revolutionize our understanding of disease diathesis, resistance, and pathognomy.

Rapid progress presupposes an ability to organize, manipulate, and analyze the data, and here, unfortunately, the problems are substantial. They arise not just from the sheer volume and complexity of levels of information, but from the inevitable variability in data structuring and annotation. A more fundamental problem is the lack of a large cadre of talented biomedical scientists trained in information technologies. The supply/demand mismatch in the United States is illustrated by current industrial salaries, which can exceed $70,000 per year for a well-trained recent MS degree recipient.

The problem in developing countries, where data access is severely limited, is even more critical. Easy access to genetic information will become increasingly important worldwide as the same techniques that have been developed for medical use are applied to plant genetics with the goal of developing high-yield grains and robustness against inhospitable climates.

Universities in developed nations of the world must immediately begin to play a substantial role in addressing these problems. At Boston University, we have recently introduced a university-wide program in bioinformatics that will span the disciplines of information science and biology. A four-day tutorial to meet the needs of local industry is planned for early September[1]. In the fall of 1997, the program will introduce a course on protein and DNA sequence analysis, focused on the theories and algorithms for sequence comparison, protein structure modeling, display, and function identification. A second new course, "Biological Database Analysis," will be offered in the spring of 1998; this will deal with database structures and methods for database integration. With three other courses already in place and more to be developed, the program will accept its first entering students for the fall semester of 1998. An important component of our program will be distance learning, or the use of interactive compressed video, delivered either via satellite or by high-speed telephone lines. The delivery will not be confined to the local Boston area, or even to the United States. An experiment with delivery to Mexico is planned for the spring semester, and we hope to broaden this to other developing countries in subsequent years.

Our commitment to educating scientific leaders who are knowledgeable in both the information and biological sciences emerges from the conviction that the requirements for advanced computational methods are here to stay, and their use will no doubt grow, evolve, and change with the biological and medical sciences. It is even possible that a new discipline is emerging, one that demands an education that cuts across biology, medicine, and mathematics. The practitioners of this discipline will be generalists rather than specialists; foxes, rather than hedgehogs, to borrow Isaiah Berlin's famous metaphor[2]. Berlin's colleague, C.P. Snow, would undoubtedly have taken heart that at least one cultural divide will be narrowed. But the driving force to do so is substantial, being nothing less than world health and a changing global economy.

*Charles Cantor is a professor, and Zhiping Weng is an instructor, of biomedical engineering, and Charles DeLisi is dean of the college of engineering, Boston University, 44 Cummington St., Boston, MA 02215 (delisi@enga.bu.edu).*
**Corresponding author.*

[1] http://engpub6.bu.edu:7105/bioinfo/workshop.html

[2] Berlin, I. *The Hedgehog and the Fox.* A study of Leo Tolstoy's view of history where Berlin distinguishes two types of thinkers: Generalists like the fox, who know many things, and specialists like the hedgehog, who know one big thing.