

BYTING THE BULLET

Computers will play a major role in the biology labs of tomorrow
p683

IN SEQUENCE

Proteomics is revolutionizing protein discovery
p684

SKILLS SINK

US universities do the training — but industry is draining the talent
p686

DATA DILEMMAS

More specialists are needed to make sense of sequencing information
p687

Next-generation biologists must straddle computation and biology

Biological sciences are being transformed by the huge amounts of data emerging from newly sequenced genomes. But do the skills exist to cope?

Like a circus performer juggling daggers, chairs and flaming clubs, the bioinformatics field stands amid elements ranging from human genetics and clinical medicine, to biophysical studies of three-dimensional structure and studies of fruit-flies, yeast and bacteria.

As the field evolves and expands, so too do the skill sets necessary to excel in it. To attain competence demands new ways of thinking about knowledge and education, both by students and universities.

“The basic mentality clash between data-driven scientists and model-driven computer and math people stems from world views so different that all too often they can’t have a conversation that makes sense,” says Mark Perlin, chief executive officer of the genomics company Cybergenetics.

Different world views must coexist in one person to bridge the gulf that separates computer scientists from life scientists, says Perlin. Because phenomena and data are indivisibly associated, he says, there’s no point in having people analyse terabytes of data if they have never spent time in the laboratory seeing what the data actually are and how they are obtained. “Until you’ve held a pipette in your hand, or run a PCR or a gel, you can’t understand what the failure modes are going to be,” Perlin explains.

Embrace the unknown

How can one achieve multidisciplinary competence in a rapidly changing field? Perlin asserts: “Those who seriously want to get into this field need to embrace what they don’t know. Take courses in things you never studied before while working on some research project, and after some time you’ll begin to understand what the people around you are talking about.”

Quality interdisciplinary training doesn’t come easily, though. It takes a month to introduce people to problems and two years to train a technician to do useful programming or lab work, but longer still to train a principal investigator.



Multiple-degree dynamics

When he considered augmenting his internal medicine studies with a mathematics PhD in 1978, Mark Perlin, now chief executive officer of Cybergenetics, a Pittsburgh-based genomics software firm, didn’t exactly receive encouraging words. “Computer science and mathematics are not, and never will be, relevant to medicine,” the dean of the University of Chicago told him then.

How times change. The path that Perlin chose over 20 years ago could be a template for today’s bioinformatics professional. Perlin, who is also an adjunct faculty member at Carnegie-Mellon University and the University of Pittsburgh, earned a PhD in mathematics from the City University of New York, an MD from the University of Chicago, and a PhD in Computer Science from Carnegie-Mellon University.

His decision to pursue mathematics and computer programming along with medicine seemed logical, not trail-blazing. “Math broadened my view of medicine, so it seemed natural to interrupt medical studies midway through and study math for three years,” he says.

He has no regrets now, but notes that some of that academic inflexibility that he faced initially still exists. “Biology is changing rapidly, but university departments are persistent forms, and rewards are distributed along department lines,” he says. Those boundaries stifle innovation: “I have more freedom for genomics research in a start-up company than in a university department.”

But earning that freedom didn’t come easily. After medical school, he did a six-month postdoctoral stint at IBM’s T. J. Watson Research Center in Yorktown Heights, New York, a one-year transitional medical residency in Pittsburgh, then joined the Carnegie-Mellon University faculty in computer science. While there, he earned a PhD in computer science, in the hope of embarking on an academic career.

But when he realized that academia could be “limiting”, he launched his own company. “I started Cybergenetics in 1994 because I had a sense that the private small business model might be a better way to innovate and do research.”

P.W.

Traditional departmental lines must become more blurred, says Jehoshua Bruck, a professor of computer science and electrical engineering at Caltech, in Pasadena, California. For example, the ideal next-generation biologist should develop both wet-lab expertise and software-writing ability.

Teamwork is key

Reorganizations along these lines are already under way at some institutions. The new Clark Center at Stanford University integrates traditionally separate disciplines, such as engineering, molecular biology, physics and computer science, to work on problems in common. Politically and financially the long-term viability of research

work groups, as opposed to more familiar departments, is likely to depend on adjusting how tenure is awarded and how funding and authorship priorities are allocated.

This more interdisciplinary, data-driven approach to biology has already begun producing results. Before Christoph Sensen, manager of the Canadian Bioinformatics Resource, in Halifax, Nova Scotia, saw a draft of the genome of *Sulfolobus solfataricus*, he thought that, on the basis of ribosomal sequences, the organism’s genome was well organized, with no space between the genes.

When he examined the *Sulfolobus* genome, he found large spaces between genes, as well as repeats comprising up to a third of the genome. “We learned that the

dogma that the prokaryotes are made with total economy is not so true," Sensen says. "This has made us change the way we think and the way we operate in the lab."

Challenges ahead

Nevertheless, the field must develop further to be more useful. In important ways, genomics has so far been built on scaled-up versions of classical methods. Although satisfactory for the linear task of factory-style sequencing, these will not accommodate the more complex problems that are emerging from a burgeoning quantity of data from a variety of sources.

At present, computational approaches to deal with these new forms of data are developed in an *ad hoc* way, says Chris Lee, at the Bioinformatics Institute in the University of California, Los Angeles. For example, to cluster genes by expression patterns, computational biologists generally try to define a metric that says "these genes express pretty similarly", Lee says. But that approach lacks both a model of the underlying phenomena and a rigorous computational method that can consider all possible interpretations.

More mundane but more immediate obstacles also clutter the way ahead. George Poste, former chief science and technology officer at pharmaceuticals giant SmithKline Beecham, frets about the lack of consistent nomenclature, and of standards to integrate research, trials and clinical databases. "It's no good realizing ten years from now that the research data can't be migrated downstream," he says.

Francis Ouellette, director of the Bioinformatics Core Facility at the Center for Molecular Medicine, in Vancouver, British Columbia, points to the problem of working digitally on tables and figures in thousands of papers that today exist only in print.

Mistrust by the public also provides a potential worry. Reaction to privacy and ownership issues, now mostly latent, could erupt, with unfortunate consequences for research funding and the licensing of new technologies. For proof of this, one need only recall the setbacks that agricultural biotechnology has suffered in Europe by its failure to allay the public's concerns about genetically modified foods.

Despite the growth of bioinformatics, biology in the age of genomics still has a great distance to go.

Getting connected to bioinformatics

For more information on bioinformatics techniques and general resources see:

- Bioinformatics courses linkage.rockefeller.edu/wli/bioinfocourse/
- Canadian Bioinformatics Resource www.cbr.nrc.ca/
- Cambridge Healthtech meetings www.healthtech.com/
- Cold Spring Harbor short courses nucleus.cshl.org/meetings/2000c-info.htm
- GOLD: Genomes Online Database geta.life.uiuc.edu/~nikos/genomes.html

- Human Genome Organisation www.gene.ucl.ac.uk/hugo/
- IBM Computational Biology Center www.research.ibm.com/topics/serious/bio/
- Kyoto Encyclopedia of Genes & Genomes kegg.genome.ad.jp/kegg/
- National Center for Biotechnology Information www.ncbi.nlm.nih.gov/
- National Center for Genome Resources www.ncgr.org/
- UCLA Bioinformatics Institute www.bioinformatics.ucla.edu/

Political conflict presents another obstacle to advancing the professions. The apparent unravelling of the public-private collaboration in genomics in the United States is demonstrated by recent letters between Francis Collins, director of the National Human Genome Research Institute, and Craig Venter, president of Celera Genomics (see www.bioinform.com).

Each side accused the other of bad faith in holding up their side of the bargain (see *Nature* **404**, 117; 2000). This exchange hints at the difficulties that lie ahead in reconciling public and private interests in a rapidly growing enterprise whose private arm is

already worth in the order of \$45 billion.

Despite the growth of bioinformatics, biology in the age of genomics still has a great distance to go. Christoph Sensen observes that genomic sequences dating from 1986 have open reading frames that are still of unknown function, and the number of these early sequences is dwarfed by those now emerging.

Resolving those problems will require a new breed of scientists who are comfortable both with the old lab skills and with new computational techniques. "To grasp the complexity will require much more wet-lab activity and much more analysis," Sensen concludes.

Potter Wickware

Companies of all sizes are prospecting for proteins

It is being called a land grab at the UK-based company Oxford GlycoSciences, but the property being prospected for is intellectual — patents on proteins. Fuelled by \$50 million raised at the end of February on the London Stock Exchange, the company is racing to patent as many proteins as possible during the next two years. In mid-March it announced that it had filed patent applications covering more than 800 different protein-use combinations.

Certainly, investors seem impressed by the company's aim, and its share price shot up from a little over £5 in early January to £32 in the second week of March, despite a warning from Michael Wandra, the chief executive officer, that extra investment is likely to be postponed to beyond 2002.

The multiplier in this equation is, of course, the Human Genome Project. The race to complete the human genome is nearing the end, and as the databases fill up with DNA sequences, and increasing numbers of genes are identified with greater certainty within those sequences, companies such as Oxford GlycoSciences and its rivals Large Scale Biology Corp. in the United States have

access to a powerful new tool for identifying proteins and protein complexes.

Why all the excitement and financial faith? Proteins orchestrate the life and death of all living organisms. They are the main product coded for by genes. When something goes wrong with a protein, it can cause disease. So identifying the proteins and protein complexes associated with all processes in healthy and diseased organisms is critical to an understanding of human biology. It is a gargantuan task, which only a decade or so ago was a highly specialized cottage industry. Now the task has acquired a new name — proteomics — and it is on the threshold of becoming a highly mechanized industry (see *Nature* **403**, 815–816; 2000).

Out of sequencing

During the cottage-industry days, a protein could be identified only by isolating it and then laboriously determining the sequence of its amino acids. Finding all the proteins with which it interacted and identifying them was several orders of magnitude more difficult. Now sequencing is not needed.

Instead, highly accurate mass spectrom-