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.

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 thousand nucleotides. He found that the genome is "open," with no special structure.
**Getting connected to bioinformatics**

<table>
<thead>
<tr>
<th>For more information on bioinformatics techniques and general resources see:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics courses</td>
</tr>
<tr>
<td>Canadian Bioinformatics Resource</td>
</tr>
<tr>
<td>Cambridge Healthtech meetings</td>
</tr>
<tr>
<td>Cold Spring Harbor short courses</td>
</tr>
<tr>
<td>GOLD: Genomes Online Database</td>
</tr>
</tbody>
</table>
| 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.bioinf.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. **Companies of all sizes are prospecting for proteins**

I t 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 £3.2 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-