Are you ready for the revolution?

If biologists do not adapt to the powerful computational tools needed to exploit huge data sets, says Declan Butler, they could find themselves floundering in the wake of advances in genomics.

SDSC

Shape of things to come: large data sets arising from genome projects demand new skills of biologists.

"We're in the middle of a messenger RNA in a gene here, *TCRA*, apparently. Before, we were at the base-pair level, now we have zoomed out to 53,000 bases. Let's zoom out again, click on the 10 × -out button — now we are at half a million bases." David Haussler, a computer scientist at the University of California at Santa Cruz, is giving me a guided online tour from my workstation in Paris of a stretch of chromosome 14 using a web-based human-genome browser.

A few minutes into the demonstration, Haussler heats up. "Hey! Look over there at that SNP track: I see one right in the middle of an exon at the bottom of the screen; there is a line-up of all those bars. I'm going to click on that to see what we get. OK, it says it's RS1197. That's a hell of an interesting SNP to look at. Hey, it's fascinating! Every time I demo this, we always discover something new and interesting! This means it's a coding SNP."

If you are one of the many biologists for whom genome databases are as comprehensible as a mass of supermarket barcodes, Haussler's enthusiasm probably leaves you cold. If so, the publication this week in *Nature* and *Science* of draft human genome sequences marks a good time to team up with a friendly bioinformaticist and join the action, before it is too late.

Brave new world

Many research leaders predict that the potential to integrate different levels of genomic data — such as raw sequence from the human genome and those of model organisms, data on genetic variability between individuals and on gene expression in different tissues — will radically change biological research. They argue that small experiments driven by individual investigators will give way to a world in which multidisciplinary teams, sharing huge online data sets, emerge as the key players. Some foresee an era of 'systems biology', in which the ability to create mathematical models describing the function of networks of genes and proteins is just as important as traditional lab skills.

If so, those who learn to conduct high-throughput genomic analyses, and who can master the computational tools needed to exploit biological databases, will have an enormous competitive advantage. Some experts even predict that the outcome of this natural selection will see many current top scientists, research groups and even whole institutes relegated to the second division.

"Every institution that expects to be competitive in this new era will need to have strengths in high-throughput genomic analyses and computational approaches to biology," says Francis Collins, director of the National Human Genome Research Institute in Bethesda, Maryland.

Søren Brunak, director of the Center for Biological Sequence Analysis at the Technical University of Denmark in Lyngby, is more blunt: "Biologists for the past 20 years have been of reasonable quality, but now we will see those who are good and those who are bad. Either you use these new methods in all aspects of your work, or you go on at your old speed and get left behind."

Given such gung-ho predictions, many researchers are becoming concerned about the fate of traditional 'wet-lab' biologists. According to David Jones of the Protein Bioinformatics Group at Britain's University of Warwick, some biologists are already having their grant proposals turned down because they lack a bioinformatics component. "This is certainly going to become more common, and for a while there will be some gnashing of teeth as biologists take stock of this," he says.

Ewan Birney of the European Bioinformatics Institute (EBI) in Hinxton, near Cambridge, identifies the need to transfer the skills of his discipline to the wider biological community as the single biggest challenge ahead. He fears that many biologists risk being "disenfranchised". "Even a biologist doing very directed experiments is not going to be able to avoid large-scale data gathering and analysis," Birney predicts.

Gene surfing

The simplest program in the bioinformaticist's tool-box is known as BLAST (basic local alignment search tool). This allows a biologist who has identified an interesting gene sequence to compare it against those held in genome and protein databases in order to check, for example, whether it corresponds to a gene that has already been characterized. Then there are computational tools such as Genscan and GeneWise, which predict the occurrence of genes with-

TEK/SPL

Petrified? Lab-based researchers will need to swallow their fears and embrace computational biology.

in raw sequence data. And genome browser programs allow users to see whole chromosomes and to zoom in on regions of interest, while simultaneously superimposing associated information such as data on homologous genes from other species.

Now tools are being developed by several groups to analyse entire genomes and biosynthetic pathways, and to integrate multiple levels of information into mathematical models of cells and even organisms — introducing to biology the high-end computing previously restricted to fields such as astrophysics and climate modelling.

So far, expertise in using these tools has tended to reside in the main genome centres, or in specialist bioinformatics institutes such as the EBI or its US counterpart, the National Center for Biotechnology Information in Bethesda. Leroy Hood, director of the Institute for Systems Biology in Seattle, describes this separation between bioinformaticists and experimental biologists as a "fatal flaw". "I would argue very strongly that you have to integrate very tightly the bioinformatics with doing experiments," he says.

Birney accepts that people like him have to redouble their efforts to reach out to experimental biologists. "Bioinformaticists must become very focused on being a service community," he says. "We have a duty to make the data more usable in simpler ways through the web."

In the meantime, many biologists who are trying to get up to speed with bioinformatics are using only the simplest tools available. Frequently, says Teresa Attwood, an expert in protein sequence analysis at the University of Manchester, they fall into "the black box trap", accepting without question the matches returned by BLAST without taking the time to learn the caveats of such techniques.

Similarly, computational tools used to assign function to sequences have inherent limitations; put them in the hands of someone who is unaware of this and you have the makings of havoc. As a result, the public genome databases are already littered with annotations that are inaccurate, misleading or downright wrong.

Dig the new breed

So if the majority of biologists are not to be disenfranchised, and the databases are not to become terminally polluted with misleading information, what is the solution? In the long run, change will come through the emergence of a new breed of biologists who are steeped in computational biology as an integral part of their education. This means that the subject must be included as a core module in all undergraduate biology courses, rather than as a specialist option. Although this is starting to happen, the availability of teachers with the appropriate expertise is still a limiting factor.

Funding agencies are also trying to drive change by ploughing money into initiatives that require a multidisciplinary approach and a strong computational component. The US National Institutes of Health, for instance, through its National Institute of General Medical Sciences in Bethesda, has created a programme of 'glue grants' for integrative and collaborative approaches to research. Under this programme, the Alliance for Cellular Signaling, headed by Al Gilman at the University of Texas Southwestern Medical Center at Dallas, will receive up to $25 million over the next five years to draw up a complete map of the interactions between some 1,000 proteins in two types of cell. The consortium unites traditional experimentalists with computational biologists, with its membership drawn from institutions across the United States. "The glue grants is an exemplar flagship initiative that is refocusing the approach to biology research," says Shankar Subramaniam, a bioinformaticist at the San Diego Supercomputing Center, based at the University of California at San Diego.

Some bioinformaticists are also working on 'outreach' activities. Birney, for instance, heads Ensembl, a joint project between the EBI and the Sanger Centre, also in Hinxton. This seeks to simplify things for mainstream biologists by making annotated genome sequences available through an intuitive web-based interface. Others, such as Subramaniam, are designing user-friendly

HEIKKI LEHVASLAIHO

Ewan Birney fears that many biologists will be disenfranchised by the changes occurring.

bioinformatics 'workbenches', which combine various computational tools.

Hood, meanwhile, hopes to drive the ethos of systems biology into the wider biological community. He plans to invite talented biologists to Seattle to take a crash course in working on large-scale data sets, so that they can then go back to their labs to use the data to pursue hypothesis-driven research. "Say you have a top-flight scientist working on the *p53* tumour-suppressor gene in Paris," says Hood. "We would bring him out; show him the kind of tools we have here, show him how you can interrogate whole systems at a time, and show him the utter importance of being able to integrate data from these different levels."

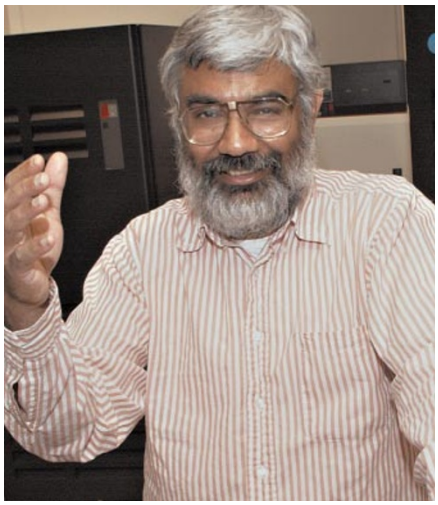
Hood's crash course may sound a little daunting. But if you decide to enter the world of bioinformatics, and find an expert who is willing to share his or her skills, it may prove easier than you think. The major problem, says David Roos, director of the Genomics Institute at the University of Pennsylvania in Philadelphia, is "lack of familiarity — not even knowing where to begin — and lack of people to provide that familiarity". Birney's advice to scientists debuting in bioinformatics is disarmingly simple: "Don't worry if you feel like an idiot, because everyone does when they first start."

In this regard, the experience of some of the biologists commissioned by *Nature* to mine the human genome for interesting nuggets of data is instructive (see pages 827–859). "The more one works, the more confident one becomes," says Rossella Tupler of the University of Massachusetts Medical School in Worcester, who searched for genes involved in processes including transcription and RNA splicing (see pages 832–833). "During the analysis for *Nature*, I increased my ability to analyse the data. I also felt very happy about the results of my search. We found several new genes."

Apocalypse postponed?

Given experiences such as these, some experts argue that the apocalyptic visions of thousands of top-notch biologists being suddenly cast aside by the genomic revolution are wide of the mark. "Molecular biology has smoothly absorbed any number of technological 'revolutions' as fast as the next generation can learn the new techniques," says Sean Eddy, a self-taught computational biologist at Washington University in St Louis. "Genomics and computational biology won't be any different. Molecular cloning, DNA sequencing and the polymerase chain reaction are all examples of revolutionary techniques. I bet you can dig up old quotes from the 1970s of people saying, 'Boy, we're all going to lose our jobs now. How will we ever retrain all the old guard to clone and sequence?'"

Stan Fields, a geneticist at the University of Washington in Seattle, adds that biologists studying the yeast *Saccharomyces cerevisiae*, which had its complete genome published in 1997, seem to be rising to the challenge. "Here, researchers from a predominantly genetic, small-laboratory approach have been confronted with the leading edge of genomic strategies," he says. "I think that this community has been highly successful."



SDSC

Easy does it: Shankar Subramaniam aims to make computational tools biologist-friendly.

When push comes to shove, says Gene Myers, vice-president for research informatics at Celera Genomics in Rockville, Maryland, molecular biologists will acquire skills in bioinformatics out of sheer necessity. "As traditional biologists see their colleagues outpace them because of technology, they will change," he says.

Nevertheless, Sylvia Spengler, a programme director at the US National Science Foundation's Division of Biological Infrastructure, believes that most biologists will need help to exploit large-scale data sets at the levels of sophistication needed. "You may not know how to use an integrated suite of tools," she says.

Among officials such as Spengler, one major concern is whether universities are geared up to provide the environment needed for bioinformatics and systems biology to flourish. This will require a truly multidisciplinary approach, cutting across departmental boundaries and embracing partnership, rather than competition, between major research universities in providing the necessary infrastructure. Many US universities have launched bioinformatics programmes and initiatives in computational biology, says Spengler. "But they are tiny, and they are very inward looking. It is not clear to me that any programme believes its mandate is to serve the entire community," she says . "Yale thinks it should serve the Yale community, Rockefeller thinks it should serve Rockefeller. But the picture is bigger."

Hood, who had to leave the University of Washington to get his Institute for Systems Biology off the ground, is more disparaging. "The leadership in academia is utterly lacking, and their scientists are going to be at an enormous competitive disadvantage if they don't see the revolution that is coming," he says. "The new biology requires teamwork and putting together different kinds of science, and the current model of tenure is very much against that; in the United States it is based on what have you done yourself."

All systems go

Similar concerns are being expressed on the other side of the Atlantic. "Increasingly, university researchers are moving further away from the 'coal face'," says Warwick's Jones. "Without some radical changes in the types of research, and the levels of funding support, university biology departments are in danger of becoming spectators in genomics." Some scientists fear that the action may become concentrated in multidisciplinary institutes with the vision to embrace systems biology, and in industry — which has wasted no time in getting up to speed with computational biology.

But once again there are some who feel that these visions are the product of excessive revolutionary zeal — or are at least premature. "The bulk of contemporary molecular, cellular and model-organism biology is a tremendously successful enterprise and is doing fine without us," says Roger Brent, associate director of research at the Molecular Sciences Institute in Berkeley, California, and a leading figure in systems biology.

Tom Pollard of the Salk Institute for Biological Studies in La Jolla, California, who was commissioned by *Nature* to mine the genome for information relating to the cytoskeleton (see pages 842–843), agrees. "I have a personal fear that full genome sequences will prove irresistibly seductive, perhaps even deferring the hard wet-lab work required to complete the mechanistic analysis needed to understand physiology," he says. "Understanding function requires lots of hard-won information that can only emerge from analysis of one or a few types of molecules in a single experiment."

But what seems clear is that the research teams that will be most successful in the coming decades will be those that can switch effortlessly between the lab bench and a suite of sophisticated computational tools. If you are not already as *au fait* with computational biology as you are with gels and culture dishes, you have been warned. ■

Declan Butler is *Nature*'s European correspondent.

University of California at Santa Cruz Genome Browser
http://genome.ucsc.edu

European Bioinformatics Institute
http://www.ebi.ac.uk

National Center for Biotechnology Information
http://www.ncbi.nlm.nih.gov

Ensembl
http://www.ensembl.org

San Diego Supercomputer Center Biology WorkBench
http://workbench.sdsc.edu