funding training slots, and a fourth branch of the NIH, the National Center for Research Resources (which is not an institute), has put itself behind shared bio-computational resources at more than a dozen centres nationwide. The Department of Energy's Argonne and Oak Ridge laboratories are also huge funders of bioinformatics work, as is, to a somewhat smaller extent, the Department of Defense.

On the private side, the Howard Hughes Medical Institute (HHMI) has declared that

Borrowing methods and models

Kimmen Sjölander's early work in using Hidden Markov Models (HMMs) in genomics illustrates how bioinformatics reaches backwards to move forwards. HMMs originated in machine learning and were developed for speech

recognition, but they turned out to be "fantastically successful" for modelling proteins, says the scientist from Celera Genomics, Foster City, California. "This typifies what bioinformatics is all about, to recognize that a method in one context can be transferred to another."

People with these skills are valuable to industry. "Pharma companies are desperately trying to organize and prioritize their drug targets," Sjölander says. "To do this you begin by making inferences about functions that are embodied in sequence."

However, that skill does not come easily. To develop fully the ability to make inferences, one needs to feel comfortable moving back and forth between computational and experimental data. "It's important for people who want to work in this field to have a true basis of understanding in more than one field. In addition to molecular biology, math, statistics and some elements of machine learning would be helpful," she says. "You can't do it with knowledge of just one specialty."

Sjölander thinks that bioinformatics should be a science in its own right. The demand certainly exists. The sheer quantity of data being generated from such a wide variety of sources requires the constant creation of new methods using mathematical or statistical modelling. "There's so much data being generated and we need to have methods that will respond to its particularities."

To increase the supply of computational biologists to deal with the demand being created by the data deluge, Sjölander believes that a cultural change is in order. "Historically, biologists haven't felt comfortable talking to mathematicians and computer scientists, who in turn don't really understand a lot of the issues in biology," she says. "That has to change."

P.W.

it will appoint investigators in computational biology, a departure for an organization that until now has avoided funding research in what it viewed as engineering disciplines. Now, however, it is becoming clear that bio-computation is not only part of HHMI's biomedical mission, but is one of its most critical elements.

Other support is also issuing from the Alfred P. Sloan Foundation, which has recently called for proposals to fund academic units that create professional master's degrees in biology. Traditionally, these degrees have not carried the same weight in biology as in engineering or business, where they are terminal degrees with full professional credentials. A two-year training cycle might help to relieve industry's acute need, where competent support staff, rather than technically savvy principal investigators, is what is needed.

Recognition of need

The NIH and the HHMI, respectively the largest public and private funders of biomedical science in the United States, have both previously acknowledged the need for more training in computational biology. But the money to pay for large-scale programmes has not materialized. Now the two agencies have signalled that this may change.

Last June, an NIH working group recommended a sweeping national programme to promote computational biology, its cornerstone being funding for up to 20 national

centres (see <http://www.nih.gov/about/director/060399.htm>). Although the goals of this Biomedical Information Science and Technology Initiative (BISTI) have not been met, Carol Dahl, NIH assistant to the director of special technologies, remains optimistic.

BISTI will be considered by all NIH institute directors this spring, and money for the programme could be in the budget for 2001. President Bill Clinton's 2001 budget proposal contains more money for large centres rather than new money for additional investigator-initiated grants.

The HHMI has sent similar signals. When Tom Cech was appointed director of the institute, he was quick to single out computational biology as an area in which it might increase its support (see *Nature* **398**, 256; 1999). But, because of the necessary shift towards collaboration and data collection, academia may need to rethink the way such work is rewarded, Cech warns. Tenure might not come easily to a researcher whose credit to a paper is lost in a sea of names. Other systems of acknowledgement, where each person's individual contributions are noted, may become necessary as authorship lists to collaborative papers become longer.

Last winter, the HHMI hosted a symposium to examine how computation and biology could be better merged and funded, from both public and private sources. However, according to an institute spokesman, no new funds have been committed for computational biology.

Potter Wickware

Training

Europe seeks solution to bioinformatics shortfall

London

A shortage of skills in bioinformatics is not limited to the United States. Europe is also feeling the pinch. The need for bioinformaticians to make sense of the data emerging from the Human Genome Project is well known among the industrialists and academics working in the field. It is something that was emphasized repeatedly to Lesley Thompson, who runs the life sciences interface programme of the Engineering and Physical Science Research Council, as she toured the United Kingdom's universities to identify ways in which the physical sciences could help life sciences.

For proteomics, bioinformaticians are, of course, essential, but they are not very easy to find. Both Walter Blackstock, of Glaxo Wellcome, and Gitte Neubauer, of the start-up company CellZome in Heidelberg, say

that when recruiting for their proteomics effort they received only ten applicants for posts in bioinformatics, compared with 100 for biology.

One way to begin to tackle this imbalance, says Graham Cameron, joint head of the European Bioinformatics Institute (EBI) in Cambridge, an outpost of the European Molecular Biology Laboratory, would be to make sure that training in bioinformatics is included in all undergraduate biology courses. "Bioinformatics needs to be so pervasive in biology that it transcends its geeky image," he says.

Another way of addressing the imbalance is by establishing more academic programmes at all levels. Five of the United Kingdom's research councils have banded together to strengthen research in bioinformatics. The Biotechnology and Biological

Mining databases will be crucial in the coming years as researchers try to pin down protein and gene function.

Sciences Research Council (BBSRC), the Engineering and Physical Sciences Research Council (EPSRC), the Medical Research Council (MRC), the National Environment Research Council (NERC) and the Particle Physics and Astronomy Research Council (PPARC) will pool their resources so that three universities will receive £1.25 million (US\$2 million) each over five years. The money will pay for a professorship, technical support and some infrastructure.

Banking on information

The EBI itself is primarily an archiving service. It looks after the gene and protein data that are emerging worldwide from the studies of the genomes of all organisms. Some 20 per cent of the institute's activity is reserved for research methods to improve the presentation and utility of its unique set of databases. Both nucleotides and proteins fall under the institute's remit, with the protein work being a collaboration with the Swiss Institute of Bioinformatics (SIB) in Geneva.

This month the two institutes are releasing a new database that comprises all vertebrate sequences that have a known function. The database is part of the institutes' Human Proteomics Initiative. If someone is seeing a sequence of human protein — derived from a gene identified in the Human Genome Project — but has no idea what the protein does, the researcher will be able to compare the new sequence with the database of vertebrate sequences. If they find a similar sequence or sequences from other species, then the researcher will have been given a clue as to what the function of the new sequence might be in humans.

Data mining

The new database is made up of information already in the public domain — in databases maintained by EBI and SIB — but it is reorganized into what is known as a cluster database. This kind of database structure allows information with differing degrees of commonality to be grouped in clusters. It is not a new tool, but it is this sort of tool that bioinformaticians develop in their efforts to find potentially meaningful correlations: hence the need for both biological and computational knowledge.

Cameron: to succeed, bioinformatics must transcend its geeky image.

The most important protein database that the EBI and SIB maintain — its information was mined for the new vertebrate database — is SWISS-PROT. This has been growing since 1986 and includes sequence data from all organisms, derived primarily from gene sequences.

Curators then ensure that there are no duplicate sequences, and research the literature to find out as much as they can about the function of the sequences. This additional information is attached as annotations to the relevant sequences. The process is time-consuming, but it makes SWISS-PROT a valuable research tool. It is free to academics, but industry pays a licence fee — the money pays for additional curators and researchers.

Aware of the need to support efforts to turn the data emerging from the Human Genome Project into applications, the EBI and SIB have stepped up efforts to annotate the human protein sequences. They are also

looking at ways to provide researchers with a database of protein-protein interactions, something that researchers such as Mathias Uhlen, head of the department of biotechnology and biochemistry at the Royal Institute of Technology in Sweden, believes is essential.

Clearly, mining databases will be crucial in the coming years as researchers try to pin down protein and gene function or spin off companies from their university departments, and as industry looks for scientific insight and checks that there is no prior art to invalidate any patent applications. Says Cameron: "The winners in the next year or so will be those who do the bioinformatics right."

Helen Gavaghan

Further information

- The European Bioinformatics Institute:
<http://www.ebi.ac.uk>
- The Swiss Bioinformatics Institute
<http://www.isb-sib.ch>
- The EPSRC
<http://www.epsrc.ac.uk>
- The BBSRC
<http://www.bbsrc.ac.uk>
- The NERC
<http://www.nerc.ac.uk>
- The MRC
<http://www.mrc.ac.uk>
- The PPARC
<http://www.pparc.ac.uk>

Sequence searching: databases are powerful tools for understanding proteins.