opment although, as Fields points out, that ranges from people who do requirement analysis and do not write software applications, to code writers, to people who do testing and documentation. The ability to function well as a member of a team is critical to software engineering, not just as a skill but as an attitude, says Fields.

There are signs investors are beginning to take notice of these bioinformatics start-ups. Both NetGenics and Structural Bioinformatics have agreed financing deals with venture-capital companies. And what these companies lack in terms of the resources and infrastructure of a large pharmaceutical company, they often make up for by a lively and enterprising work environment. The fact that everybody gets a piece of the action through stock options can also be a draw. Although the world of large pharmaceutical companies is comfortable, it is sometimes "hard to move things quickly and in new directions", says Susan K. Burgess, vice-president of corporate development at Structural Bioinformatics and formerly with Wellcome and Glaxo .

But job seekers would also do well to take a look at what is happening inside some of the pharmaceutical companies.

Although generally slow to move into this area, SmithKline was one of the first companies to recognize the importance of integrating genomic data into the drug-discovery process. In 1993, it entered a $125 million alliance with HGS, which gave it access to the company's vast database of human gene sequences. Searls says the link forced SmithKline into bioinformatics ahead of others and may have been one of the best side-effects of the HGS deal.

Indeed, it was the sheer scale of the bioinformatics enterprise at SmithKline and the company's commitment to academic-style research that lured Searls away from the University of Pennsylvania. SmithKline's bioinformatics group, mainly US-based, now numbers in the high 50s. In addition, the company has recently established a sizeable gene informatics group in the United Kingdom under the direction of Peter Goodfellow, consisting mainly of PhD-level biologists whose primary function is to discover new genes and genetic markers in the databases.

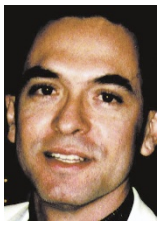
Searls currently has at least six job openings and is keen to recruit a number of senior researchers who he says would typically be academic types capable of running independent research activities.

Bioinformatics is clearly an area where supply has not yet outstripped demand and where people with the right skills can do very well — whether in drug or biotechnology companies or, now, in this new set of 'vendor' companies servicing the pharmaceutical and biotechnology industries. □

Diane Gershon is assistant editor, new technology, of Nature Medicine.

Common language of bioinformatics

Bruno W. S. Sobral



Bruno Sobral

There's no denying that bioinformatics is an exploding field. And as team leader for agricultural and environmental genomics at the National Center for Genome Resources (NCGR), a non-profit bioinformatics organization, I am struck by the lack of qualified people to meet the challenge.

The most effective solution for overcoming this limitation, as vast quantities of genomic data stream in from private industry and the public sector, are partnerships between groups specializing in bioinformatics and those in genetic mapping, physical mapping and genomic sequencing. But before going further, it is necessary to define some terms in the quest for a common language.

With the development of molecular biology over the past couple of decades, ever more powerful tools have been developed to dissect the role of genes in a wide assortment of traits, such as diseases of various organisms. In addition, new information has helped in the understanding of inherited traits and their influence on behaviour. More recently, as a result of computational and engineering advances engendered by the Human Genome Project, the field of genomics has emerged. Genomics has two major components: DNA sequencing in the molecular genetics laboratory; and bioinformatics, another new term which was spawned by genomics. Thus, genomics is a combination of complete nucleotide sequences for any organism and the information tools required for analysis of data. Bioinformatics is defined as the computational systems used to collect, store and analyse biological information. These include software systems that take in DNA sequence data, database systems that store data, and software systems that analyse stored data.

One of the key challenges in bioinformatics as a field is to move smoothly through the transitional period of collections of incompatible components to integrated systems. In the next few years, people working in bioinformatics will be tackling issues of interface homogeneity, development of scaleable software toolkits for data analysis, inter-database connectivity, consistency of terminology and long-term funding of public repositories of information.

Universities have not generally played a major part in bioinformatics because they have not had the resources. As genome projects broaden their scope from the human genome to model systems and agricultural and food genomes, universities need effective bioinformatics support. One solution is to enter partnerships with groups already experienced in bioinformatics, especially within the public sector. In this manner, DNA sequencing/physical mapping groups at universities can quickly develop large datasets without having to create their own bioinformatics infrastructure. NCGR, for example, has recently established a partnership with New Mexico State University to develop a national biotechnology information facility (NBIF — http://nbif.org/), a five-year, $8.5-million partnership. NBIF is developing a plant-specific metabolic pathways database while providing training and bioinformatics support for graduate students and postdoctoral fellows. NCGR's strategy is to provide similar bioinformatics support to leading agricultural research institutions.

Through such partnerships, collaborative development of human resources in bioinformatics can become reality. Not only will this provide much-needed bioinformatics expertise in universities; it will also provide the private sector with individuals trained to make use of genomic information. As genetics and biology enter the twenty-first century, it is clear that genomics is changing the way biological research is done.

Public information and biological reagent repositories are a decentralizing and democratizing force in research, much as the Internet was and is for computers. Eventually, any scientist may have direct and rapid access to information and reagents without needing to create and maintain a large-scale operation. I hope that this will also mean that scientists will be rewarded more for creativity and less for their fund-raising capabilities. □

Bruno W. S. Sobral is at the Genomics National Center for Genome Resources (NCGR), Santa Fe, New Mexico 87505, USA. e-mail: bws@ncgr.org.

Crossover in interest and skill

Brendan Horton

Barry Robson and Roger Pain wrote one of the first papers discussing how to store protein sequences and recover information (*Nature* **227**, 62–63; 1970). As he has done throughout his career, Robson prefers to erase the artificial lines that separate and define disciplines, describing himself on his *curriculum vitae* as "an experimental medical scientist and a computational chemical physicist". This mixture of 'hard'

Barry Robson

and 'soft' science gives him much enjoyment in his work.

Because of mild dyslexia, Robson was channelled away from maths and theoretical sciences by the age of 16 in the UK education system. Unrecognized in his youth, Robson's dyslexia affected his ability to spell, make left and right spatial judgements and do arithmetic. In particular, he notes that at that time it was not appreciated that mathematics and programming are not the same as arithmetic.

It was not until after Robson had been trained as an experimental medical scientist that he discovered he could do theoretical science after all. In his experience, students were channelled into subjects depending on their early academic performance, as if the proper intellectual hierarchy dictated mathematics for the most able, followed by physics, chemistry, biology, psychology and finally sociology. Robson's first self-directed step to theoretical biology came when, at 18, he developed mathematical functions for modelling molecules in three dimensions.

"A lot of protein modelling code sprang from those early routines, which were first implemented to draw molecules with the help of a mechanical calculator," he says. Realizing that the mathematics would save him work, he began "to collect maths tricks like a magpie collects shiny objects". He remains, however, envious of those who have been formally trained and have a full repertoire of mathematics skills.

Yet it was the lack of formal training that benefited his work in a series of information theory papers with Pain, Eicichi Suzuki and others from 1970 to 1976, leading to the Garnier–Osguthorpe–Robson (GOR) secondary-structure prediction method. To deal with very sparse X-ray data from 1968–1972, Robson used bayesian statistical methods, which were largely ignored at the time.

"There was in fact a very bad reaction to anyone using such techniques, and except in very rare and very advanced courses, they were not taught at all." They were regarded as 'subjectivist', he says: "I would never have chosen that worthy approach if I had been formally trained in the approved statistical philosophy." Today, he says, bayesian methods are the mainstream of bioinformatics.

It was these early successes, perhaps, that strengthened Robson's belief that lines drawn between fields were artificial and limiting. The 1980s were revolutionary, bringing the ability to engineer proteins through genetic engineering; large increases in requisite data by breakthroughs in NMR, crystallography and monoclonal antibodies; and supercomputers to run these programs, as well as cheap, networked computing to process and share data locally.

Robson's involvement in commercial and semi-commercial ventures led to the formation of the Epsitron peptide and protein engineering research unit at Manchester University, and the Proteus group of pharmaceutical companies. He stayed as reader in biochemistry at Manchester while cofounding and becoming scientific director of Epsitron and Proteus. This year he became a visiting researcher at Stanford, as well as a principal scientist at MDL, where he helps to acquire and develop new tools, trains staff and customers, assists marketing and sales, and advises on general strategy with MDL's Bioinformatics Workbench and Molecular Informatics' BioMerge systems.

Robson participates in a course at Stanford which is open to those from other departments, including medical students. He says that programmes like Stanford's are mainly concerned with preparing the next generation of bioinformaticists and tool providers through research (see Internet table). Robson is unaware of any example where someone from industry can take a short course in an academic institution and then return to industry, but notes that the master's programme at the University of Manchester is mainly populated by industry researchers, indicating that there is such a need.

There should be rich opportunities in this field, but when will it reach saturation? Will all the hyperbole result in the overproduction of bioinformaticists similar to the glut of molecular biologists that resulted from unchecked growth in that field in the 1980s and 1990s? The expertise shortage is a real bottleneck, but there are ways in which computing and artificial intelligence (AI) can help, says Robson. These include building problem-solving environments to make best use of a limited number of experts; automating these to capture expertise (moving towards 'expert systems'); adapting them to smarter, more human-independent AI systems, still further reducing the dependence on experts; and adapting 'expert' systems as teaching machines.

Ultimately, advances of this kind will limit and eliminate jobs. The National Academy of Sciences' office of scientific and engineering personnel met earlier this month to examine

Table 1 Internet resources for bioinformatics

Organization	http://...
Affymetrix	www.affymetrix.com
Human Genome Sciences	careers.hgsi.com
Incyte Pharmaceuticals	www.incyte.com
Molecular Informatics	www.molinfo.com
MDL	www.mdli.com
Molecular Applications Group	www.mag.com
National Center for Biotechnology Information	www.ncbi.nlm.nih.gov
National Center for Genome Resources	www.ncgr.org
National Center for Supercomputing Applications	www.ncsa.uiuc.edu/Apps/CB
NetGenics	www.netgenics.com
Merck	www.merck.com
Pangea Systems	www.PangeaSystems.com/0_home/index.htm
Pfizer	www.pfizer.com
SmithKline Beecham	www.bioinformatics.com
Structural Bioinformatics	www.strubix.com
The Institute for Genomic Research	www.tigr.org
Stanford Groups	
Sam Karlin Group (mathematics)	genomic.stanford.edu
Michael Levitt (structural biology)	hyper.stanford.edu/faculty.html
Section on Molecular Informatics – (acad. progs)	www-SMI.Stanford.EDU/smi/academics
Douglas Brutlag (bioinformatics group)	motif.stanford.edu
Section on Molecular Informatics	camis.stanford.edu
Program in Molecular and Genetic Medicine	mgm.stanford.edu/br/
General	
Weizmann Institute Genome and Bioinformatics	bioinfo.weizmann.ac.il/
Bioinformatics Services	www.bioinformaticsservices.com/
Bioinformatics at University of Wisconsin	www.bocklabs.wisc.edu/Home.html
BioInform	www.bioinform.com/
Biocomputing	www.cryst.bbk.ac.uk/BCD/bcdgloss.html
BioComputing Hypertext Coursebook	merlin.mbcr.bcm.tmc.edu:8001/bcd/Curric/welcome.html
Genomics	www.phrma.org/genomics/index.html
KEGG Encyclopedia of Genes and Genomes	ep3000.scl.genome.ad.jp/kegg
The CMS Biology Resource	www.unl.edu/stc-95/ResTools/BioTools/biotools4.html
The DOE Biology Information Center	www.er.doe.gov/production/oher/bioinfo_center.html
United States Department of Agriculture (Budget)	www.usda.gov/agency/obpa/Home-Page/obpa.html
Agricultural/Biotechnology Resources in Education	
Colorado State University, Deptartment of Entomology	www.colostate.edu/Depts/Entomology/ent.htm
University of Arizona, Department of Entomology	ag.arizona.edu/ENTO/entohome.html
Iowa State University, Department of Agriculture	www.ag.iastate.edu/
University of Maryland, Center for Agricultural Biotechnology	www.umbi.umd.edu/~cab/index.html

methodological problems related to forecasting supply and demand for scientists and engineers and the use of such forecasts in policy making, as well to finalize the agenda for a workshop on these issues to be held early next year. According to the NAS, "Conflicting assessments that have emerged from recent analytical efforts have resulted in a considerable amount of confusion about the future state of labor market conditions for scientists and engineers." □

Brendan Horton is in Nature's *Washington Office.*

Choices and challenges

Potter Wickware

Computers have changed biology forever, even if most biologists don't yet realize it, says Michael Levitt, a structural biologist at Stanford University and the founder of Molecular Applications Group (MAG), in Palo Alto, California. Already, drug discovery is driven by the need to apply powerful computers to voluminous data sets, and the trend, he says, is certain to extend into all other disciplines in biology.

Chris Lee, Levitt's former graduate student and co-founder of MAG, agrees, noting that most biologists today use computers only in the most elementary way as a typewriter and graph-paper substitute. "Bioinformatics is really going to surge when biologists realize that there's a lot of value, and a lot of new insights, in being able to work across large amounts of data that they and all the other scientists in the world have produced," says Lee.

Levitt began using computers to solve problems in protein folding when working under John Kendrew, Max Perutz, Francis Crick and other eminent molecular biologists in the "golden age" of the 1960s at the Laboratory for Molecular Biology at Cambridge, United Kingdom. Today the 7lb laptop he carries in his backpack has more than a thousand times the computing power, at less than a thousandth of the cost, of the punch-card behemoths of 30 years ago.

Accompanying the relentless increase in computing power is a breathtaking expansion of biological data from the human and other organism genome-sequencing projects. Complementary information from the pharmaceutical chemistry, neuroscience, microbiology, immunology, clinical trials, toxicology, teratology, epidemiology and other disciplines waits to be integrated with the genetic and structural data. There is no way to obtain a global view of all this information, to establish links between disparate fields of knowledge, without the computer.

Myra Williams, MAG's new president, has a PhD in biophysics from Yale and was hired this summer from Glaxo Wellcome to launch the company's GeneMine Pro suite of bioinformatics tools. She observes that rates of data acquisition, far from levelling off, are accelerating. Soon innovations like Affymetrix's high-density oligonucleotide array microchip will come online, generating terabytes ('terror-bytes') of new sequence information. How scientists navigate this ocean of biological information will be crucial.

"To be effective," Williams says, "bioinformatics tools must not merely automate data retrieval, but give researchers the information in usable form, through clustering, filtering, analysis and visualization, allowing them to perceive insights which might well have eluded them had they attempted to process the information manually."

The bioinformatics capabilities of MAG's GeneMine Pro program are built around its Discovery Engine, an automated Web browser which retrieves biological information in 22 categories from servers worldwide. After processing, the information is presented at the user interface, where it can be visualized in the context of sequence alignments and three-dimensional protein structures, or read as text. Large pharmaceutical companies, challenged by expiring patents and the high cost and slow pace of conventional drug development, are now the main source of sustenance for bioinformatics. The fact that genomics and bioinformatics are creatures of big industrial research environments inevitably leads to a blurring of distinctions between academic and industrial science.

Despite a strong and growing demand for bioinformaticists, there are few established training centres, perhaps 20 in the world, estimates Levitt. The field is still defining itself, and those who do have formal training are quickly snapped up by industry on hefty pay scales, leaving a deficit in the numbers of those available to train the next generation.

Nevertheless, Lee believes that the candidate who, on his or her own initiative, "can demonstrate the ability to cross over and generate results — not even necessarily original ones — is the one who will capture the recruiter's attention."

Levitt adds that the shortage of formal training slots coincides with exceptional opportunities for self-learning: with an Internet connection and inexpensive computer, one can download all the databases, programs and papers needed to undertake an original research project. "Spend two days looking at the results and thinking about what you don't understand. Use e-mail to contact someone who does, and ask, 'what should I be doing'?" he recommends. Prize-winning discoveries can be made in this way. "The problems are so difficult, and there is so much to be analysed, that the boom will not go away. It's a great time to be getting in. There's a wonderful lightness to the field." □

Potter Wickware is a science writer in Oakland, California, USA. e-mail: wick@netcom.com

Running to catch up in Europe

Helen Gavaghan

Across Europe, the story is the same. Demand for those skilled in bioinformatics exceeds supply. Like biochemistry and biophysics before it, bioinformatics is crushing the barriers between traditional academic fields, and demanding flexibility and a new way of thinking from its adherents.

Computational biology has meant different things to different people. Not too long ago, says Hans Prydz of the University of Oslo's Biotechnology Centre, it meant handling NMR data or analysing Doppler echograms. Now renamed bioinformatics, it means looking for patterns in DNA and RNA, predicting protein structure, modelling proteins and mining massive databases that continue to grow. When the DNA database run by the European Bioinformatics Institute (EBI) was first set up, it contained 700,000 nucleotides: now there are more than a billion.

Driven by the scientific and commercial importance of bioinformatics in genomics and drug discovery and development, governments, universities and industry are responding with varying degrees of vigour and success to the skills shortage and are seeking ways to cross the boundaries between disciplines as diverse as engineering, physics, mathematics, computer science, statistics, protein chemistry, genetics and molecular biology.

At European level, the EBI, based near Cambridge (United Kingdom), is funded to the sum of about DM 9 million ($5 million) by members of the European Union and Israel via their contributions to the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Contributions from the pharmaceutical and agricultural industries roughly double the institute's income. The EBI, an offshoot of EMBL, develops tools for bioinformatics, seeks innovative ways to apply the tools, and runs training courses for academics and industrialists. Initiatives with industry include the Industry Affiliates Initiative, which helps small and medium-sized companies to identify and apply new techniques; the BioTitan Project, running nodes to enable faster access to databases; and the Biostandards project, funded jointly by industry and the European Union for promoting and developing standards.

National initiatives also exist, particularly in the United Kingdom and Germany. Says Andrew Lyall, responsible for bioinformatics at Glaxo Wellcome, "I think the UK is in pretty good shape." There are two government-financed initiatives in the United