Labs and companies seek their niches as work continues after the draft

Specialization may be the key to success as divisions between big and small genomics centres continue to grow, say Potter Wickware and Paul Smaglik.

Now that the human genome has been sequenced in draft form, will there be less work in sequencing centres? Far from it, says Francis Collins, director of the US National Human Genome Research Institute (NHGRI). The biggest sequencing labs will switch from human to a list of other organisms. Smaller labs will focus on finishing and annotation. The output of both will crank the demand for bioinformaticians even higher. The common denominator? A strong computational background, lab heads say.

TIME PIX/REX

Get your motor running: there's plenty of work for self-starters, says Francis Collins.

Opportunities, big and small

The big-science model, in which large sequencing centres dominate production of sequence data, will continue at least in the near term, even though large-scale sequencing of the human genome is now drawing to a close. Those centres — at Washington University in St Louis; the Sanger Centre in Hinxton, Cambridgeshire; the Whitehead Institute at the Massachusetts Institute of Technology; Stanford University; and the US Department of Energy's Joint Genome Institute (JGI) in Walnut Creek, California (a consortium of university and national lab sequencing centres) — will be turning their high-throughput capacities towards other organisms.

A vast supply of other genomes beckons, representing a mostly untapped resource with huge potential payoffs for agriculture, environmental remediation and medicine. Dan Rokhsar, director of bioinformatics at the JGI, says that advances in sequencing and assembly mean 95% of a typical microbial genome can now be sequenced in a day and a half. The JGI sequenced 15 last October, and proposes to do three times that number each year, in a set of 'microbial marathons'.

But the big centres will not eclipse the smaller labs — as long as they do not try to compete in the high-throughput sequencing game. "The smaller centres will need to develop themes, as it is unlikely that they can match the largest centres on a pure cost-per-base-pair basis," Collins says.

'Finishing', or filling in the many gaps in the draft version of the human genome, is a growth industry, he says. Only about a quarter of the genome can be considered finished, and the publicly funded project is committed to completing the rest by the end of 2003. Maynard Olson's University of Washington lab is actively engaged in closing gaps in the human genome that result from

Following the growth of data

In the early 1980s David Haussler worked with pattern recognition on some of the earliest available DNA, the phage ϕX174.

"A single seamless career working with viruses and ribosomal binding sites would have been possible, but the lack of data was frustrating," he says. It took an inordinate amount of bench work to get more, so he felt it was more logical to go off and develop fundamentals in statistics and machine analysis methods.

"If the data had been there I would have stayed with bioinformatics, because the early work we were doing with promoters was so exciting," recalls Haussler, who is now director of the Center for Biomolecular Science and Engineering at the University of California at Santa Cruz.

Later, of course, the problem became, if anything, too many rather than too few data. When he returned to sequence analysis, Hidden Markov Modelling — a robust set of pattern recognition methods that was developed by his group — was in place to help deal with the task.

"We're just beginning to get to know the genes. There's a huge amount to be discovered as we push on to link genes with disease and basic molecular biology," Haussler says. P. W.

the high-throughput approach. But the sheer size and difficulty of the task means ample opportunities for latecomers.

Other labs may specialize further, finishing parts that resist being sequenced by machines. Regions near telomeres and centromeres may be well suited to small labs, as these stretches of DNA require time-consuming manual techniques to decipher, says David Haussler, director of the Center for Biomolecular Science and Engineering at the University of California at Santa Cruz.

Funding abundance

Specialized labs not directly associated with sequencing will also emerge. The NHGRI received a 15% increase for the fiscal year 2001 (which started last October), increasing its budget to $382 million. Most of the increase will go towards the new Centers of Excellence in Genomic Science programme.

This will fund multidisciplinary centres focusing on genome-wide analysis and technology development. Although applications will not be reviewed until May, the programme is likely to create several multimillion-dollar centres for analysing gene function and gene expression; studying population genetics and sequence variation; and performing comparative genomics and complex trait analysis, among other things. Most will require scientists with sophisticated computational skills.

Other agencies also benefit from federal largesse. The Department of Energy allocated $117 million to genomics in 2001 — up 14% from last year — and the National Science Foundation, which got a rise of $500 million this year to $4.4 billion, is supporting genomics-related research. Funding for plant research, still far behind human-related work, is also on the rise, with a current US Department of Agriculture genomics budget of $85 million, up from $79 million last year.

The states are being generous, too. California's new Institute for Bioengineering, Biotechnology and Quantitative Biomedical Research, announced in December, will receive $100 million in public and $200 million in private donations. It will be based at the University of California at San Francisco's new Mission Bay campus, with major components at Santa Cruz and Berkeley, and will tie into Berkeley's $500 million Health Sciences Initiative, which seeks to advance health science research through multidisciplinary collaborations.

Bioinformatics needs

All these initiatives will need people with the computational skills to analyse increasingly complex data sets. But bioinformaticians, like smaller sequencing centres, may find that it pays to specialize.

Mark Gerstein, who leads a bioinformatics group at Yale University, sees opportunities for the independent researcher to carve a niche in a particular type of promoter or structural motif, then carry out the analysis on the entire genome.

"It doesn't require a huge amount of apparatus, and the full annotation of the genome is such a huge task that there's room for everyone," he says.

Genomics and bioinformatics companies in North America

Company	Activity	URL
Affymetrix*	Expression chips, analysis services;	www.affymetrix.com
AP Biotech*	Comprehensive drug development	www.apbiotech.com
Base4	Pharmatrix database; pharma and IT consulting	www.basefour.com
Caprion Pharmaceuticals	High-throughput cell maps, proteomics	www.caprion.com
Celera*	Databases, genotyping, target discovery	www.celera.com
Cellomics	Whole-cell assay chip	www.cellomics.com
CuraGen*	SNPs, expression, low abundance genes, drug-induced changes in gene expression	www.curagen.com
Deltagen	Functional genomics	www.deltagen.com
Digital Gene Technologies	TOGA system correlates expression with anatomy; also has LIMS	www.dgt.com
DNA Sciences (was Kiva Genetics)	Gene Trust patient database	www.dna.com
DoubleTwist	Web-based softbot "agents"	www.doubletwist.com
Exelixis	Pathway and target finder software	www.exelixis.com
Gemini Genomics*	Genotyping, SNPs	www.gemini-genomics.com
First Genetic Trust	Database; patient information encryption	www.firstgenetic.net
Genaissance	Population genomics, "personalized medicine"	www.genaissance.com
Gene Logic*	Expression analysis	www.genelogic.com
Genicon	RLS method of labelling & detecting biomaterials	www.geniconsciences.com
Genomica*	Software for family studies, epidemiology	www.genomica.com
Genomics Collaborative	SNP genotyping	www.getdna.com
Genomics Institute (a division of Novartis)	Comprehensive genomics, proteomics; mouse genetics	www.gnf.org
Genomic Solutions	Biochips & arrays	www.genomicsolutions.com
Human Genome Sciences*	Sequencing, expression analysis, proteomics, preclinical & clinical testing	www.hgsi.com
Hyseq	Genomics, high-throughput sequencing	www.hyseq.com
IBM*	"Blue Gene" supercomputer, computational biology group	www.research.ibm.com/topics/serious/bio/
Integrated Genomics	Microbial genomes, pathways	www.integratedgenomics.com
Incyte*	Databases, software	www.incyte.com
Informax*	Vector NTI software	www.informax.com
LabBook	Genomic XML browser	www.labbook.com
Lexicon Genetics	Biochips & arrays; OmniBank knockout mouse clones	www.lexgen.com
Lynx Therapeutics	SNP genotyping; MegaClone expression method	www.lynxgen.com
Maxygen	"Gene breeding" directed evolution compounds	www.maxygen.com
Motorola BioChip	Expression arrays	www.motorola.com/biochipsystems/
Molecular Simulations	Software	www.msi.com
Myriad Genetics*	Sequencing, proteomics	www.myriad.com
Nanogen*	Biochips & arrays	www.nanogen.com
NetGenics*	Software	www.netgenics.com
Orchid Bioscience	SNP genotyping	www.orchid.com
Paradigm Genetics	SNP genotyping	www.paragen.com
Phylos	"HIP" protein chip	www.phylos.com
Promega	SNP genotyping	www.promega.com
Prospect Genomics	Structure prediction	www.prospectgenomics.com
Proteome (division of Incyte)	Worm, yeast databases	www.proteome.com
Qiagen Genomics	SNP scoring	www.qiagen.com
Rosetta Inpharmatics*	Expression analysis software	www.rosetta.com
Senomyx	Smell & taste genes	www.senomyx.com
Sequenom*	SNP analysis	www.sequenom.com
Silicon Genetics	Array analysis software	www.sigenetics.com
Spotfire	Analysis "siftware"	www.spotfire.com
Syrrx	High-throughput protein structure prediction	www.syrrx.com
Structural Bioinformatics	Patient-specific structural variants	www.strubix.com
Structural GenomiX	Proteomics, structure	www.stromix.com
Zyomyx	Protein biochips	www.zyomyx.com

(* = public company)

How many people practise the hard-to-define trade is hard to pinpoint, but a good estimate is attendance at the annual Intelligent Systems for Molecular Biology conference, which began with just 200 in 1993. The figures started zooming up three years ago,

rising to 1,300 last year. Worldwide, the number is probably two or three times greater. Bioinformatics may be a small field, but few areas can boast such a steep rate of gain.

The jobs are spread across industry, academia and government labs. Enterprises ranging in size from Fortune 100 corporations to tiny boutique companies all have the 'Help Wanted' sign up. At least 50 companies devoted to genomics, proteomics or bioinformatics are doing business in North America (see table). Many corporations have their own groups and list openings on their web pages; collective listings are also found on bulletin boards such as http://scijobs.org and http://www.genomeweb.com.

Major pharmaceutical companies are a rich source of bioinformatics jobs. Terry Gaasterland, director of the Laboratory of Computational Genomics at Rockefeller University in New York, thinks they lead in bioinformatics analysis. They have far more data than anyone else, and because of the many specific tasks they address they are ahead with their methods as well. "For example, pharmas have much better methods of dealing with Affymetrix expression data than the academic sector," Gaasterland says.

Reaching this stage of the human genome project has created ample opportunities for people adept at adapting — and adapting to — new technologies and computational approaches, says Collins. "If you're bright, motivated, creative, and can set up automated PCR and write a Perl script, there's a promising future for you in genomics." ■

Potter Wickware is a science writer in San Fransisco, Paul Smaglik is editor of *Nature*jobs.

SNP consortium ▸ **http://snp.cshl.org**

Centers of Excellence in Genomic Science ▸ **http://www.nhgri.nih.gov/Grant_info/Funding/Research/CEGS_synopsis.html**

Current role suggests the shape of future work opportunities

Mappers, cloners, sequencers, finishers and annotators — each of the five major sequencing centres and the nine responsible for smaller portions of the genome employ such staff. How do their tasks fit together? And, now that the project is racing to the finishing stages, what will people at centres worldwide turn their attention to? One way to get an idea is to look at their role in the human genome project up to now.

Before they could proceed into the uncharted territory of human high-throughput sequencing, geneticists needed maps. First, scientists created a genetic map, which plotted the approximate location on each chromosome of genetic features such as a gene or particular base sequence, determined by studying genetic material from small populations of families. Then they sketched a physical map, using molecular biology to determine the location of specific sequences.

Combining, then refining, those maps has been critical to the production of the draft human genome published today. By locating a number of DNA sequences on the chromosomes, biologists provided themselves with a framework on which they could place sequence data. Without maps, accurately positioning DNA sequences would have been akin to assembling a 3-billion-piece jigsaw with no guiding picture, using pieces of jigsaw characterised by only four shapes. Even the private sequencing venture run by Celera Genomics has turned to the publicly published maps as a guide to assembling their sequence.

SPL

Computers will play an increasing role in biology.

Mapping jobs at the Sanger Centre, in Hinxton, Cambridgeshire, are less in demand now that the genome has reached draft quality. At the height of the human mapping efforts, there were some 60 to 70 dedicated mappers. Now that the Sanger has turned to sequencing pathogens and smaller model organisms such as zebrafish, the centre only needs 18 to 20 mappers. They are focusing on verifying the draft and finished version of the human genome, and on learning as much as possible from its raw data. Panos Deloukas, a project leader and one of the key players in producing the physical map of the human genome, is studying the occurrence of disease in three different populations (African Americans, Asians and Caucasians) and matching disease phenotypes to genetic mutations.

The kind of work Deloukas does requires a PhD and experience as a postdoc. Universities, institutes and industry are finding that candidates with these qualifications are in short supply. Says Michael Ashburner, joint head of the European Bioinformatics Institute (EBI): "There are very few good people at a senior level, and there are very good posts we cannot fill." Yet it is not just those with PhDs who find work at sequencing centres.

Clone rangers

Preparation of the draft published today was generated using cloned contigs. In this approach, the DNA is fragmented into sequences that contain markers from the published genetic and physical maps, and inserted into bacteria. Each clone is then randomly broken into smaller segments which are sent to the shotgun teams for sequencing, then to finishing and annotation. Finally, the finished contigs are placed on the master map in the position defined by their markers.

The range of qualifications needed for shotgun sequencing, finishing and analysis vary considerably. Bev Mortimore, a shotgun team leader, was one of the original 17 staff to accompany John Sulston's team when it moved to the Genome Campus at Hinxton (the Sanger now employs 570 people). She says that the sequencing work can be repetitive and she prefers not to hire graduates who find the routine tedious. It is suited to intellectually curious school-leavers (perhaps with GCSEs, perhaps 'A' levels) who learn on the job.

Finishing touch

Once each fragment is sequenced, it is reassembled by running computer programs that seek out areas of overlap. But there are always gaps where it has not been possible to identify the sequence correctly and find areas ▸

Who makes the best bioinformaticians?

Bioinformatics careers can be divided into two paths: developing software, and using it. The field, catalysed by the rapid accumulation of genomic data, has attracted attention as a salvation for jobs in biology. But that sentiment may not provide an accurate assessment of job opportunities, at least for career prospects on both paths. For example, InforMax, one of the largest bioinformatics companies in the United States, generally doesn't hire biologists-turned-programmers, says Alex Titomirov, chairman and chief executive officer of the company, based in North Bethesda, Maryland.

InforMax has about 95 programmers, almost all of whom come from a maths, physics or computer-science background. Titomirov says it is "much easier" to teach people with those skills about biology than to teach biologists how to code well. However, as the company turns to developing software to handle functional genomics and protein data, it may draw on more biologists to help design new software modules. **P.S.**