Hacking the genome

Don't be surprised at the strength of the open source ethic in the human genome project and in the coming bioinformatics explosion, says **Damian Counsell**: science has always been about public collaboration

The planet is inhabited by elaborate, water-rich organisms of various sizes, every one built to a tight specification. But even the most sophisticated of these creatures is merely a container. The inhabitants of this world are soft, fleshy bags. Inside these vessels are the true masters: stored programs of massive complexity. These encoded intelligences influence their hosts' appearance, the time of their deaths and much of their behaviour.

Recently some of the beings have become aware that they are enslaved. They may soon have the technology to escape from internal domination. However, from our privileged viewpoint outside this ecosystem, we can see an obstacle. Some of the slave organisms are witholding from their fellows the information they all need to escape. Restricting this information may give their rulers, the selfish replicators, the upper hand. We observe the alien slaves as their hopes for freedom hang in the balance.

If you haven't guessed yet, the planet is Earth and we, *Homo sapiens sapiens*, are the 'slaves'. To quote Steve Jones, Professor of Genetics at University College London - presumably writing of a relatively rich population like the British – "most people die because of the genes they carry." In case you haven't noticed a few months' worth of headlines, the 'good' draft database of those genes, the human genome, is out.

This final draft is pretty small by modern storage standards: three million MB – that's megabases (2-bit) not megabytes (8-bit) – of our own specifications. What is amazing is that this library of biological information is available for download to your own PC and can be browsed for free from the comfort of your own home using the Human Genome Browser. To cap this, in the process of bringing home the genome, researchers have given brought a new science to maturity.

Bioinformatics is born

Even if you are familiar with the parents – computing and biology – you shouldn't be ashamed if you find the work of their offspring – bioinformatics – incomprehensible. The text of human DNA makes past winners of obfuscated Perl competitions look like Plain English Awardees. Some of the biggest brains in science are working on an interpretation even as I write. For that job most of them are using the same sorts of open source tools and open source operating systems - including Linux - that they applied to liberating our genomes in the first place.

The human genome is the software that programs the building and maintenance of your body. Your body's 'listing' of the code is written in a language of just four letters, four different chemical modules. The names of these chemicals are usually shortened to their initials: A, C, G and T. When you look at an image of DNA, the double helical 'staircase' stuff of your genetic material, complementary pairs of these four letters form the 'stairs'.

DNA is supercoiled and packed with proteins into the complex structures we call chromosomes. If a genome is a library of code, each chromosome can be thought of as a shelf full of books of program printouts. One of the most amazing discoveries of the genome project has been that most of the pages in the human library are nonsense. Less than five per cent of all the code can be shown to make something tangible. These biological algorithms describe, for example, the haemoglobin that carries oxygen in your blood (and colours it red), the keratin that makes up your skin and hair and even the receptors in your brain that collect the message that you are bleeding.

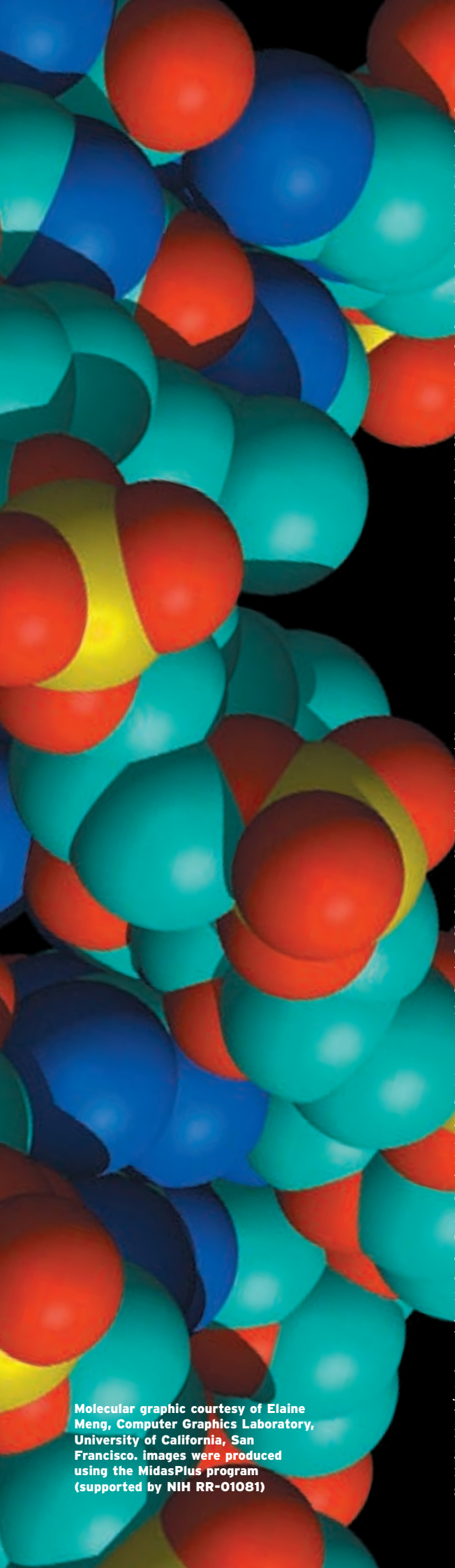
Why are we so interested in the few pages that actually make sense? One reason is inherited genetic disease. If one letter in the copies of your 'book' for haemoglobin is wrong, for example, then, from early childhood on you can lurch into excruciating sickle-cell disease 'crises', caused by the distortion and destruction of your blood cells by the products of your own genes. Disorders like this derive from errors in the copying or repair of the books of life or even of the 'shelves' - people with Down's Syndrome carry an extra copy of chromosome 21.

It's not just medical research funds and universities who are interested in the contents of the library – the drug companies see massive potential too. If they can identify, say, a gene that codes for an important component of the brain they might be able to develop drugs which bind to it and alter the way it works in whole humans, perhaps to treat schizophrenia or depression. More controversially, if a company patents its knowledge of that gene, it can demand payment from anyone else who targets that gene product to treat the same illnesses. The genome is a matter of life, death and big money.

In the beginning was The Tape

There's nothing new about the open source model; science has always been about public collaboration, even between rivals. Anyone who has ever been assigned a partner for teaching practicals will understand that the participants in this sharing of data are not always entirely willing, of course.

Sharing used to be the rule in computer science too. In the beginning was The Tape. And The Tape was distributed to anyone who needed it. The users contributed their fixes back to the programmers and everyone benefitted. With the advent of the microcomputer came the earliest suggestions of a mass market. With the possibility of serious money came serious change. Many people with their roots in that



Molecular graphic courtesy of Elaine Meng, Computer Graphics Laboratory, University of California, San Francisco. images were produced using the MidasPlus program (supported by NIH RR-01081)

early telecoms, academic computing and hobbyist world felt payment for their hard work was due from their peers – including a certain William H. Gates III. Perhaps this was more about pride than cash. Perhaps the programmers no longer felt that the users were truly their peers. Such disgruntled quasi-hackers and their employers felt the best way to ensure their just reward was to close the source of their programs and charge for the executable. This was a 'business model' that was to prove extremely lucrative. Some say this is a model with strong parallels in biotechnology.

Many scientists don't like this model. Science works because you can only publish experiments that other people can do too. If there are any mystical tools needed to perform your tricks you have to be prepared to distribute them to fellow professionals. The white-coated conjurors of science have to hand over their wands to anyone in the magic circle who asks. This is crucial; if an experiment is not repeatable than it might as well not have happened - as the 'discoverers' of cold fusion might tell you.

In biology the classic openly distributed tools are antibodies, exquisitely specific biological molecules which can be used to recognise and label parts of living things. Interestingly biotech companies still manage to make money from selling antibodies in bulk, despite their relatively free circulation among scientists.

Nowadays, the open tools of science include software like Linux. Even when essential programs are 'closed', like the Staden programs used since the early days by single labs sequencing genes, they are distributed at cost to other academics – who identify bugs and suggest features.

A cunning plan

For years scientists had been mapping and sequencing single genes in small groups storing, assembling and manipulating their data on single workstations running programs like 'Staden'. Often these investigators were trying to find out whereabouts and on which chromosome the gene for a particular disease was hiding away. This was a long and difficult process and it was rarely in any one group's interest or power to map and sequence huge regions of the genome – especially as much DNA turned out to be 'rubbish'.

From the mid '80s, a small, ambitious group of scientists, including James Watson (one of the discoverers of the structure of DNA) believed that having the whole map of the human genome at much higher resolution was not only possible, but would massively simplify gene hunting, increase the power of gene research and reduce the duplication of effort.

Not only would scientists be able to look up their own gene of interest, but they would immediately be able to find the relationship between it and all the others. This is very important; neighbouring genes are far more likely to be inherited together for example. Just like Linux, the combined effort of geeks worldwide would add up to something much greater than the work of isolated geeks.

The HuGeP, as it is affectionately known among researchers, was carried out at 16 centres around the world, including substantial proportions at the Sanger Centre on the 'Genome Campus' near Cambridge, UK, the Whitehead Institute and Washington University in the US and Keio University, Tokyo, Japan.

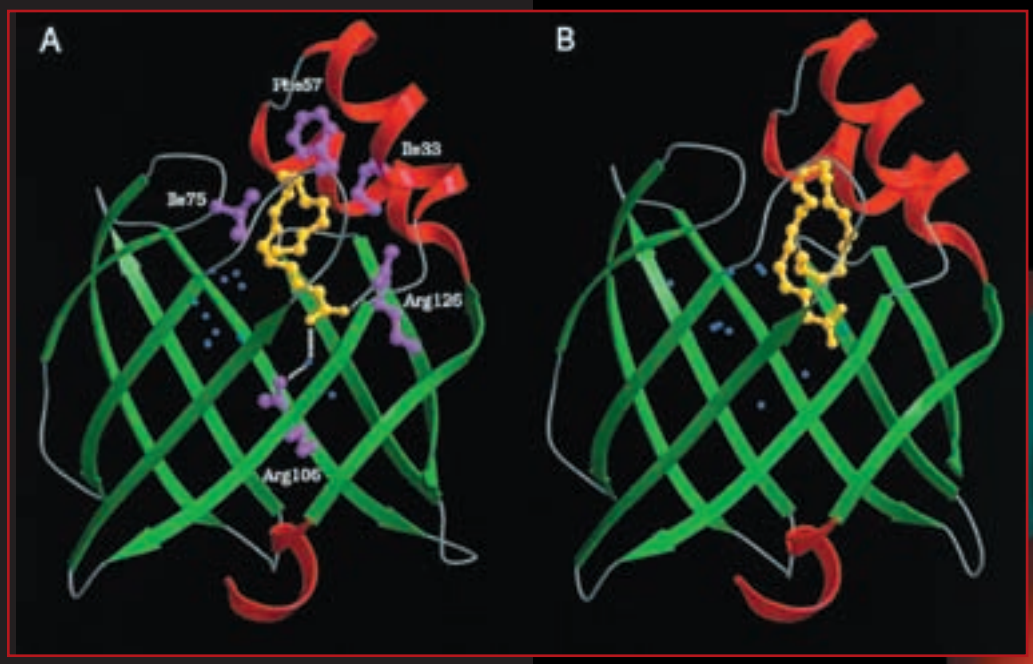
The human genome is actually one of several different species' genomes you can download to your hard disk today. The previous projects on various bugs and simple animals had already given us some idea of approaches to reading a genome, but on a much smaller scale.

You don't just put some cells into a machine, crank the handle and read off the code. A genome has to be assembled at various levels, like building a gigantic jigsaw. As you might separate out the sky from the sea, from the people and from the buildings, the 23 kinds of chromosome found in humans were parcelled out to various groups of scientists. Further, the various groups had to break their designated chromosomes into chunks small enough to be stuffed into an unsuspecting yeast or bacteria. The cells of these creatures would then be hijacked to make trillions of copies of the human genes.

Putting it all together

This wasn't smallest level of the puzzle. Although the Internet was crucial to the co-ordination of the sequencing effort and the sharing of its fruits, the biggest single technological advance making the large-scale sequencing of humans possible was the development of brilliantly engineered laser fluourescence sequencing robots. Ironically, these were sold first to private genome hunters who had been closely involved in their development. The main customer was Celera Genomics. When Craig Venter, currently Celera's president and chief scientific officer, vowed in 1998 that he would use these robots to complete a draft of the human genome by 2001, the public project knew it had a race on.

Many perceived Celera's US$70 million challenge as a threat. Although there had already had been a controversy over the US National Institutes of



A stylised image representing the structure of a gene product, a protein, found in the brain, showing normal (A) and mutant (B) forms of the proteins. Although the sequence of both gene products is known, no one has yet determined the exact structure of the protein in the state shown in the images. Instead the researchers involved (Liang Zhong Xu and colleagues at Howard Hughes Medical Institute and The Rockefeller University, New York, NY) have published a representation of a model, rather than an image of a real object. The model was generated using a Linux cluster running the program MODELLER. MODELLER was written by Andrej Sali at the Rockefeller and the program was distributed across a cluster administered with the Clustor system (currently part of TurboLinux's enFuzion 6.0).

Health and its intention to patent certain genes identified by Venter's so-called 'EST' method of gene finding, this was nothing compared to the growing fear of a large-scale stakeout of our genetic heritage by corporate interests. Despite Watson's initial scathing comments on robot sequencers, leaders of the public project knew they had to have the best engines to have any chance in the race. To quote the chairman of Perkin Elmer who made the robot sequencers, "The day we announced Celera we set off an arms race, and we were in the arms business." But even the amazing PRISM 3700 sequencers he was referring to could not read any more than 550 bases of genomic sequence in each 'lane' or column of output.

The public first draft of the project emerged in the 15 months to February 2001 – just before the private effort. Many believe this version of the genome to be better than the 'commercial product' – even allowing for the fact that Celera was free to use information from the public effort while the project's researchers could not do the same themselves. The Genome Center at the Whitehead Institute for Genome Research made a map of the sequence freeze made on 24 May 2000. For this most recent draft a new and more rigorous map was needed. 96 per cent of the sequence of this working draft has a lower than one in 1,000 error rate.

Enter perhaps the most exciting open source contribution to the genome Project race: James Kent and David Haussler, respectively from the Biology and Computer Science Departments of the University of California at Santa Cruz, developed an algorithm called GigAssembler for the final assembly of the overlapping fragments of the code.

Students hacking code is nothing new, but this was rather special code for a rather special code and James Kent is no ordinary student, more of an academic late developer, as one might say. His first career was running a computer animation business. At the time of his involvement last summer he was Haussler's 40-year-old PhD student. For one month he coded the algorithm in C to run on a cluster of 100 800MHz Pentium boxen Haussler had persuaded the University to buy especially for the purpose. To quote Haussler: "Jim in four weeks created the GigAssembler by working night and day. He had to ice his wrists at night because of the fury with which he created this extraordinarily complex piece of code."

Even one error in 1,000 is not good enough, but now we are at a stage where the humble Linux enthusiast can not only read the code, but actually contribute to its interpretation.

Bigger, slicker and free

One of the most interesting parts of bioinformatics is the what comes after you have your sequence. The first tools to manipulate this DNA data *in silico* also started in academia. A collection of scientists in the Department of Genetics at the University of Wisconsin-Madison in 1982 created a bunch of command line tools which went on to be used by computer-savvy biologists everywhere to manipulate sequence data. GCG (Genetics Computer Group) continued as a service of the UW Biotechnology Center from 1985 to 1989. Then it went commercial.

Since then most academic centres have continued to use it at (what most academics thought was) a reasonable price. But a recent price hike of the sort that would shame Microsoft pushed the Human Genome Mapping Project Resource Centre (the HGMP-RC) to the point where they recommend that researchers give up paying them for GCG licences and turn to a free alternative. Worse, even programs written by outsiders to extend the original package started to get mired in the complexities of commercial licences.

Ironically it's taken another group of academics to 're-free' the same algorithms that made GCG worth selling. The HGMP-RC lives on the same Genome Campus outside Cambridge as the Sanger Centre where about a third of the human genome was sequenced. It is part of EMBnet, "a science-based group of collaborating nodes throughout Europe". They used to write extensions to

Further reading

Genome

For a fuller account of the race to sequence the genome itself try 'The Sequence' by Kevin Davies [Weidenfeld; ISBN 0297646982].

Find out more about the biology behind the genome in Matt Ridley's 'Genome' [Fourth Estate; ISBN 185702835X] – both an interesting layman's introduction to the issues raised by the bioinformatic revolution and an overview of its enormous scope.

Go back to the 'real beginning' of the race and read James Watson's entertaining and indiscreet memoir of his and Francis Crick's determination of the structure of DNA, 'The Double Helix' [Penguin; ISBN 0140268774] – now updated with an introduction by media don Steve Jones.

Hardcore

It's notoriously difficult to find any books on bioinformatics itself that cater well for all of those coming from computing, from mathematics and from biology backgrounds. The few textbooks available in the field tend to be eyewateringly expensive as well. If you are a hardcore maths/computing person Michael Waterman's 'Introduction to Computational Biology' [Chapman & Hall/CRC Statistics and Mathematics; ISBN 0412993910].

Alternatively, Pavel Pevzner's 'Computational Molecular Biology – An Algorithmic Approach' [The MIT Press (A Bradford Book); ISBN 0262161974] will give you all the discrete maths you can shake a stick at, but perfunctory introductions to the biology. If you're coming to the subject as a Linux user with a biological background, looking to exploit the many tools available, you might want to try Terry Attwood and David Parry-Smith's 'Introduction to Bioinformatics' [Longman Higher Education; ISBN 0582327881], or Des Higgins and Willie Taylor's 'Bioinformatics: Sequence Structure and Databanks' [Oxford University Press; ISBN 0199637903].

Cancer

For a gentle introduction to cancer biology and its links with genetics try Robert Weinberg's 'One Renegade Cell' [Phoenix Press; ISBN: 0753807866].

Key links

Genome

Zoom in and out on our genetic heritage with the Genome Browser
http://genome.ucsc.edu/goldenPath/hgTracks.html
Visit the focus of the UK's human genome effort at the Sanger Centre
www.sanger.ac.uk/Info/Intro

Bioinformatics

Try your hand at some real-life bioinformatics questions; visit the Open Bioinformatics Project at the Institute of Cancer Research...
www.icr.ac.uk/cmb/bioinformatics/Open
...or Damian Counsell's own website at the Institute
www.icr.ac.uk/cmb/bioinformatics
For answers to more general questions visit The Open Lab's Bioinformatics FAQ
http://bioinformatics.org/FAQ

Selected bioinformatics Linux clusters

The BLAST farm at the Sanger Centre
www.sanger.ac.uk/Info/IT/sld012.htm
The cluster at the Jena Genome Sequencing Center
http://gen100.imb-jena.de/cluster
The YAC at the Mount Sinai Hospital at the Samuel Lunenfeld Research Institute
http://bioinfo.mshri.on.ca/yac
The Collective at the University of Idaho
www.cs.uidaho.edu/~beowulf

GCG when it was cheap now they're writing the free (as in beer and speech) alternative. Being hardcore nerds, these guys have eschewed the messy-but-fast approach of the average open source project for a very disciplined, object-oriented form of C. Not only does it work, but it looks pretty too.

The applications themselves are a boon to practising biologists less interested in bioinformatics for its own sake. Many of them simply replace the functionality of their GCG predecessors and, unlike GCG and, it must be said, a lot of previous biological software packages, they have been released under the GPL.

Even better, EMBnet's hackers have, in passing, had to build a set of amazingly powerful biological libraries which are available under the LGPL. Slowly and surely users are migrating to the new system and I have given EMBOSS my personal 'growing frog' award: it gets bigger and slicker with every change.

Post-genomic hacking

One of the biggest surprises of the Final Draft is that humans appear, according to the latest estimates, to have about half as many genes again as a fly. If we want to refine this count we need to figure out better ways of recognising genes whose instructions are followed by the body to make bits of ourselves. Biologists talk about such parts of the code being 'expressed'. How can we find the 'good bits' that are in the genetic code? Ensembl is another European molecular biology project that is completely open source and uses various statistical techniques to spot the characteristic signatures of actual genes in the vast amounts of noise in the DNA signal.

DNA has been sexy for half a decade, but it's proteins that do the work. If the genes are the software then the proteins are the hardware. Whenever a gene is expressed, it is 'made flesh' as a protein. Last year IBM promised a tidy sum to the development of a machine which can take the one-dimensional DNA code for a gene and predict the elaborate three-dimensional shape of the protein machine it specifies.

Blue Gene, the resulting new supercomputer, will follow IBM's attempt to beat man at chess by tackling this, one of the classic questions in biology, the so-called 'protein folding problem'. In the first week of April 2001 Hitachi and Oracle announced that they were about to spend US$500 million on the human proteome - a detailed description all the products of all the genes in the human genome.

The proteome is a much bigger and much fuzzier target to throw hard cash at. The goal with sequencing the genome was pretty well defined by comparison. People in research were similarly sceptical about private efforts to tackle the genome and the cloning of large mammals, however. The new race is on.

Damian Counsell does bioinformatics in the Protein Folding and Assembly Group at the Institute of Cancer Research, London. He has been using Linux since 1995.