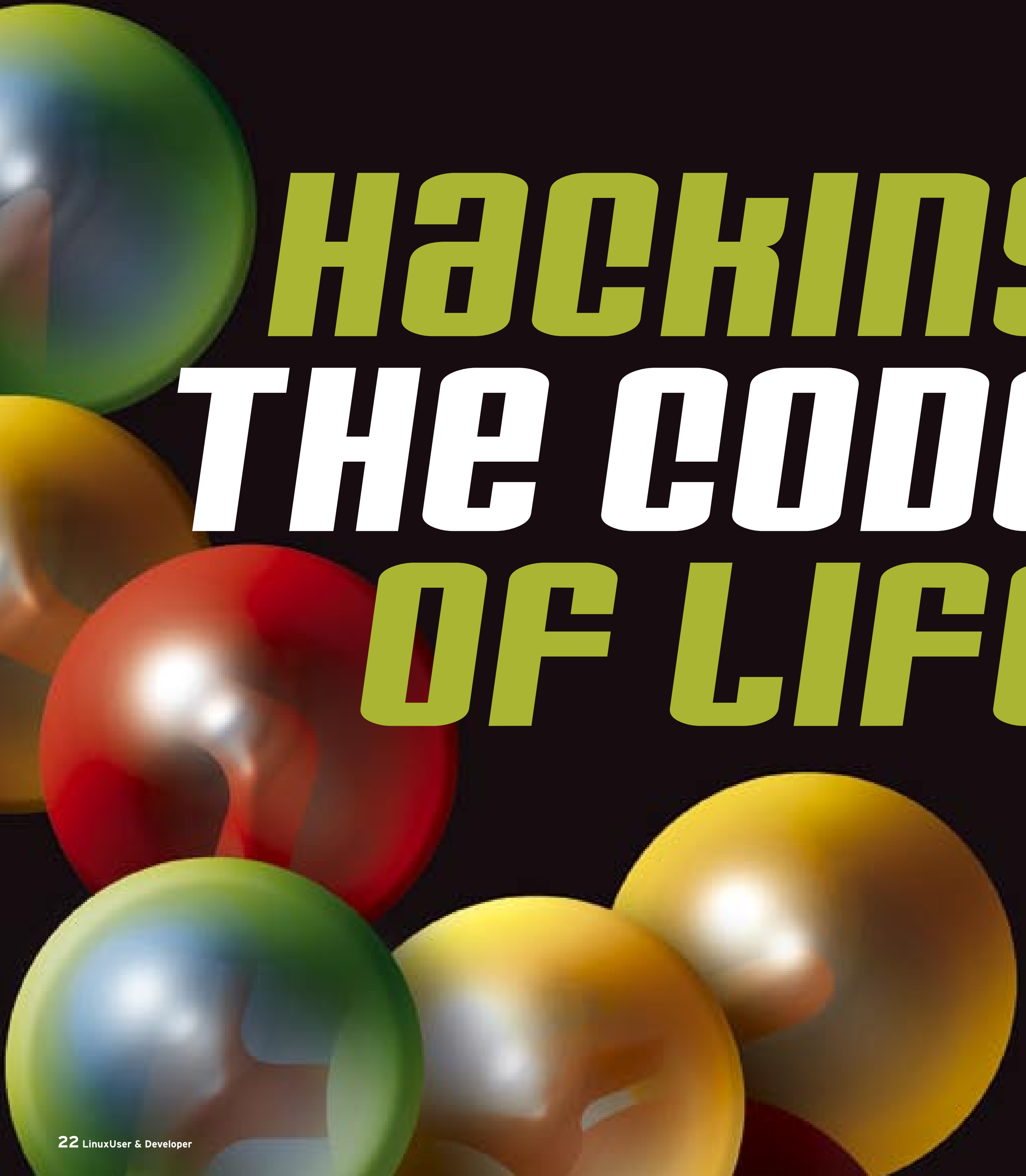
Open source and the bioinformatics revolution

Open source bioinformatics is not going to save the world, but it has been essential to biology's first true big science project, says Damian Counsell. But the project is under threat from proprietary interests and withdrawal of funds



Hacking the code of life

As you read this, your body is writing out billions of copies of your death sentence. If you are holding a paper copy of LinuxUser & Developer, you probably live in relative comfort in a richer developed country. Provided you make it through your next few decades without being electrocuted by a faulty power supply or fatally mauled by your robot dog, the chances are that there will be a significant genetic element in whatever finally kills you: cancer, stroke, heart disease. That mortal curse is embedded in your genes right now in a simple chemical code. The code is written in four compounds, which we represent as four letters: A, C, G, and T. And those letters are the chemical rungs, three thousand million of them, on the ladder of DNA. Each rung is copied into, coiled and supercoiled, and packed with other molecules to make chromosomes. And those chromosomes sit in the control centre, the nucleus, in nearly every one of the trillions of non-blood cells in your body.

Because this information about how to build your body is everywhere in your body, a scraping from your cheek or tissue from the base of a strand of your hair will carry a trace of your genetic identity. As we discover more and more about the meanings those four letters can carry, it will not only be possible to use such traces to put criminals at the scene of a crime, but also to estimate when you will become a victim yourself - a victim of your own nature. Like the bits on your computer's hard drive, the chemical letters of your DNA can be corrupted. Mistakes in the way that your genetic make-up is stored, the way it is packaged, and the way it is copied can lead to disease too, but catastrophic failures of the human operating system can be written into its source from the moment of its conception. Blame your parents.

This is a caricature. When you weigh up the contributions to your demise, it's important to balance the wounds from your genes against your liking for, say, smoking cigarettes or free climbing or eating cheeseburgers. For example, although there is a genetic element in breast cancer, fewer than ten percent of those who die from breast cancer carry one of the two inherited alleles known to be associated with the disease. Even scientists find it very difficult to tease out the relative effects of nature and nurture on your well-being. It doesn't help that questions of human genetics are matters of great political passion and that pseudo-scientific ideas of genetic purity have been used to justify horrific crimes against humanity.

The genome and the arms race

Three years ago LinuxUser carried an article (LinuxUser 11: "Hacking the genome") about the part played by open source software and the open source philosophy in the global scientific mission to read every last chemical letter of every last human gene - and read the apparently useless and often repetitive DNA that fills nine-tenths of the space between those genes. The result of that mission was "The Book of Life": the human genome. Researchers across the planet collaborated across a decade to map and record the information content of DNA taken from a small group of human volunteers. In June 2000, Tony Blair and Bill Clinton made the most public of several announcements of that project's "completion"; in April 2003 99% of the genome was declared to have been sequenced with 99.99% accuracy. About a third of this sequencing was done in Britain at the Sanger Institute, on the Wellcome Trust Genome Campus near Cambridge. We now live in what is often called a 'post-genomic' era.

In 2001 I wrote about how the technology of robot sequencers had transformed the tedious and labour-intensive task of reading genes and fed into an arms race between the public genome project - dedicated to the wide and free dissemination of its results - and the private project, led by Craig Venter of the Celera Corporation - dedicated to the successful commercial licensing of its output, mainly to drug companies.

The race to the human genome sequence was a race for the rights to our genetic heritage. Most of us in the academic biological community wanted this information to be free as in free beer - to ensure that we could all access the fruits of this huge scientific endeavour without having to pay - and free as in free speech - to ensure that we could use that information however we wanted, including in our pursuit of new treatments for genetic diseases. Celera aimed to secure as much intellectual property as possible from the human genome and offered access to their data under conditions or at a price.

The business of applying computers to biology, especially to the biology of genomes is called bioinformatics. The bastard offspring of biology and computing - or, at the least, the product of their shotgun marriage - is itself now the subject of a custody battle between public and private interests. Now I want to argue that we need to keep the tools of bioinformatics as free as the data that we analyse with those tools. Further, we should make sure that the results of our analyses are open for all to read and exploit further in the pursuit of new knowledge and treatments.

The new fight for freedom

Some parts of this battle have been won. The public and private genome projects accepted a fudged public tie in their race to a good rough draft, though there has been plenty of mutual bitching since about their rival definitions of "complete". Data from the public genome projects is freely available. You could download a complete sequence to your PC tomorrow or explore the parts that interest you via one of several Web-based genome exploration tools. And it is more-or-less impossible to "patent genes" in the purest sense of that phrase. The successes of the public project are not just the result of funding from government bodies like the UK's Medical Research Council and the US Department of Energy, but come from aggressive and imaginative support from one of the largest charitable funds on the planet, the Wellcome Trust.

In other ways we are still fighting. Many doubtful gene-related patents remain to be tested in the courts and still hamper both public and private drug research. Perhaps most importantly, we lack open, user-friendly ways to make the Book of Life truly free to the biologists who could most productively use it, analyse it, and manipulate it. (I believe that we have failed in a deeper and more serious way to use computers to change more radically the way biology is done - but that's another story.)

What are the strands to the new battle for genomic freedom?

- Biologists need computational tools to handle their data
- Geeks - tool-makers - need to share with each other and work with the biologists to build those tools well
- The biologist tool-users need to learn to exploit fully the tools the geeks make for them
- All scientists need to be able to publish their work with genomes in a form that is safe and accessible to everyone and that can be mined by machines
- The open source ethos is central to these challenges. It's possible that you could help too

Embossing and blasting genes

The information content of the Book of Life is pretty small by modern computing standards - a genome easily fits onto a CD - but it's the information about the information and the information needed to collect the information that make genomic storage such a challenge.

Running sequence analyses is not usually as computationally demanding as running analyses of particle physics experiments; but it can involve intensive computing, especially when - as drug companies seeking gene patents do - we want to exhaustively search a newly discovered gene against the databases of existing gene for anything remotely similar.

Most bioinformatics isn't complicated and subtle; a lot of it is what people in the business call "blasting". Just as you might "google" the Web for your name, biologists "blast" a gene sequence against a genome, using the eponymous, crafty, corner-cutting search algorithm. BLAST is such a pervasive piece of software that it has even earned itself an O'Reilly book. The many variations on this theme go by exotic names like WU-BLAST and PSI-BLAST, but all of them, important as they are, have one main use and a similar underlying idea.

Life on a CD

If you'd like to play around with some open source bioinformatics software there is already a range of pre-packaged compilations available for you to try. As well as programs that need hardcore science knowledge to use, the collections also usually include friendlier and more straightforward applications plus bundled tutorials and guides.

BioKnoppix [bioknoppix.hpcf.upr.edu] is perhaps the most famous of these open source bio-bundles [though at the moment the home site for the system is unavailable] and is exactly what you would expect from its name: a live CD of biologically-oriented software based around a Knoppix distribution. It's a creation of bioinformaticians at the University of Puerto Rico [www.upr.edu] and packs out a full 688 MB if you download the group's free ISO CD image - they don't sell or give away ready-burned CDs.

DNALinux [www.dnalinux.com] is a more command-line oriented system, prepared in association with the Universidad Nacional de Quilmes [www.unq.edu.ar] in Buenos Aires. Currently at version 0.37, alongside the software common to many of the other CDs - EMBOSS, BLAST, primer3 - DNALinux actually comes with sequence data, for a plant (thale cress), for an insect (the fruit fly) and for a bacterium (E. Coli). Don't be alarmed by the last inclusion. E. coli lives happily and harmlessly in your gut most of the time; it's only occasionally that nasty strains appear in undercooked Scottish burgers and poison people. DNALinux is based on SLAX, a small live version of Slackware.

Vigyaan CD [www.vigyaancd.org] describes itself as "an electronic workbench for bioinformatics". It's based on Knoppix 3.3 so also runs straight from the disc and disappears from your machine when you finish with it. The distribution leans toward chemistry and molecular structure, but the creators claim it is designed with both beginners and experts in mind. The producer is Pratul Argarwal - "vigyaan" means "science and knowledge" in Hindi - and he releases it from the Oak Ridge National Laboratories in Tennessee under an umbrella GPL licence.

Britain's National Environmental Research Council (NERC) [www.nerc.ac.uk] funds the Environmental Genomics Bio-Linux project [envgen.nox.ac.uk/biolinux.html]. Their customised, Red Hat-based Linux comes with pre-built bioinformatics, development, and security tools. The downside is that this all-in-one set-up is targeted to a specific configuration of Dell Precision workstation. In October, developer Dan Swan predicts NERC's Environmental Genomics Thematic Programme Data Centre, sited at Oxford University, will make a Knoppix-based version available.

Like a lot of scientific disciplines, and like most open source projects, bioinformatics started with itches that needed scratching

Another acronym embraces a far wider range of bioinformatics goals: EMBOSS is among the most widely used collections of bioinformatics programs. Unlike BLAST, whose many variants are associated with several American research groups, EMBOSS (European Molecular Biology Open Source Software Suite) has a much more British and European flavour. Its spiritual home is with programmers at the aforementioned Genome Campus, whose landscaped grounds are home to ducks and geese, and hordes of rabbits that line up to stare at programmers as they drive off the site after late nights at their keyboards.

Most academic software is flaky: "read-only" nests of fragile code built to do an experiment or write a paper, but never built to last or to be built upon. It is, however, often willingly shared by scientists. Unfortunately, their employers are usually universities who have one eye on such programs' potential for technology transfer.

Contrary to popular belief, licensing the fruits of academic research brings relatively small sums in to even the most successful universities, but it does force other scholars to sign licence documents so restrictive that many IP lawyers would wince to read them and, ironically, this can lead to not-for-profit institutions having to pay commercial interests to use the results of 'their own' work.

EMBOSS is different. It is written in well thought out C and released, like the Linux kernel, under the GNU General Public Licence (GPL). It is robust and fast. It includes tools to compare, convert, manipulate and analyse gene sequences and now has routines to do similar things with the structures of gene products. Perhaps more importantly, it is constructed on libraries protected by the GNU Lesser Public Library (LGPL) - an open licence that more conventional private companies are not scared of. Other bioinformaticians, programmers and scholars can also use these EMBOSS libraries (called "Ajax" and "Nucleus") to build with and learn from.

Like Linux, EMBOSS began because the commercial alternatives, including some that started in academia, were too expensive and prevented scientists from getting on with the job of decoding life.

Not only can you can download the human genome directly to your PC, you can download this set of tools - and others - to analyse and manipulate the sequences and structures of genes and gene products yourself. Britain is a

LIKE THE BITS ON YOUR COMPUTER'S HARD DRIVE. THE CHEMICAL LETTERS OF YOUR DNA CAN BE CORRUPTED

global centre for the creation of open source tools for hacking biological data, or bioinformatics. Bioinformatics courses have grown up at UK universities more rapidly than any other country. Not only did we use open source software to read the Book of Life, but people involved in the decoding of its contents have contributed to open source development, made the results of their work available openly, published scientific research done with that data in an open-source-like way, and are now pressing for the opening of all scientific literature so that it is permanently accessible for the benefit for all the public and scientists.

THE BIRTH OF OPENSOURCE BIOINFORMATICS

Compared to their brothers and sisters in physics and chemistry, biologists have been relatively slow to take up the tools of mathematics and computing. Like a lot of scientific disciplines, and like most open source projects, bioinformatics started with itches that needed scratching. First biologists, biochemists, and biophysicists worked out how to read sequences and determine the shapes of the molecules they coded for. They needed computers to calculate, assemble, and store the data. They needed computers to analyse, share, and compare it. (In the early days, “computers” meant actual human beings, sitting in labs and offices, doing sums.)

Scientists who understand biology, chemistry and physics, and have real competence in computing are unusual. One of the biggest problems for the field has been the very individuality of the practitioners. Often they have been isolated from other bioinformaticians - or didn’t even know what they were doing was “bioinformatics”. A typical worker would be “the one in the lab who knows about computers”. As gene data grew through the late eighties her task was made easier by the arrival of friendlier computing tools: mini-computers, cheaper micro-computers, the BASIC language, and Apple’s HyperCard. Later these technologies were joined in mainstream bioinformatics by Perl - built to manipulate more mundane character data of course - and the Web - built to share those kinds of data and others.

With the expansion of the academic Net, that guy-in-the-lab often found out (often too late) that she was not alone and that her nifty bit of Perl was out-performed by another, niftier bit of Perl that someone else had written months before. This existing code would probably turn out to do half-a-dozen other things that were equally useful as well. Today there is almost certainly a chunk of C or C++ that does the same job too. What is refreshing is that, these days, it is becoming more and more likely that such code is available under a libre licence and can be taken down from the virtual laboratory shelf and poured into an experiment.

Bioinformatics has come of age, but is only just starting to grow up. By that I mean that the completion of the human genome demonstrated bioinformatics could contribute to important science, and that the successful creation of centres like the European

PUTTING IT ALL TOGETHER

HOW COMPUTERS ASSEMBLE GENES AND GENOMES

Imagine there is a display in a bookshop made up of a stack of paperback copies of “The Satanic Verses”. Someone has broken into the shop and torn every single copy to tiny, but not always identical fragments and stuffed them between the pages of other books. You walk onto the shop floor and have to assemble at least one complete and correct version of the book. This task is analogous to assembling a genome, except the words are invisible to the naked eye. This is also one of the classic problems of bioinformatics.

The data demands of the 3000 million characters of the human genome sequence seem, on paper at least, to be relatively small and, in the recent past, the chemical technology available only permitted a tiny, tiny subset of these characters to be harvested. Why has sequencing been so heavily dependent on computers? One main reason is that we can’t read off the data of the genome like a very long piece of ticker tape. Even the body itself doesn’t copy or read DNA continuously. We don’t even read the code directly from cells of human beings. Instead we have to extract it from human tissue, break the DNA up into short strips, fool bacteria into making copies for us, chop out and chop up those duplicated strands, and then reassemble them - without any labels telling us which fragment follows which other fragment.

Imagine you are a scientist working in the dark days before industrial-scale sequencing - just a few years ago. You suspect that a particular region of the, so far sparsely mapped, genome contains a disease gene. You would only have a small team of researchers available to you and you have to do most of the genetic engineering and sequencing yourself. You don’t have the time or the experience to learn to program or learn UNIX. You need an intuitive interface to assemble stretches of overlapping DNA sequence, correct errors, filter out experimental noise, and remove the messages from the other organisms you had taken over with foreign DNA. Perhaps you used an expensive, possibly Mac-based, commercial software package or the refined and public package known as “Staden”, after its original creator Rodger Staden. Until recently Staden worked at the legendary Laboratory of Molecular Biology in Cambridge UK, working home to thirteen Nobel prizewinners in its fifty or so years.

As the public genome project raced the private project to release a complete rough draft and put it into the public domain as quickly as possible, this very problem - the final assembly of the overall structure of the genome - gave rise to one of the real legends of open source programming and of the genome project itself. A former animation programmer and mature PhD student, James Kent spent a month writing GigAssembler, a C program to produce the final assembly of the human genome for the public project. He soaked his wrists in ice in the evenings as he cranked out code to run on a cluster of 800MHz Pentiums bought by the University of Santa Cruz specifically for the job. For his efforts Kent collected Bioinformatics. Org’s Benjamin Franklin award. You can read the story in more detail here in the UCSC magazine here: **www.ucsc.edu/currents/00-01/02-12/genome.html**

BUGS IN YOUR CODE - GENES AND DISEASE

Long before we could actually "read" the letters of genes, we managed to trace the steps of some of the simplest and most damaging genetic diseases. Now, we know exactly what changes in what characters of what genes account for most of these illnesses. Any given gene can come in different forms, called "alleles". In most cases, you receive one copy from you father and one from your mother. The exact version of the gene passed on to you is usually an allele that's common in the general population. When people talk about someone "having the gene for", say, sickle cell disease it's misleading; everybody has the gene for haemoglobin, which is what the sickle cell gene makes. Some people have one or more copies of a potentially damaging version of that gene. A person who is said to "have the sickle cell gene" has (at least) one of the alleles that makes the "wrong" product, that is, one that causes problems for the person who inherited it - even if it didn't cause problems for the parent who passed it on.

Often, one copy of a gene can be seen as a back-up of the other. Sometimes they combine to give a summed overall effect. Sometimes a mistake in one can have serious consequences that the presence of a spare copy cannot correct. Even more strangely, a supposedly dodgy allele from one parent might give you an advantage over your less fortunate friends, but only if it is combined with a sound allele from the other parent - two bad copies and you're in trouble. Alleles that are fatal to people unlucky enough to inherit two copies can hang around in a population, thumbing their noses at Charles Darwin. Darwin devised his theories, of course, without knowing that such things as genes existed. We now have answers to many of the questions that the continued existence of such diseases posed for his theories. Ironically, the persistence of the gene variants that causes them has confirmed the soundness of several of his fundamental ideas.

Although scientists and doctors have firmly attached many clear-cut and easily detected genetic disorders to particular forms of genes or particular kinds of chromosome damage, it's probably true that the bulk of genetic disease and death results from the combined effects of many as-yet-unknown bugs in our code and the combined effects of many inputs from our environment. Reading the genome has shifted the emphasis of research into genes and disease. Now we have a reference copy of the human genome, we can, in theory, read what we believe are the critical parts of the code of many individual humans, examine their lifestyles, and see what kind of lives are associated with what permutations of nature and nurture. At the same time, we can look, not just at the genes themselves, but at how much different parts of the body use these genes at different times and under different conditions, such as when they are malignant or infected.

Bioinformatics Institute (the EBI), also at the Genome Campus, shows that bioinformatics could develop a body of thought, like biochemistry and molecular biology before it. Sadly, in Britain at least, our lead in this field could be threatened by recent developments. Next year, the Medical Research Council will shut down the Human Genome Mapping Project Resource Centre (the HGMP-RC, recently renamed the "Rosalind Franklin Centre for Genomics Research") where several of the key developers on the EMBOSS team have been employed.

BIOINFORMATICS.ORG

You don't need to work in a lab or be a professional scientist to jump straight into bioinformatics. Even if all you want to do is find out what bioinformatics is about you don't need to be in a big university or research centre. A great place to start is the relatively free-standing and non-academic site of "Bioinformatics.Org". Located in Massachusetts, Bioinformatics.Org was set up in 1998 by Jeff Bizzaro, a PhD student. It's now officially an independent not-for-profit corporation. It hosts my "Bioinformatics Frequently Asked Questions". This work-in-progress which will, one day concentrate more on day-to-day questions from people established in the field - just as Linux "How-To" documents are used by even UNIX gurus. For now, because these are literally more "frequently asked", it devotes most of its space to answering requests for definitions of the field and its cousins and to offering advice on how newcomers can get trained and get involved in bioinformatics.

THE RACE TO THE HUMAN GENOME SEQUENCE WAS A RACE FOR THE RIGHTS TO OUR GENETIC HERITAGE

After you've browsed the basics and realised that there are ways anybody can contribute to the work, you should explore some of the many projects that Bioinformatics.Org hosts and read about the organisation's campaigns for freedom of information in biology. Every year Bioinformatics.Org gives its Benjamin Franklin award "to an individual who has, in his or her practice, promoted free and open access to the methods and materials used in the scientific field of bioinformatics". You can think of Bioinformatics.Org as combining aspects of Sourceforge, providing a home and an infrastructure for open projects, and the GNU Foundation, advancing the philosophy of open source in science, bio-computing, and scientific, technical, and medical (STM) publishing.

GENOMICS HACKATHONS

A parallel not-for-profit is the Open Bioinformatics Foundation (OBF). It has a narrower focus than Bioinformatics.Org, being concerned mainly with the building of frameworks for bioinformatics programming: BioPerl, BioJava, BioPython and so on. These give genome hackers whole suites of pre-built biologically oriented modules, objects, and libraries, ready to use and (to a lesser extent) ready documented. If, for example, you need to read various formats of DNA files (and there are lots of formats for storing gene sequence data) then you can use a BioPerl module rather than have to figure out how to parse tens of different variant file formats. Every year, members of the OBF teams meet each other to discuss their plans and the requirements of their users. They also hold hackathons, intensive coding sessions in locations such as Arizona and South Africa where bioinformatics programmers get together to build on the existing code and grow their ponytails and goatees.

PUTTING EEVERYTHING TOGETHER: ENSEMBL

If you want to wander around your genome (and those of other species) from the comfort of your own computer, you are spoilt for friendly guides. You can use the UCSC (University of California Santa Cruz) Genome Browser [genome.ucsc.edu/cgi-bin/hgGateway] or the NCBI's MapView (NCBI is the US National Center for Biotechnology Information). The local choice, however, is the European Bioinformatics Institute's Ensembl [www.ensembl.org]. All of these genome exploration tools build on the same basic map of the human genome, but all have their own slightly different view of the landscape. Ensembl, in particular, attempts to work out the structure of every gene automatically. Like the others, Ensembl's interfaces is straightforward enough for even complete newcomers to tour their own genes.

IT DOESN'T HELP THAT QUESTIONS OF HUMAN GENETICS ARE MATTERS OF GREAT POLITICAL PASSION AND THAT PSEUDO-SCIENTIFIC IDEAS OF GENETIC PURITY HAVE BEEN USED TO JUSTIFY HORRIFIC CRIMES AGAINST HUMANITY

Ensembl and its creators have a strong open source ethos. One of Ensembl's founders, Tim Hubbard, has even proposed a framework for open source drug development. The Ensembl project is a result of work at the Sanger Institute, funded by the Wellcome Trust, and at the EBI, funded by the European Union, via a virtual international laboratory called EMBL. On top of an open source infrastructure (Apache and Perl and friends), Ensembl weaves together a whole range of different kinds of information about genes, their sequences and positions in the genome, what they make, and about, for example, the diseases associated with them. It applies a wide range of open source programs to gene information and it offers a single window to these data with full documentation and a polished look. It manages to be automated, comprehensive and scientific, but at the same time presents a friendly face to biomedical scientists who are not necessarily computer geniuses.

Sacking the hackers: Staden and EMBOSS

Imagine if, within a period of two years, the leading developers of both BSD and Linux lost their jobs and could no longer afford to work on the projects. Last year the UK Medical Research Council (MRC) decided to withdraw its (already somewhat indirect) support for both the Staden programs [see BOX: "PUTTING IT ALL TOGETHER"] and the EMBOSS system.

Staden was a closed package whose copyright was owned outright by the MRC, but one that, until relatively recently, was distributed according to time-honoured scientific tradition: on a collection of tapes. There was a nominal fee, but the real pay-off for the "purchaser" was the amazingly high level of support that comes from having friendly collegiate access to the actual developers, who promptly fix problems and improve "the product" out of intellectual pride (rather than in return for the promise of stock options).

When Rodger Staden lost funding for his staff of developers, he chose retirement and climbing mountains, rather than continuing on the academic treadmill. One of those principal developers, James Bonfield, is now at the aforementioned Sanger Institute, where parts of the Staden package are used heavily in a production environment. He continues to maintain and improve those parts, but the other pieces of the suite would benefit from volunteers to take them on and keep them going. You can visit the Staden package's Sourceforge page here:

staden.sourceforge.net

THE BUSINESS OF APPLYING COMPUTERS TO BIOLOGY, ESPECIALLY TO THE BIOLOGY OF GENOMES IS CALLED BIOINFORMATICS. THE BASTARD OFFSPRING OF BIOLOGY AND COMPUTING - OR, AT THE LEAST, THE PRODUCT OF THEIR SHOTGUN MARRIAGE...

Bugs and bad guys

Open genomics and terrorism

Many of the genomes we have sequenced are from killer bacteria and viruses, "pathogens". We didn't only do this because we are interested in the diseases they cause, but for other practical reasons: they are small and simple, have little "nonsense" in their DNA (or RNA), and they are often parasites anyway - so need only contain code for functions they haven't already borrowed from their unhappy hosts: us.

The data for the blueprints of these potential bioweapons are as freely available as those for humans. We already know the codes for over a hundred species of bacteria and more than a thousand viruses. It is possible to assemble large chunks of pathogen genomes from public DNA data, but it's a pretty silly way to go about making your own biological weapon. More frightening, perhaps, is the possibility that armed fundamentalist religious and political groups might insert dangerous new genes into existing, less harmful bugs in order to make them more effective killers.

The US government funds sequencing of pathogens and often requires that those who receive its money make the data they accumulate accessible. Given the new security climate in the World, they want to weigh the risks. In September 2004 the results of a report commissioned by the CIA and the US National Science Foundation were released. Stanley Falkow, the Stanford microbiologist leading the study stated that "open access is essential if we are to maintain the progress needed to stay ahead of those who would attempt to cause harm" and agreed that mass murderers would be unlikely to gain much from our keeping raw sequence data in public. Even if it made sense to keep such information off limits, any plan to hide it would be extremely difficult to work out and to make work. And, of course, the government can still declare certain results it has paid to obtain "classified". For what you would expect to be a highly controversial area of discussion, the scientific and government consensus about what to do with pathogen data was surprisingly strong and broad.

sciencenow.sciencemag.org/cgi/content/full/2004/909/1

In one sense the Staden package was a victim of the success of large-scale sequencing. Small sequencing projects are nothing like as common as they used to be. In another sense the "death" of Staden was a small victory for the open source spirit. Instead of being transferred to the private sector and shrink-wrapped in a restrictive licence by a technology transfer process or left to decay unmaintained, the MRC were, at least, persuaded to open the code of the Staden system under a BSD-style licence so that others, both public and private can now use and extend it.

The more recently condemned members of the EMBOSS team are now beginning to scatter. They are moving to other jobs before the MRC's Human Genome Mapping Project Resource Centre closes and/or looking for alternative funding. It's not just support from the MRC that may disappear this year; the Biotechnology and Biological Sciences Research Council, the BBSRC, also used to back EMBOSS-related activities. A founder researcher and programmer on the project, Alan Bleasby, has been trying along with colleagues to obtain new funding for EMBOSS. Another EMBOSS founder, Peter Rice (who has worked both in not-for-profit and private bioinformatics, and is now a group leader at the EBI) has been guiding a team in an already-funded effort to integrate EMBOSS and other bioinformatics programs and services into workflows, so that biologists and bioinformaticians can pick, mix, and bolt together useful functions and create computational pipelines, to process streams of genetic data.

What's next?

Now the principal individual who drove the private genome project, Craig Venter, is pushing forward his plans for the "thousand-dollar genome" with a new company, US Genomics. This company and others want to develop technology that would permit ordinary humans to read the contents of their genetic code - or the important bits at least - for a matter of hundreds of US dollars. In a country with a healthcare system that is largely private and with many commercial health and life insurers such a development would have profound effects on millions of lives.

Strangely, one of the things that we have learned from existing screening for genetic disorders is that knowing you have a potentially dangerous genetic pre-disposition does not necessarily improve your life - even when you have access to preventive treatment. Health is as much about the way you feel about your life as the life is treating you. Considerations like this are deeply important and we have to think now about what will happen when we can afford to carry our entire genetic make-up with us, not just in our bodies, but in machine-readable form on a smartcard. We also have to think deeply about what seems at first to be a less important issue: who will own the code that processes that data? It's one thing to archive all of your office documents in a closed, changing, and cryptic file format (like ".doc", say), and quite another to store your medical details that way. Do we want to hand over analysis of our genetic data to a Microsoft of bioinformatics?

DO WE WANT TO HAND OVER ANALYSIS OF OUR GENETIC DATA TO A MICROSOFT OF BIOINFORMATICS?

Open source bioinformatics is not going to save the world, but it has been essential to biology's first true big science project - biologists had never previously operated at the fly-a-man-to-the-moon, build-a-super-collider scale. Like Linux, the initial pay-off will come from our being able to do vital work - in this case biomedical research - far more cheaply, quickly, and reliably. The deeper reward lies in the philosophical and cultural changes it will bring. The Web was invented by a Briton and was also a by-product of another big science project (CERN). When it came to exploiting the WWW, we often wasted our natural advantages - for example, through our initial failure to build and open a broadband telecoms infrastructure.

In the future, not only will many people have some inkling of when Mother Nature plans to recycle them into daisies, they will have a much better chance of postponing it. Open source bioinformatics will bring insights into our genetic fates closer. While that might be a mixed blessing, it should help us to benefit from that knowledge sooner too. The danger now is that, by failing to support the inexpensive and successful open bioinformatics projects we have already started, the UK will let others seize the prizes of our country's huge contribution to the genomics revolution.