Abstract
In June 2000, the draft sequence of the human genome was announced. It is, and will be for some years, incomplete, but the vast majority is now available. Currently about a third is finished (including two complete chromosomes); the rest has good coverage, but not long-range continuity. First-pass analysis indicates, among other things, fewer genes than expected: about 40,000 now looks a likely number. This uncertainty illustrates the difficulty of interpretation: the sequence is not an end in itself, but a resource to be continually reanalysed as our biological understanding increases. That is the scientific reason for releasing it promptly, fully and freely. The social reasons for doing so are even more compelling.

Introduction
Looking back over history, we see that humanity has gone through a series of episodes, where new methods of observation and thinking have led to rapid changes in our perception of ourselves. One such was the cosmological revolution of the 17th century, in which a combination of the invention of the telescope and mathematical theory led to an understanding that we stand not at the centre of the universe, but in a nondescript corner of it.

Again, the enunciation of the theory of evolution in the 19th century made (most of) us believe that we are simply a particular sort of animal, evolved from the same common ancestor as all other living organisms, and not set apart from them in an absolute way.

We are now in the midst of another revolution – that of molecular biology. The great pioneers of the mid-20th century uncovered the basic molecules and mechanisms of life, and we are all now engaged in discovering exactly how it all works. Reading out the human sequence is really just one more step in this process.

However, some celebration for this particular step is appropriate. After all, it's the first time (as far as we know) that a life form has read out the code of instructions for its own assembly. The event has, rightly, captured the popular imagination, as have the potentials for medical advances and the possibilities for harm. This opens up valuable opportunities for public debate and increased understanding, all of which will be essential for the maximization of benefit.
From worm to human

The origins of genomics go back to the early 1980s, when it became apparent that it would be useful to tackle large genomes as a whole, rather than relying on gene-by-gene analysis based on classical genetic methods alone. Apart from small viruses, the initial effort was directed towards mapping, it being impractical at the time to sequence at the multi-megabase scale: the dideoxy sequencing method invented by Fred Sanger was as yet only 5 years old. The nematode worm Caenorhabditis elegans was one of the first organisms to be studied in this way, and by the late 80s we had built a map of cosmids and yeast artificial chromosomes (YACs) covering almost all of the genome. Crucially, we recognized that data sharing was essential. Two labs, ours and Bob Waterston’s, focused on physical mapping, while the necessary correlation of clones with genes came about by continuous collaboration with the whole worm community.

By 1989, we were ready to start sequencing. By that time, the value of tackling the 3000 megabase human genome had been recognized on both sides of the Atlantic in the form of the Human Genome Project (HGP), and we were funded to begin sequencing the worm as a pilot project under that umbrella.

At this point, there was some debate about the value of sequencing whole genomes compared with simply sequencing coding regions in the form of cDNAs. Two crucial arguments prevailed, however: one was the impossibility of actually finding all the genes by cDNA technology alone; the other was that the non-coding regions contained important information, including control sequences and structural elements. So, at least for a few organisms, it would be essential to read out everything.

The debate has now been concluded in favour of a combination of approaches, since the cDNA sequences are invaluable for interpretation, while the necessity of genomic sequence for completeness is no longer challenged.

Worm sequencing was done by breaking a minimal set of mapped clones into small random fragments (‘shotgunning’ in Sanger’s parlance), which were cloned and thus provided the templates for sequencing. The sequence reads were assembled to give the clone sequences, which in turn were overlapped to give the genome sequence. Apart from being important in its own right, the worm sequencing project helped to drive the technology and motivation for human sequencing. The same process was used for humans, except that the mapped clones were the newly invented BACs (bacterial artificial chromosomes), rather than cosmids and YACs.

Both worm labs (WashU and Cambridge—now transferred from MRC-LMB to the Sanger Centre) had begun sequencing human clones by 1993. In late 1994, Bob Waterston and I started pressing for human sequencing to accelerate, by producing an initially imperfect product that would be finished later. This scheme did not at first find favour, but in 1996 the first of the ‘Bermuda’ meetings was held. These gatherings brought together groups from all over the world who were interested in large-scale human sequencing. A significant development at the first meeting was the formulation of the Bermuda principles. These came directly from our experience with worm genomics, and stated that large centres would release all their data immediately and would not claim patent rights. It was important to have a written code of conduct for human sequence data, because of the high stakes.

In 1998 the worm sequence was largely complete and was published. But earlier that year the international HGP was challenged by a rival in the form of Celera, a private corporation that intended to sequence the human genome rapidly to a limited accuracy, and then profit from patenting and sale of access rights. The mapping step would be bypassed; the entire genome would be shotgunned directly into small cloned fragments and the sequence of the genome would be assembled from the resulting reads in a single step. Celera stated that the world would get the human genome free of charge, so why should the HGP continue? Some of us had doubts (well-founded, as it turned out) both as to the practicality of the methods proposed by Celera, and as to the wisdom of leaving so important a piece of information in the hands of a private corporation. Fortunately, this view rapidly prevailed.

Thanks to the earlier pressures to accelerate, the funding agencies were ready, and a very rapid increase in public output began in 1999. In both public and private efforts, the acceleration was predicated upon a new generation of sequencing devices using capillaries, which reduced labour costs and allowed rapid scale-up. Furthermore, in order to meet the now sharpened expectations of biologists, the notion of an intermediate goal was resurrected, and named the draft sequence. It would start from the mapped clones, but to save
time they would initially be shotgunned to half the usual depth. Thus the draft would have temporary gaps but would provide good overall coverage at an early stage. This is the product that has been announced this year, and so the great majority of the human genome sequence is now accessible from public databases to anyone with a PC. A third of it is already finished, and all will be finished within two years.

Interpretation

Table 1 summarizes some of the ways in which genomic information can be used. The key point is that the sequence is not an end, but a beginning. There are orders of magnitude more work to be done in biology than went into sequencing. Yet the sequence — even when imperfect, but much more so when finished — is an immensely powerful foundation. Possession of it is not merely a convenience, but provides entirely new points of departure for the experimenter.

As yet, the detection of genes, even in finished sequence, is still in its infancy. The low coding density of the human genome, and the fragmentation of genes, mean that the signal/noise ratio for analysis ab initio is greatly inferior to that in model organisms like yeast, worms and flies (e.g. human coding density is around 2%, worm around 30%). Matches to cDNAs and similarities to related genes allow many predictions to be made, but this initial annotation is inevitably imperfect. There are therefore active experimental programmes to test regions where genes are weakly predicted for actual transcription, so as to consolidate the verified gene set.

In addition, comparative genomic analysis will be an increasingly powerful interpretative tool as more vertebrate genomes become available. Consider, for example, the evolution of the mouse and human genomes after they diverged some 100 million years ago. Mutations and copying errors caused all the genomes in the two populations to diverge gradually from one another. Within each species constant mixing during meiosis and reproduction, followed by natural selection, kept the genomes within a narrow range (about 0.2% divergence, in our case). Between species there is no such cross-checking, but there is the overarching requirement to make a viable mammal. We therefore find that the key features — coding regions and nearby sequences that are required for control or other purposes — are strongly conserved, while the intervening sequences have drifted apart to a much greater extent. The comparison provides a wonderful tool with which to identify and delineate genes and other functional elements.

A public-private consortium is now generating mouse sequence from a whole genome shotgun. The data is released publicly and freely on to special repositories at EBI and NCBI, and will be at a depth of 3× by March 2001. In time, this will be combined with data from mapped clones to generate a complete mouse sequence, but meanwhile the random reads provide a valuable annotation tool. Random sequence reads from the fish Tetraodon are also publicly available, and are

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value of the genomic sequence</strong></td>
</tr>
<tr>
<td><strong>Materials</strong></td>
</tr>
<tr>
<td>Sequence known in advance</td>
</tr>
<tr>
<td>Gene-hunting by computer</td>
</tr>
<tr>
<td>Tools for manipulation</td>
</tr>
<tr>
<td>Total gene inventory</td>
</tr>
<tr>
<td><strong>New entry points</strong></td>
</tr>
<tr>
<td>Matches → protein types</td>
</tr>
<tr>
<td>Gene disruption → function</td>
</tr>
<tr>
<td>Gene expression → patterns</td>
</tr>
<tr>
<td>Organism specific genes</td>
</tr>
<tr>
<td>Comparative genomics</td>
</tr>
<tr>
<td><strong>Genome structure and archaeology</strong></td>
</tr>
<tr>
<td>Long-range structure</td>
</tr>
<tr>
<td>Evolutionary studies</td>
</tr>
<tr>
<td>Draws everything together</td>
</tr>
<tr>
<td><strong>Complete information</strong></td>
</tr>
<tr>
<td>Archive for the future</td>
</tr>
</tbody>
</table>

© 2001 Biochemical Society
similarly valuable. The rapidly increasing stocks of public sequence from a variety of organisms will drive progress.

Despite the current limitations, a reasonable overall picture of the human genome can already be obtained. One surprise has been that the estimated number of genes is lower than generally expected. Around 40000 is a good guess at the moment. Of course the number of gene products will be much larger, due to alternative splicing and a variety of post-translational modifications. But it is striking that a human can be constructed from only twice as many genes as a nematode.

The protein analysis group of the HGP has begun to investigate what the extra genes look like. For the most part they are not entirely novel, but more commonly consist of new combinations of old domains. The origins of these novel architectures will become clearer as more genomes are studied.

On a larger scale, it is very striking to see the overall landscape of the genome unfold. Such features as the regional variation in GC content, the fossil transposons that can be arranged into lineages by their sequence variation, and the patterns of duplication in gene families are being catalogued. Much can be learned in this way about the evolution of the genome, but, again, comparative genomics will provide much greater analytical power.

**Application**

The sequence is already proving its worth for gene finding. Even now there are more than 30 publications on genes involved in genetic disease that have been isolated using HGP data, and vastly more studies are in progress or unpublished. It should be remembered that not only is this the only data that is freely available, but also that much of it has been available for some time.

So far, by human genome we mean a reference genome, not that of any particular individual. What are the genetic variants in the population, and what are the differences between us to which they give rise? Identification of variations at the single nucleotide level, single nucleotide polymorphisms (SNPs), is proceeding apace, with well over 1 million variants now catalogued. The SNP consortium is again a public–private partnership that releases its data freely.

Although we do not of course actively experiment on human beings, it will be possible to deduce a great deal from the vast genetic exercise that 6 billion people are engaged in as they procreate. By collecting large numbers of genotypes and comparing them with medical records it will be possible to home in on the variations that are medically important.

However, in order for this to be successful, it is of critical importance that genetic discrimination of any kind is outlawed, in the same way that the specific discriminations concerning skin pigmentation and the number of X chromosomes one possesses has been outlawed. Otherwise, people will, very sensibly, refuse to be tested and much potential benefit will be lost. For this reason, as well as ethical considerations of elementary human rights, governments need to give a lead.

**What medical gains can we expect?**

Once a gene has been sequenced and recognized as having variants responsible for a particular medical condition, it is straightforward to screen patients for that condition. This is more difficult when multiple genes are responsible, but still achievable if enough data are collected for the combinatorials to be worked out (e.g. programmes on diabetes, heart disease). Hence the importance of screening large populations.

However, diagnosis without cure is not of much value to the patient, except for the limited case of prenatal diagnosis and termination for fatal conditions. The ideal is to proceed to therapy.

The pharmaceutical companies are hoping that drugs can be chosen or modified to suit the patient’s genotype, and no doubt this will be of some value. People will be given advice on the basis of genotype – perhaps controlling their diet or other aspects of lifestyle. This does rather conjure up the prospect of a nation of hypochondriacs anxiously consulting their genotypes before choosing a restaurant. Probably most of us won’t bother.

True gene therapy is going to be harder, requiring delivery of a gene to the right place and having it expressed correctly and stably. But there are some encouraging early results, and in time there will be many successes.

Cancer treatment is likely to benefit relatively early. This is a matter of killing rogue cells, which can be an easier task than improving debilitated ones. It is hoped that detailed analysis of tumour genomes will reveal new targets for toxic drugs more specifically toxic to tumours, and therefore with fewer side effects for the patient, as well as higher cure rates.

Beyond analysis and treatment of genomic defects lie the rewards of having a complete parts
Society and the human genome

list for the human body – the set of proteins and RNAs that actually do the work. Already the gene list is extensively used to manufacture chips that detect expression, and so can be used to monitor the function, as well as the presence, of a gene. Structures of proteins will be uncovered systematically, as will interactions between them. Control mechanisms will be hard, but buried in the genome they are all there. The workshop manual for the human body will become a reality.

There is a hierarchy of difficulty in these developments, and it is crucial that the easy ones are not used, through excessively broad patent claims, to impede progress on the difficult ones.

Freedom of information
Come what may, the human genome sequence is now out in the public domain. The HGP consortium will finish it to a high standard over the next two years, and along with other groups will continue to refine it. The interpretation, however, will not be the preserve of the few but of the many, working with the freely accessible basic information. This is crucial, because we are no longer talking about annotation, but about the complete understanding of the human body in molecular detail. Many more large-scale projects, involving proteomics, structural biology, neurobiology, and so on will be needed to even come close to this end; even more needed will be countless individual problem-solving efforts. The more communication we can have the better and faster it will go.

In order to ensure its continued availability, genomic sequence and, as far as possible, related information needs to be held in public repositories. It is unlikely that a private database can fulfill the essential requirements, as well put by the journal Nature (23 March, 2000), of being ‘reliable, publicly available, unrestricted and free’.

To maintain the viability of the public databases, as well as to advance genome knowledge in general, I hope that Europe will continue to develop as a strong co-operative grouping. In order to guarantee stable and open databases, there is a need for more than one centre in the world, and support for the EBI is crucial in this regard.

In conclusion, I want to thank the hundreds of devoted people who have contributed to the HGP and made it speedy, efficient and effective; they have done this not for unusual wealth or for unusual recognition, but to benefit human-kind. For links to public resources see: www.ensembl.org/genome/central

Note added in proof (received 2 March 2001)

The HGP has now published the draft sequence in Nature [1]. Another draft version of the human sequence has been published by Celera, in Science [2]. It draws heavily on the public data, and cannot be regarded as an independent product. This frozen version is available free to academics, but stringent restrictions on use limit its value.

My admiration and thanks to all my colleagues in the International Human Genome Sequencing Consortium for the collective achievement. I would like to thank the Wellcome Trust and the Medical Research Council for their generous support.

References
1 Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium (2001) Nature (London), 409, 860-921. In the same issue, there will be several papers on the interpretation and use of the sequence.

Received 1 February 2001

© 2001 Biochemical Society