The turning point in genome research

One could argue that, until recently, the impact that genome research has had upon the rest of biology has been rather modest. Human genome research has predominantly concerned itself with large-scale mapping and technology development[1], activities that require an industrial mindset best left undistracted by that smaller-scale, more intimate approach to Nature that most biologists have traditionally pursued. This is not to say that the wider community has not benefitted from the technology development (mostly in the areas of automated sequencing and robotics). Nor can one deny the excitement and insights that have resulted from positional cloning[2], an approach to gene isolation that relies heavily upon the infrastructure that the genome program has created. Some fundamental new biology has come from this; the involvement of trinucleotide repeats in the pathophysiology of fragile X syndrome, Huntington's disease and other neurological disorders is a striking example[3]. But even here the outcome has been slight by one measure: to date only about 40 or so human genes have been cloned by virtue of their location in the genome[2], compared with perhaps 36 000 sequences in the primate division of GenBank® that are the result of 'functional cloning' experiments, whereby one starts with the biology first and only then isolates the genes or gene products involved. Insofar as the major data products of genome research to date have been genetic and physical maps, the genome program has been largely irrelevant to biochemists, cell and developmental biologists, neurobiologists and other biological scientists for whom phenotype, not genotype, is the alpha and the omega.

But all of this is about to change. The vital link between genome research and the rest of biology is DNA (and inferred protein) sequence data, a commodity that the genome program will soon be producing in abundance. More about that later, but first let us consider 'expressed sequence tags' (ESTs), partial cDNA 'survey sequences' that were conceived as a short cut to the finish line[4], but that were initially viewed by genome researchers as a diversion, indeed even an obstacle, to achieving the final goal: the complete three billion nucleotide sequence of the human genome.

EST data accumulation

Space does not permit a detailed review of EST research over the past four years, but the field got off to a rocky start in 1991, owing to a distraction over patent issues and lack of public funding. Large-scale EST sequencing, championed by Craig Venter of the Institute for Genomic Research in Gaithersburg, MD, was readily taken up by the private sector, and proprietary data collections rapidly surpassed what was trickling into the public databases[5] from small and under-funded, but dedicated, groups. Gradually though, it became apparent to most biologists, even former skeptics, that this type of incomplete and inaccurate, but rapidly and cheaply obtainable data was indeed quite useful: for 'phylum-hopping' among organisms[6]; for discovering new members of gene families involved in human disease[7]; for the identification of exons in vast expanses of genomic DNA[8]; and as a plentiful source of gene-based mapping reagents with which to populate physical maps[9]. Public data collections got a tremendous boost when Merck & Co. launched a project to compile a comprehensive sample of expressed human genes, represented by 5′- and 3′-sequences from 200 000 cDNA clones, all for the public domain. This program is being carried out by a consortium centered around the Genome Sequencing Center at Washington University in St Louis (directed by Robert Waterston), with cDNA libraries constructed by Bento Soares at Columbia University and arrayed for high-throughput processing by Greg Lennon at the Lawrence Livermore National Laboratory. To date, more than 150 000 sequences have been released and new data are being submitted to GenBank at a rate of 1500 sequences per day. These sequences,

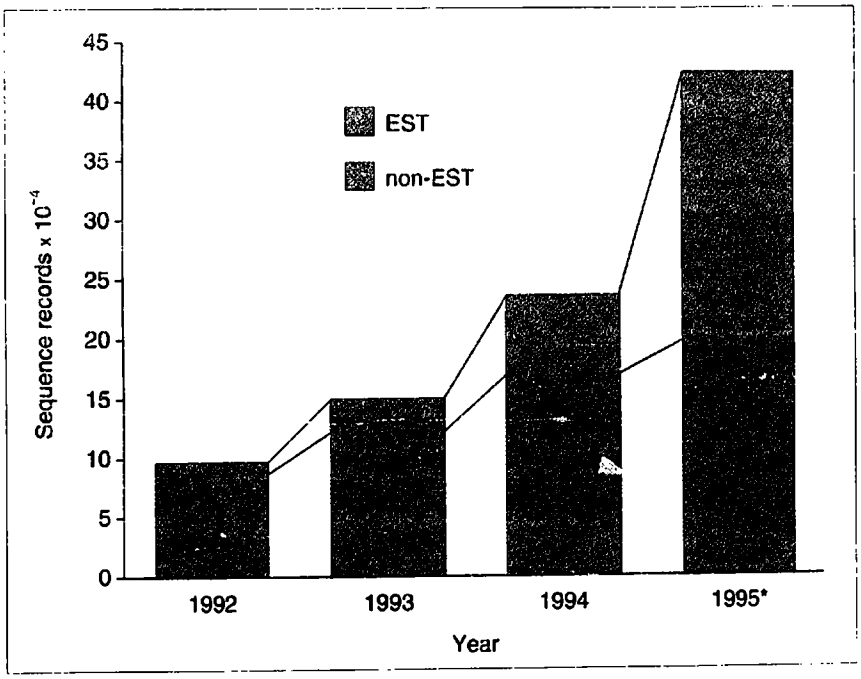

Figure 1

Growth of GenBank and its expressed sequence tag (EST) division. Release 89.0 of GenBank (June 15, 1995) contains 425 211 sequence records totaling 318 624 568 bases. Of these sequence records, 255 244 represent partial cDNA sequences or ESTs. The data are taken from the GenBank release notes available by anonymous ftp from **ncbi.nlm.nih.gov** (the file is gbrel.txt in the /genbank directory). More detailed information on the contents of the EST Division can be found on the World Wide Web at: **http://www.ncbi.nlm.nih.gov/dbEST/index.html**. The asterisk after 1995 indicates that the figure represents data accumulated only in the first half of that year.

Box 1. How to search and retrieve EST data

Data from the WashU-Merck cDNA project is available from three primary sources at the following World Wide Web addresses:
http://www.ncbi.nlm.nih.gov/
(sequences, contact information, homology data, etc.; mirrored at **http://www.ebi.ac.uk/**)

http://genome.wustl.edu/est/esthmpg.html/
(trace data, library information, experimental protocols)

http://www-bio.llnl.gov/bbrp/image/image.html
[integrated molecular analysis of genomes and their expression (I.M.A.G.E.) consortium library arrays]

For other forms of access and additional information, send email to: info@ncbi.nlm.nih.gov, est@watson.wustl.edu, info@image.llnl.gov or telephone +1 301 496 2475 and ask for the service desk.

combined with those derived from previous and ongoing efforts by other groups, now account for more than half of the sequence records in GenBank (Fig. 1). These data are being heavily used by biomedical researchers: so far, the EST division of GenBank has processed more than 150 000 queries consisting of homology searches, email retrievals, anonymous file-transfer protocol (ftp) transfers and access via the World Wide Web (WWW; Box 1).

Plans to produce the complete sequence

As previously mentioned, the focus in the first phases of the genome program was on mapping and technology development[1]. The rationale was that detailed physical maps were the essential infrastructure to provide 'sequence-ready' genomic DNA clones, and that new sequencing technologies would dramatically reduce the cost and increase the speed at which DNA could be sequenced. At some point, it was envisioned, progress in both areas would be deemed sufficient to embark on the final phase of the genome program, namely DNA sequencing on a massive scale. In December 1994 a meeting was held that brought together the directors of all of the NIH-funded Genome Science and Technology Centers to take stock of the situation and propose appropriate action. Robert Waterston argued that sufficient experience and economies of scale had been achieved with existing sequencing technology (gel-based, fluorescence detection), that concerted human genomic sequencing should be commenced without further delay, and that the project could be completed by 2001, four years ahead of schedule. This bold proposal incubated in the hearts and minds of the genome community over the next few months, and by May 1995 (the occasion of the annual exodus of these researchers to the Genome Mapping and Sequencing meeting at Cold Spring Harbor), this community sensed that, as Maynard Olson put it, the climactic phase of the Human Genome Project was about to begin.

Concepts and costs

Maynard Olson, of the University of Washington in Seattle, is considered the sage of the genome program. In the early 1980s, he proposed to physically map a eukaryotic genome by constructing restriction maps of all of the chromosomes of the yeast *Saccharomyces cerevisiae*[10]. This idea was greeted with considerable skepticism at the time, to put it mildly. But now that we are only six months away from a *complete genomic sequence* (approximately 12.5 megabases) of this organism, one can appreciate the delphic proportions of Maynard's vision. Maynard also pioneered the use of yeast artificial chromosomes (YACs) as cloning vectors[11] and co-invented the concept of 'sequence-tagged sites' (STSs), an idea that revolutionized physical mapping[12].

Both Maynard and John Sulston (of the Sanger Centre at Hinxton Hall, Cambridge, UK) were called upon to present some closing remarks at the Cold Spring Harbor meeting. Sulston, together with Waterston, is sequencing the 100 megabase genome of the nematode *Caenorhabditis elegans* (projected for completion by the end of 1998). Sulston shares Waterston's view that human genome sequencing should proceed without delay and that the project can be completed in five years for a total expenditure of $300 million. This assumes that costs can be reduced from the current price of 49 cents down to 10 cents per base (although this is controversial[13]).

In his closing remarks, Maynard gave his opinion on what the final, irreducible product of the genome project would be. There are three components: (1) *the genetic map* (because recombination frequencies can never be reduced to a physical distance in base pairs); (2) *the sequence map* (i.e. the complete sequence or a sequence with so few remaining gaps and errors that it can be considered complete for all practical purposes); and (3) *the peer-reviewed scientific literature* (traditional or electronic, with sequences acting as specific retrieval keys and cross-references).

At these three points, the genome project and the rest of biology will merge. At 10 cents per base, annotation of the sequence by its producers will be minimal and largely automated, creating the opportunity for other biologists to make discoveries and to provide detailed interpretations, based upon sequence analysis and experimentation. The complete genomic sequences of *Homo sapiens* and several other organisms will provide a whole new framework for biological investigation into the 21st century.

References

1 Collins, F. and Galas, D. (1993) *Science* 262, 43–46
2 Collins, F. S. (1995) *Nat. Genet.* 9, 347–350
3 La Spada, A. R., Paulson, H. L. and Fischbeck, K. H. (1994) *Ann. Neurol.* 36, 814–822
4 Adams, M. D. *et al.* (1991) *Science* 252, 1651–1656
5 Boguski, M. S., Tolstoshev, C. M. and Bassett, D. E. (1994) *Science* 265, 1993–1994
6 Tugendreich, S. *et al.* (1994) *Hum. Mol. Genet.* 3, 1509–1517
7 Papadopoulos, N. *et al.* (1994) *Science* 263, 1625–1629
8 Brody, L. C. *et al.* (1995) *Genomics* 25, 238–247
9 Boguski, M. S. and Schuler, G. D. *Nat. Genet.* (in press)
10 Link, A. J. and Olson, M. V. (1991) *Genetics* 127, 681–698
11 Burke, D. T., Carle, G. F. and Olson, M. V. (1987) *Science* 236, 806–812
12 Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science* 254, 1434–1435
13 Marshall, E. (1995) *Science* 268, 1270–1271

MARK S. BOGUSKI

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

Note: an electronic copy of this article can be found on the WWW at **http://www.ncbi.nlm.nih.gov/**