A Cell Biological Perspective on Genome Research

J. Richard McIntosh and Robert R. West

Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347

The genomes of several organisms are currently being sequenced at impressive rates, but some scientists have asked whether the final products will be worth the effort. Here we argue from both the history of structural biology and some encouraging initial results that the answer is a resounding yes.

There have been five overlapping stages in the growth of structural biology. The ancients learned much about animals and plants through comparative anatomy. A systematic study of animal anatomy through dissection began in the 16th century. Microscopic anatomy began in the mid-17th century, and by the early 1900's there were detailed descriptions of a wide diversity of organs and organisms. 1940–1970 saw a comparable flowering of knowledge about cellular fine structure and macromolecular assembly, thanks in part to the electron microscope. The details of macromolecular structure were, however, largely invisible without crystallography, and the fruits of this technology are still ripening.

Each of the above stages in the development of “anatomy” has depended on a new method or approach whose availability led to rapid growth in structural information at a novel level of size or detail. An analogous situation now pertains to determination of DNA sequence. While the technology for sequencing proteins and nucleic acids has been available for decades, recent improvements have made it possible to learn the primary structure of DNA comparatively quickly. As a result, there has been a huge increase in the number of labs that are sequencing DNA and an almost incredible increase in the rate of sequence acquisition. The availability of computers fast enough to deal expeditiously with all the resulting data has permitted efficient analyses and comparisons of DNA sequences and of the polypeptides predicted from them, so the flow of information from acquisition to utilization for biological comparisons has become both rapid and objective. The resulting catalogues of genes and gene products are facilitating a new version of all the discriminations that have traditionally made anatomical information useful; clearly the sixth age of anatomy is at hand.

During each stage in the development of structural biology, the practitioners have been delighted in their new view of nature. Others in the scientific community have generally been less enthusiastic, finding structure itself of little interest unless it can be related to function. When this goal has been achieved, then anatomical detail has been invaluable for elucidating biological mechanisms. For example, a mitochondrion makes no sense until its enzymes are localized relative to its membranes; a catalytic mechanism is usually revealed by comparing the structure of its active site with different substrate analogues bound.

When study of the human genome was introduced essentially as a grand project in anatomy, many of us were highly skeptical of its value. Doubt arose in part because sequencing for its own sake seemed so mindless, in part because the human genome seemed likely to contain a large amount of DNA whose sequence would convey little information, and in part because it seemed that allocations to support this work would likely divert resources from the already competitive RO1 grants that support more conventional biomedical science. Several things have happened, however, to change the face of things.

Genome Research Has Been Organized in Two Useful Hierarchies

Rather than choosing a few pieces of the human DNA and beginning to sequence, genome researchers have organized their work in two important ways: (1) they have focused on model organisms with small genomes where there was already sufficient genetic and molecular information to give context and meaning to sequence data; and (2) genome work on each organism has begun with the generation of physical maps, so sequence information can be related to genetic maps and studies of mutant phenotype.

One of the first major sequencing efforts has been with the yeast *Saccharomyces cerevisiae*, whose genome was known to be small (~12.5 megabases [Mb]). Many yeast genes had already been identified by mutation, and several regions of this genome had been analyzed by saturation mutagenesis, which identifies most genetically recognizable loci. The frequency and size of intervening and spacer DNA sequences were known to be small, and physical maps of the genome had already been constructed. There were also methods for manipulating this genome experimentally, e.g., gene replacement by homologous recombination, so one could anticipate that sequence information could be related to the functional consequences of mutation. When systematic sequencing began, the 16 *S. cerevisiae* chromosomes were undertaken by several yeast labs throughout the world. As of October 1995, over three-quarters of the
The Department of Energy’s Biological and Environmental Research Program is manageable (8). Meanwhile, DNA sequencing has been ongoing in two centers, and the rate of new finished sequence (six determinations to minimize the chances of error) now exceeds one cosmid/day (7). Completion of this genome was sequenced (11) and completion of the project is expected by the beginning of 1996. For detailed information about the current state of work on this and other genomes, please see the Internet addresses given in Table I.

Genome research on more complex model organisms has again begun with the formation of detailed physical maps and limited genomic sequencing. For the nematode, *Caenorhabditis elegans* (genome size ~100 Mb), a collection of “cosmid” clones has been constructed, each of which carries tens of kilobases (kb) of exogenous DNA. Many of these sequences are from contiguous loci, yielding an ordered library that covers much of the genome. There is now a *C. elegans* genomic library in yeast artificial chromosomes (YACS) as well; each of these clones contains hundreds of kb of DNA, so the number needed to cover a larger genome is manageable (8). Meanwhile, DNA sequencing has been going on in two centers, and the rate of new finished sequence (six determinations to minimize the chances of error) now exceeds one cosmid/day (7). Completion of this genome is expected by the end of 1998 (Table I).

For model organisms with larger genomes, like fruit flies (genome 1.5 × that of *C. elegans*) and mice (30 × the nematode), the above strategy is being employed in an expanded form. Both YAC and cosmid libraries are now available to help map genes to specific pieces of DNA. Indeed, much of the recent work on the human genome has been focused not on DNA sequencing, but on detailed chromosomal mapping. Microsatellites (short segments of repetitive DNA) are being used to identify unique molecular markers with better than 1 centimorgan resolution, and physical maps are being generated with sequence-tagged sites (STSs) on YAC and cosmid clones. This work has significantly refined results from the classic methods of cytogenetics, like Q or G banding, as well as older molecular mapping methods, like restriction fragment length polymorphisms. The result is both a useful map of genes related to human diseases and tools that will be necessary for molecular studies of particular chromosomal regions. Systematic sequencing of the human genome is just now getting underway (Table I).

### Table I. World Wide Web Addresses for Information about Specific Genome Projects

<table>
<thead>
<tr>
<th>Genome of Interest</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td><a href="http://genome.wustl.edu">http://genome.wustl.edu</a></td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td><a href="http://genome/wustl.edu/gsc/gschmpg.html">http://genome/wustl.edu/gsc/gschmpg.html</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.sanger.ac.uk/~sjl/C.elegans_Home.html">http://www.sanger.ac.uk/~sjl/C.elegans_Home.html</a></td>
</tr>
</tbody>
</table>

**National Center for Biotechnology Information**


**The Dept. of Energy’s Biological and Environmental Research Program**

- [http://www.er.doe.gov/production/oher/](http://www.er.doe.gov/production/oher/) (and then either hug_top.html (for human) or mgh_top.html (for microbial))
- [http://gdbwww.gdb.org/omim/docs/omintop.html](http://gdbwww.gdb.org/omim/docs/omintop.html)
- [http://gdbwww.gdb.org/genomic_links.html](http://gdbwww.gdb.org/genomic_links.html)

**Online Mendelian Inheritance in Man**

- [http://gdbwww.gdb.org/omim/docs/omintop.html](http://gdbwww.gdb.org/omim/docs/omintop.html)
- [http://gdbwww.gdb.org/genomic_links.html](http://gdbwww.gdb.org/genomic_links.html)

**General List of Many Other Genome Projects**

- [http://genome.wustl.edu/gsc/gschmpg.html](http://genome.wustl.edu/gsc/gschmpg.html)
- [http://www.sanger.ac.uk/Nsjj/C.elegans_Home.html](http://www.sanger.ac.uk/Nsjj/C.elegans_Home.html)
- [http://genome.wustl.edu/gsc/gschmpg.html](http://genome.wustl.edu/gsc/gschmpg.html)
- [http://gdbwww.gdb.org/omim/docs/omintop.html](http://gdbwww.gdb.org/omim/docs/omintop.html)
- [http://gdbwww.gdb.org/genomic_links.html](http://gdbwww.gdb.org/genomic_links.html)

**Sequence Data Are Telling Us Surprising Things about Genome Structures and Functions**

Information from *S. cerevisiae* and *C. elegans* is already telling us interesting things about genome structure and organization. In addition to the obvious features, like the number of genes in a given amount of DNA, the statistics of gene size, and the distribution of genes among major families defined by sequence, things are coming to light that few would have predicted. The sequences from regions subjected to saturation mutagenesis have revealed many more genes than were found in searches for mutants with phenotypes. For example, genetic estimates predicted ~3,000 essential genes in *C. elegans* (3), but estimates from genome sequence suggest ~13,000 (10). Presumably, the inactivation of the additional genes by mutation is either irrelevant for growth as we assess it in the laboratory or is “covered” by the functions of other, unmutated genes. The work has also revealed novel features of genome organization: there are genes within introns, genes that appear to encode giant proteins, and clusters of genes with unknown functions (12). While more than half the genes so far discovered by sequencing are of unknown function, ~40% of the genes from *C. elegans* (7), ~55% from *S. cerevisiae* (11), and 58% from the prokaryote, *H. influenzae* (6) encode proteins whose sequences are similar to the products of genes already identified in other organisms. Clearly, there is much still to do in tracking down and analyzing unknown genes, but it is encouraging that so much sequence information can already be understood in some functional context and related to processes that are under study elsewhere.

Large scale studies of gene sequences have also opened up a new approach to genetic comparisons. Many genes from humans and other organisms have been partially characterized by sequencing only a small part of a cDNA to yield an “expressed sequence tag” (EST).1 These have

---

1. Abbreviations used in this paper: EST, expressed sequence tag; Mb, megabase; YAC, yeast artificial chromosome.
have now been identified in model organisms for most of
significant similarity, thereby identifying a mammalian gene
genes identified by mutation are compared with those in
the dbESTs, it has sometimes been possible to find a sig-
Mysterious Aspects of Organismic Biology Are Likely to
processes and pathologies in higher eukaryotes; and (b) it
positional cloning (1, 2, 4, and Table I). There are two im-
the N50 human diseases genes that have been mapped by
information about DNA sequence and organization is
likely to play big dividends, not only in knowledge of ge-
structure but in the saving of countless hours of mo-
lar drudge work. A well set up sequencing lab can pro-
produce ~2 kb of finished sequence per worker per day, while
labs that are sequencing by hand are doing well if they get
1/10th that amount. Moreover, the cost per base drops by
~10-fold when the work is done in a lab with proper
and experience. The current rate of sequence
collection will probably increase as technology improves,
implying that we should continue with significant invest-
ments in the study of genome structure for the near future.
While those of us doing experiments may resent the expen-
diture of money that might have helped us to stay funded,
the resulting information will almost certainly be worth the
investment, and the approach will probably save money.

Even with All This Success, There Are Problems That
Must be Addressed by Our Scientific Community

Enthusiasm for genome research does, however, lead to
three problems that our community must address: (1) since resources are limited, the dollars spent on genome research
must be consistent with our current funding constraints.
The 1995 investment in extramural research by the National
Center for Human Genome Research (NCHGR) was
~$114 million, while the Departments of Energy and of
Agriculture spent ~$69.5 and $3.8 million, respectively, on
work. The Small Business Administration spent
$3 million, and comparatively small sums that are hard to
determine came from other Institutes at the National In-
stitutes of Health (NIH) and from the National Science
Foundation. These investments in aggregate are small rela-
tive to the whole NIH budget (~$12 billion for 1995), but
they are significant compared with the budget of the Na-
tional Institute for General Medical Sciences (just over $900
million), the agency upon which many of us rely. Given the
importance of genome work and given the investments by
other institutes and agencies in basic biomedical research,
broadly defined, the current balance seems to us reason-
able, but a greater investment in genome work would tip
the balance.

(2) The results from sequencing are first available to the
sequencers themselves. With genome labs funded by in-

McIntosh and West A Cell Biological Perspective on Genome Research
dustries, as well as the government, it is critically important to maximize the free and rapid flow of sequence information into the public domain, otherwise knowledge will be either wasted or improperly used. Some sequencing labs have been models of efficiency in making their data available, but others have not. The NCHGR has stipulated that sequence data from the labs they fund must be made public within 6 mos, but some labs in foreign countries are slower than that, and some industries attach strings to sequences they share. We believe that efforts should be made to minimize the time to public access in order to maximize the rate at which good science can be done. Moreover, the push by both individuals and companies to profit from sequences and/or to restrict their use is a problematic trend that will take some years to sort out in the courts. In the interim we must work to minimize the chances that profit motives retard scientific progress.

(3) In the future, sequencing labs will certainly provide important information, e.g., the genome structures for model organisms, humans, and important pathogens, as well as the mapping of human disease genes. With such longevity, however, sequencing labs may tend to become entrenched, so their lifetimes exceed their utility. It is impossible to say now when enough genome work will have been done, but there is an obvious analogy with a learning curve. Sequencing centers may want to go on from human to aardvark, caiman, and dodo, expanding our knowledge of evolution, phylogeny, and extinction. The community of biologists will have to recognize when new knowledge from new sequences has begun to plateau and rein in the rate of spending. Sequencing programs must be structured so they can shrink as their value diminishes, giving way to more experimentally oriented biology.

That said, it is clear that future biologists will be working in an environment defined by a wondrous wealth of information about genome structure. It is mind boggling to think of the ways in which our experimental lives will be changed as a result. No field of biology will be untouched, and a whole new generation of experimental approaches will likely emerge to help justify the investments that are now being made in sequence acquisition and analysis.

Received for publication 1 September 1995 and in revised form 23 October 1995.

References