Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 5506–5507, May 1997

Commentary

Molecular linguistics: Extracting information from gene and protein sequences

David Botstein and J. Michael Cherry

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305

Highly controversial only a few short years ago, the human genome project has spawned a vigorous new science called genomics. A decade ago a National Research Council (NRC) Report (1) came out with a compromise 15-year plan to produce comprehensive genetic and physical maps of the human genome, the sequence of the human genome and, surprisingly to many, the sequences of the genomes of a number of so-called "model genetic organisms," generally understood to comprise, at least, a bacterium (*Escherichia coli*), a yeast (*Saccharomyces cerevisiae*), a nematode worm (*Caenorhabditis elegans*), a fruitfly (*Drosophila melanogaster*), and a rodent (*Mus musculus*). The rationales given for the necessity to sequence the genomes of model organisms were quite diverse, and skeptics abounded, suspecting that tradition and politics might have played some role in this potentially diversionary recommendation.

At the heart of all the NRC recommendations was the understanding that the sequence of the human genome would require interpretation. Biological experimentation was seen as the only realistic means of interpretation. The experimental tractability of the model organisms, it was hoped, would facilitate elucidation of the functions of genes and proteins. Taking advantage of the slow rate of protein evolution, the understanding obtained in the model organisms might allow reliable inferences concerning possible roles of the cognate human genes and proteins (see ref. 2 for an example of this argument at that time). In short, the model organisms were to serve as the "Rosetta Stone" that would allow us to understand the human genome sequence, just as the original Rosetta Stone allowed decipherment of the ancient Egyptian hieroglyphics. It was understood that the requisite sequence comparisons and sequence analyses would absolutely require development of algorithms, software, and computation facilities well beyond what then was available. Indeed these needs drove the invention of another new field, now usually called bio-informatics.

Today, at the midpoint of the 15-year plan, the science of genomics is well established. It boasts more than a few dedicated journals, ranging from the archival to the determinedly trendy, scores of meetings every year, an National Institutes of Health institute of its own (the National Human Genome Research Institute), and even a handful of start-up companies organized specifically to exploit the commercial potential of this newest of sciences. A solid infrastructure is in place for molecular and genetic (i.e. linkage and association) studies of the human genome. The databases bulge with more than 20,000 mapped polymorphic DNA markers useful in genetic mapping and more than 30,000 sequence-tagged-sites (STSs) (3–5) suitable for physical mapping using yeast artificial chromosomes (6) or, more conveniently, radiation hybrid mapping (7). A single investigator today can genetically map and even hope to positionally clone a gene in a reasonable time, a task requiring dozens of investigators and many millions of dollars just a few years ago. Thousands of human disease genes have been mapped and hundreds of thousands of short segments of expressed human genes (expressed-sequence tags, or ESTs) have been sequenced (8, 9). On the order of 100 human disease genes have been positionally cloned, beginning with nothing more than evidence of a genetic etiology. The reader is referred to on-line databases devoted to human gene mapping (Whitehead/Massachusetts Institute of Technology, Centre d'Étude du Polymorphisme Humain (Paris), Cooperative Human Linkage Center, TIGR Human cDNA Database, Washington University/Merck, Stanford Human Genome Center, and Genome Data Base; Table 1) for up-to-date information and documentation.

In the model organisms effort, the sequences of a number of bacterial species became available; Table 1 lists databases in which these sequences can be found. *Hemophilus influenzae* (10) was first, and several were finished, including some *Archaea* (11), well before the *E. coli* sequencers finally got the job done. The complete sequence of the first eukaryote, *Saccharomyces cerevisiae*, appeared on the Worldwide Web a year ago (ref. 12, see the *Saccharomyces* Genome Database and Yeast Genome from MIPS sites given in Table 1). Consultation of the relevant Internet sites (Table 1) will confirm that the nematode worm is more than half done and *Drosophila* is moving right along.

What of bio-informatics? If anything, this has been an even bigger success than genomics. Statistics cited in the paper by Mushegian *et al.* in this issue of *Proceedings* (13) attest to this. In a sample of 70 positionally cloned (and sequenced) human disease genes, they found that 36% had orthologs (i.e. genes encoding proteins likely to be identical in function) in *C. elegans*, despite the fact that only half the worm genome had been sequenced at the time of the comparison. More than 60% of the disease genes had close homologs for at least one of their encoded protein domains in yeast. Mushegian *et al.* also cite the remarkable fact that 29 genes have been cloned by functional complementation of yeast genes, which again illustrates that the rate of evolution of proteins has been slow enough to permit functional interchangeability even after divergence times measured in the billions of years.

The paper of Mushegian *et al.* is notable in another way: it contains no experiments, and all of its results are from analysis of molecular sequences using computational methods, algorithms, and even words (e.g. "ortholog" and "paralog") not known to the NRC committee. Many of the authors belong to an already indispensable organization (the National Center for Biotechnology Information, or NCBI; see also Table 1) consisting entirely of bio-informaticians or, as we would prefer to think of them, molecular linguists. As the steward of GenBank, NCBI has illustrated brilliantly the reality that simple storage of sequence information is grossly inadequate to the needs of the scientific community—organization and assimilation of the data (in a word, curation by experts) is at some point indispensable.

The rise of genomics and bio-informatics has had another consequence: the increasing dependence of all biology on results available only in electronic form. Most of the useful

0027-8424/97/945506-2$2.00/0
PNAS is available online at **http://www.pnas.org**.

Table 1. Some DNA sequence and genomic databases.

Database	Web address
Human	
CEPH Généthon Integrated Map	http://www.cephb.fr/bio/ceph genethon map.html
The Cooperative Human Linkage Center (CHLC)	http://www.chlc.org/
MIT Center for Genome Research	http://www-genome.wi.mit.edu/cgi-bin/contig/phys_map
Stanford Human Genome Center	http://shgc.stanford.edu/
Washington University-Merck Human EST Project	http://genome.wustl.edu/est/esthmpg.html
The TIGR Human cDNA Database	http://www.tigr.org/tdb/hgi/hgi.html
National Center for Biotechnology Information (includes GenBank)	http://www.ncbi.nlm.nih.gov/
The Genome Database	http://gdbwww.gdb.org/
XREFdb, Cross-referencing Model Organisms	http://www.ncbi.nlm.nih.gov/XREFdb/
Model organisms	
Saccharomyces Genome Database	http://genome-www.stanford.edu/Saccharomyces/
Yeast Genome from MIPS	http://speedy.mips.biochem.mpg.de/mips/yeast/
The *C. elegans* Genome Project	http://www.sanger.ac.uk/worm/C.elegans_Home.html
Berkeley *Drosophila* Genome Project	http://fly2.berkeley.edu/
FlyBase	http://morgan.harvard.edu/
Mouse Genome Informatics	http://www.informatics.jax.org
Arabidopsis thaliana Database	http://genome-www.stanford.edu/Arabidopsis/
MaizeDB	http://teosinte.agron.missouri.edu/
Archaea and eubacteria	
The *Mycoplasma genitalium* Genome Database (MGDB)	http://www.tigr.org/tdb/mdb/mgdb/mgdb.html
The *Mycoplasma pneumonia* Genome Project	http://www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html
The *Methanococcus jannaschii* Genome Database (MJDB)	http://www.tigr.org/tdb/mdb/mjdb/mjdb.html
The *Haemophilus influenzae* Rd Genome Database	http://www.tigr.org/tdb/mdb/hidb/hidb.html
CyanoBase, The Genome Database for *Synechocystis* sp.strain PCC6803	http://www.kazusa.or.jp/cyano/cyano.html
SubtiList Web Server	http://www.pasteur.fr/Bio/SubtiList.html
E. coli Genome Project	http://www.genetics.wisc.edu/index.html
MycDB, The Integrated Mycobacterial Database	http://www.biochem.kth.se/MycDB.html

genomic data, notably genetic maps, physical maps, as well as DNA and protein sequences, are available only on the World-wide Web. Not only are these data unsuited, because of their very bulk, to print media, they are of very little use in print because this kind of information can only be truly assimilated, used, and appreciated with the aid of computers and software.

This trend is rapidly being extended to nonsequence data such as mutant phenotypes, gene expression patterns, and gene interactions, whose complexity defies simple description. In all such descriptions, there are at least as many data points as there are genes in an organism, meaning that we can look forward to data sets comprising literally millions of data points. Of necessity, results will only be summarized in print; the real data will reside as binary strings on electronic media. As a result, databases of genomic information for a variety of organisms have been organized (i.e., *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Methanococcus jannaschii*, *Haemophilus influenzae* Rd, Cyanobacteria, *Bacillus subtilis*, Mycobacteria, yeast, worm, *Drosophila*, *Arabidopsis*, maize, mouse, and human; see Table 1).

To conclude, at its halfway point the human genome project already has transformed biological science. We are now in a period of unification among sub-fields of biology too long fractured along organismal lines. There is no longer any doubt that the model organism sequences are effectively providing information about human genes and proteins to a level of detail and specificity beyond the dreams of the most optimistic members of the NRC committee. The meaning of the sequence of the disease genes is routinely deciphered using information from yeast and worms. We all have had to become molecular linguists, to learn to respect the unity of biology. We can reflect on our good fortune that Mother Nature has given us, through the slow pace of protein evolution, such a good Rosetta stone.

1. Alberts, B. M., Botstein, D., Brenner, S., Cantor, C. R., Doolittle, R. F., Hood, L., McKusick, V. A., Nathans, D., Olson, M. V., Orkin, S., Rosenberg, L. E., Ruddle, F. H., Tilghman, S., Tooze, J. & Watson, J. D. (1988) *Report of the Committee on Mapping and Sequencing the Human Genome* (Board on Basic Biology, Commission on Life Sciences, National Research Council) (National Academy Press, Washington, DC).
2. Botstein, D. & Fink, G. R. (1988) *Science* **240,** 1439–1443.
3. Green, E. D. & Olson, M. V. (1990) *Science* **250,** 94–98.
4. Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., *et al*. (1995) *Science* **270,** 1945–1954.
5. Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J. & Weissenbach, J. (1996) *Nature (London)* **380,** 152–154.
6. Bellanne-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., *et al*. (1992) *Cell* **70,** 1059–1068.
7. Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J. F., Dib, C., Auffray, C., Morissette, J., Weissenbach, J. & Goodfellow, P. N. (1996) *Hum. Mol. Genet.* **5,** 339–346.
8. Adams, M. D., *et al*. (1995) *Nature (London)* **377,** 3–174.
9. Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., *et al*. (1996) *Genome Res.* **6,** 807–828.
10. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al*. (1995) *Science* **269,** 496–512.
11. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al*. (1996) *Science* **273,** 1058–1073.
12. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. & Oliver, S. G. (1996) *Science* **274,** 546.
13. Mushegian, A. R., Basset, D. E., Jr., Boguski, M. S., Bork, P. & Koonin, E. V. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 5831–5836.