Mathematics and the Human Genome

The genetic material in our chromosomes is our genome. Chromosomes are long DNA molecules, that is, double-helices consisting of long paired polymers of the bases Adenine, Cytosine, Guanine, and Thymine bound together in sequence, attached to sugar-phosphate backbones. The twenty-two regular human chromosomes and the X and Y chromosomes that are a human genome comprise about three billion base pairs of DNA. Genes are segments of chromosomes of differing lengths, with the typical human gene being a few thousand base pairs in length. There are believed to be thirty to forty thousand genes in the human genome, making up about 3 percent of the total base pairs. Identifying all the genes and understanding their function is the major task of genomics.

For most of the twentieth century mathematics played a minor role in genomics, but its role has expanded greatly over the last twenty years. With the ability to sequence complete genomes, the expansion accelerated. Why? What kinds of mathematics arise here? Why might mathematicians find the field interesting? What are the opportunities? Are there barriers to the involvement of mathematicians in genomics?

Why mathematics and the human genome? The answer is: data. Early genetic data were either counts, typically of genotypes at a few loci or discrete phenotypes, or measurements of continuously varying phenotypes. We now have very large quantities of genotypic data on a genome-wide scale: many loci and many phenotypes on many people. There is a vast and growing amount of DNA, RNA, and protein sequence data in public databases, much of it on a genome-wide scale (e.g., billions of bases of human DNA sequence). Human sequence variation data is now accumulating at the level of the individual. The generation of massive amounts of messenger RNA expression data on a genome-wide scale has just begun; similar data at the protein level is the next great growth area. The amount can hardly be imagined: In principle we could generate data on the expression of all genes in all cell types under all conditions for individuals of all genotypes.

What kinds of mathematics? This is a hard question to answer briefly, but let's consider the recent human genome sequencing and the initial follow-up. Several different human genomes were copied many times and broken into millions of small fragments. Fairly accurate but imperfect DNA sequence was obtained for each of these fragments. The algorithms which turned laser-scanned electrophoretic traces into sequences of A's, C's, G's, and T's, along with quality scores, involved dynamic programming, machine learning, Fourier analysis, and other signal processing techniques. Then came the massive assembling of these millions of fragments into large pieces. There a wide range of combinatorial, graph-theoretic, probabilistic, and statistical tools were combined into algorithms which ran for thousands of hours on some of the world's largest computers.

After the genome was assembled—finished data in one sense, raw data in another sense—the task of interpretation began. Identifying the 3 percent comprising the genes was the first priority. Protein-coding genes have a modular structure, including several highly but not completely conserved signals, and DNA sequences that have small but consistent compositional differences from nongenic DNA. The task was to use this information, together with that on known genes in sequence databases, to identify new genes. The best current methods use hidden Markov models. Stochastic context-free grammars play an analogous role in identifying RNA genes. Comparing new DNA sequences to those in databases involves string search algorithms (e.g., finite-state machines), sequence alignment algorithms (more dynamic programming), and statistical scoring. When we have predicted genes, we want to learn about the proteins they encode. Alignment, hidden-Markov models, clustering, and phylogenetic tree algorithms all play roles here. Having predicted proteins we next want to learn their structure and function. How do they fold? When and where are they expressed? With which other proteins do they interact? In which pathways are they involved? More questions, more data, more mathematics.

If your interest is in algorithms, combinatorics, machine learning, optimization, probability, statistics, knot theory, to mention the obvious areas, there is a good chance that there are problems in genomics to interest and challenge you. If you like biology as well, the opportunities are endless. What about barriers? They are the usual ones when boundaries are crossed: language and culture. For mathematicians getting into this area there is much to learn. And as in any application of mathematics, the problems mathematicians want to study may not be those biologists think important. The genome community may not embrace your contribution in the way you wish. But none of this is new or unique to genomics. The challenges and satisfaction are there for anyone wishing to get involved; the benefits far outweigh the risk of investing time and effort to find out what is needed.

—Terry Speed
University of California at Berkeley

About the Cover

April is Mathematics Awareness Month (MAM). The theme for MAM 2002 is "Mathematics and the Genome". Visit `http://www.mathforum.org/mam/02/` for more information and to download the poster. Images: *Protein bound to DNA*, courtesy of Akinori Sarai, Ryken Bioinformatics Database, `www.rtc.riken.go.jp/jouhou/image/dna-protein/all/N1hjb.gif`; *DNA, the molecule of life*, courtesy of U.S. Department of Energy Human Genome Program, `www.ornl.gov/hgmis/`.

—Annette Emerson, AMS Public Awareness Officer