

■ MONUMENTAL SCIENTIFIC ADVENTURE

THE HUMAN GENOME PROJECT AND INFORMATICS

KAREN A. FRENKEL

Many Human Genome Initiative (HGI) researchers believe they are shaping the future. To Nat Goodman, a senior research scientist at the MIT-affiliated Whitehead Institute in Boston, the HGI is "one of the monumental scientific adventures of history." Goodman is creating a genetic and physical map database of the mouse genome. "When we are able to decode, in a routine way, the genes that make life possible," he says, "that will be one of the milestones in science. I can't imagine why everyone's not jumping up and down to work on this, because this is only going to happen once."

Los Alamos National Laboratory staff computer scientist Robert Pecherer is part of a team in the Theoretical Biology and Biophysics Group at work on a Human Genome Information Resource. Its goal is to develop information management and analysis tools for physical mapping data. The project is aimed at linking databases from two or more biological disciplines with the long-term objective of extending similar research to other related data sets such as nucleotide sequences and genetic maps. For Pecherer, the "driving forces behind the HGI have to do with the fact that we're entering a new century and the technological groundwork for what will be important then will be genetically based." Looking at the range of govern-

ment agencies involved, Pecherer probes for their common interests. The reasons for preventing genetically linked diseases like muscular dystrophy, cystic fibrosis, sickle cell anemia, and Tay-Sachs are easily justified, he says. The Department of Agriculture is also interested because the U.S. used to be the leading goods exporter and could return to that role with genetically engineered crops. "Despite the fact that we think of the U.S. as being really high tech, we made a lot of money selling food in the 50s, 60s, 70s and even into the 80s," Pecherer comments.

But Pecherer believes there is more to the HGI than medical and agricultural motivations. The knowledge of biotechnology and bioengineering will make one country more successful than the other, Pecherer claims. It is not just a matter of whether its automobiles have better paint jobs, or its workers can assemble stereos more cheaply. "So one of the reasons for pursuing the HGI here aggressively is that the technology that will flow from it will make the U.S. a more competitive member of the global business community in the next century."

The Human Genome Initiative

The large-scale Human Genome Initiative (HGI), which began about three years ago, is aimed at determining the location of the estimated 100,000 human genes that comprise the entire human genome. The purpose is also to ana-

lyze the structure of DNA, the famous double helical strand of base pairs discovered by James D. Watson and Francis Crick in 1953. In parallel with the human DNA effort, the DNA of model organisms, such as bacteria, yeast, fruit flies, and mice, is being studied to provide comparative information necessary for a better understanding of how the human genome functions. Human genes are distributed over 24 pairs of chromosomes; 22 pairs of autosomes and the two sex chromosomes, X and Y. They exist in cell nuclei and contain protein and DNA. The order of the base pairs, that is, their sequence, determines the information content of a particular gene or of a piece of DNA. That order is the genetic code. Thus one goal is to generate "maps" of varying resolutions. Biologists literally divide and conquer this enormous task by breaking up the DNA into fragments, sequencing and mapping those sections, and trying to reassemble them.

With approximately 100,000 genes ranging from 2,000 to 2 million base pairs, the human genome is estimated to be 3 billion base pairs. In 10-point type, like the type on this page, the genetic code would be a sequence of letters 5,000 miles long. In contrast, the much simpler bacterium *Escherichia coli* (*E. coli*), which has 4.8 million base pairs, is only seven miles long in 10 point type. Under investigation for many years and the most completely known organism, this crea-

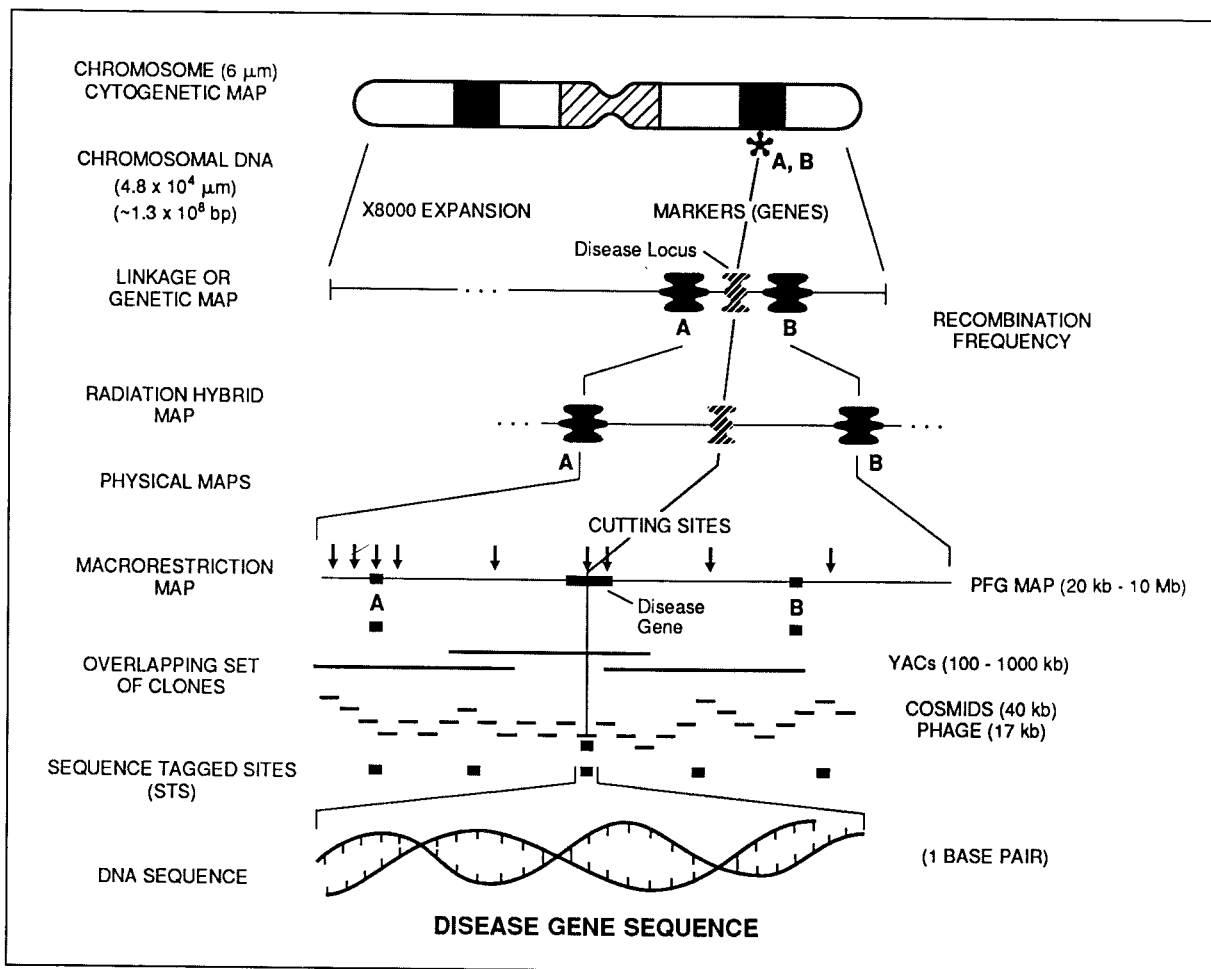
nally, it would cost too much to "read" the byte sequence of every one of these fragments directly, so first you must determine the minimum spanning set of fragments. This requires you to devise and perform some partial characterization of each fragment that can be used in an exhaustive set of pairwise comparisons to generate a 60-billion-

entry probability-of-overlap matrix from which you can attempt to deduce the minimal spanning set. All of your characterization, comparison, and assembly algorithms must take into account the possible occurrence of random or systematic errors at every point.

differ by about one in 1,000 nucleotides. So another complicating factor is how to—or whether or not to—use the *average* of those several trillion copies as the input into this reverse engineering project. In the classic medical text *Gray's Anatomy*, illustrations of body parts are idealized composites, approximately true of everybody and precisely true of no one. Although essential learning aids at an introductory level, such approximations are inadequate preparation for detailed work like surgery. Says Robbins, "If you really need to know how human hands are built—if you're a surgeon about to operate on a human hand—you don't look at *Gray's Anatomy*. You look at advanced works that emphasize not the similarities, but the differences and variations from hand to hand."

As if assembling the sequence were not enough of a genuine challenge, Robbins notes that the millions of ordered nucleotides in potentially a 3.5 gigabyte database are not themselves the end product. "That's the raw input for a tremendous reverse engineering project. And that's just one copy. You have several trillion, slightly different versions scattered in the cells in your body, and so do I, and so does everybody else," he says. Humans

Chromosome-mapping resolutions
Lower-resolution maps include physical maps of the human genome showing the locations of landmarks on the DNA, such as restriction enzyme sites, and genes. The sequence map—the highest-resolution map possible and the ultimate physical map—shows the actual order of nucleotides in a nucleic acid or the order of amino acids in a protein.



Source: Human Genome 1989-90 Program Report, Mar. 1990, U.S. Department of Energy Office of Energy Research, Office of Health and Environmental Research.