Scientific American, December 1999, pages 50–55.

Genomics and the Future

The study of all the genes of various organisms — a field called genomics — will yield answers to some of the most intriguing questions about life

by Francis S. Collins and Karin G. Jegalian

When historians look back at this turning of the millennium, they will note that the major scientific breakthrough of the era was the characterization in ultimate detail of the genetic instructions that make a human being. The Human Genome Project—which aims to map every gene and spell out letter by letter the literal thread of life, DNA—will affect just about every branch of biology.

The complete DNA sequencing of more and more organisms, including humans, will revolutionize biology and medicine. In the spirit of this special issue of Scientific American, we predict that genomics will answer many important questions, such as how organisms evolved, whether synthetic life will ever be possible, and how to treat a wide range of medical disorders.

The Human Genome Project is generating an amount of information unprecedented in biology. A simple list of the units of DNA, called bases, that make up the human genome would fill 200 telephone books—even without annotations describing what those DNA sequences do. A working draft of 90 percent of the total human DNA sequence should be in hand by the spring of 2000, and the complete sequence is expected to be available in 2003. But that will be merely a skeleton that will require many layers of annotation to give it meaning. The payoff from the reference work will come from understanding the proteins encoded by those genes.

Proteins not only make up the structural bulk of the human body but also include the enzymes that carry out the biochemical reactions of life. Proteins are composed of units called amino acids that are linked together in a long string; the string folds back onto itself to form a three-dimensional structure that determines the function of the protein. The order of the amino acids is set by the DNA base sequence of the gene that encodes a given protein. Genes dictate the production of proteins through intermediaries called RNA; those that actively make RNA intermediates are said to be "expressed."

The Human Genome Project seeks not just to elucidate all of the proteins produced within a human being but to comprehend how the genes that encode the proteins are expressed, how the DNA sequences of those genes stack up against comparable genes of other species, how genes vary within our species and how DNA sequences translate into observable characteristics. Layers of information built on top of the complete DNA sequence will reveal the knowledge embedded in the DNA. This information will fuel advances in biology for at least the next century. In a virtuous cycle, the more we learn, the more we will be able to extrapolate and hypothesize and understand.

By 2050, we believe that genomics will be able to answer the following major questions affirmatively.

- Will the three-dimensional structures of proteins be predictable from their amino acid sequences?

The six billion bases of the human genome are thought to encode approximately 80,000 proteins. Although the sequence of amino acids in a protein can be translated in a simple step from the DNA sequence of a gene, we cannot currently elucidate the three-dimensional structure of a protein on purely theoretical grounds, and determining structures experimentally can be quite laborious. Still, a protein's structure is conserved—or maintained fairly constantly throughout evolution—much more so than its amino acid sequence. Many different amino acid sequences can lead to proteins of similar shapes, so we can infer the structures of various proteins from studying a representative subset of proteins in detail.

Recently, an international group of structural biologists have begun a Protein Structure Initiative to coordinate their work. Structural biologists "solve" the shapes of proteins by either making very pure crystals of a given protein and then bombarding the crystals with X-rays or by subjecting them to nuclear magnetic resonance (NMR) analysis. Both techniques are time-consuming and expensive. The consortium intends to get the most information out of each new structure by using sequences and existing knowledge about related structures to group proteins into families that are likely to share the same architectural features. Then they plan to target representatives of each family for examination by painstaking physical techniques.

As the catalog of solved structures swells and scientists develop more refined schemes for grouping structures into a compendium of basic shapes, biochemists will increasingly be able to use computers to model the structures of newly discovered—or even wholly invented—proteins. Structural biologists project that a total of about 1,000 basic protein folding motifs exist; current models suggest that solving just 3,000 to 5,000 selected structures, beyond the ones already known, could allow scientists to routinely deduce the structures of new proteins. With structural biologists solving more than 1,000 protein structures every year, and their progress accelerating, they should be able to complete the essential inventory not long after the human genome itself is sequenced.

• Will synthetic life forms be produced?

While structural biologists strive to group proteins into categories for the practical aim of solving structures efficiently, the fact that proteins are so amenable to classification reverberates with biological meaning. It reflects how life on Earth evolved and opens the door to questions central to understanding the phenomenon of life itself. Is there an essential set of proteins common to all organisms? What are the core biochemical processes required for life? Already, with several fully sequenced genomes—mostly from bacteria—available, scientists have started to take inventories of genes conserved among these organisms, guided by the grand question of what constitutes life, at least at the level of a single cell.

If, within a few years, scientists can expect to amass a tidy directory of the gene products—RNA as well as proteins—essential for life, they may well be able to make a new organism from scratch by stringing DNA bases together into an invented genome coding for invented products. If this invented genome crafts a cell around itself and the cell reproduces reliably, the exercise would be the ultimate proof that we understand the basic mechanisms of life. Such an experiment would also raise safety, ethical and theological issues that cannot be neglected.

• Will we be able to build a computer model of a cell that contains all of the components, identifies all of the biochemical interactions, and makes accurate predictions about the consequences of any stimulus given to that cell?

In the last 50 years, a single gene or a single protein often dominated a biologist's research. In the next 50 years, researchers will shift to studying integrated functions among many genes, the web of interactions among gene pathways, and how outside influences affect the whole system.

Of course, biologists have long endeavored to understand how components of a cell interact: how molecules called transcription factors bind to specific scraps of DNA to control gene expression, for example, or how insulin binds to its receptor on the surface of a muscle cell and triggers a cascade of reactions in the cell that ultimately increases the number of glucose transporters in the cell membrane. But the genome project will spark similar analyses for thousands of genes and cell components at a time. Within 50 years, with all genes identified and all possible cellular interactions and reactions charted, pharmacologists developing a drug or toxicologists trying to predict whether a substance is poisonous may well turn to computer models of cells to answer their questions.

• Will the details of how genes determine mammalian development become clear?

Being able to model a single cell will be impressive, but to fully understand the life forms we are most familiar with, we'll plainly have to consider additional levels of complexity. We will have to consider how genes and their products behave in place and time—that is, in different parts of the body and in a body that changes over a lifespan. Developmental biologists have started to monitor how pools of gene products vary as tissues develop, in an attempt to find products that define stages of development. Now, scientists are developing so-called expression arrays that survey thousands of gene products at a time, chart which ones turn on or off, and which ones fluctuate in intensity of expression. Techniques like these highlight many good candidates for genes that direct development and establish the animal body plan.

As in the past, model organisms—like the fruit fly Drosophila, the nematode Caenorhabditis elegans and the mouse—will remain the central workhorses in developmental biology. With the genome sequence of C. elegans completed, Drosophila's also near completion, the complete human sequence on the way by 2003 (with a working draft expected by spring, 2000), and the mouse's likely within four to five years, sequence comparisons will become increasingly commonplace and thorough and give biologists many clues about where to look for the driving forces that fashion a whole animal. Many more complete genomes representing diverse branches of the evolutionary tree will be derived as the cost of sequencing moves steadily downward.

So far, developmental biologists have striven to find signals that are universally important in establishing an animal's body plan, the arrangement of its limbs and organs. In time, they will also describe the variations—in gene sequence, perhaps in gene regulation—that generate the striking diversity of forms among different species. By comparing species, we'll learn how genetic circuits have been modified to carry out distinct programs, so that almost equivalent networks of genes fashion, for example, small furry legs in mice and arms with opposable digits in humans.

• Will understanding the human genome transform preventive, diagnostic and therapeutic medicine?

Molecular biology has long held out the promise of transforming medicine from a matter of serendipity to a rational pursuit grounded in a fundamental understanding of the mechanisms of life. Molecular biology has begun to infiltrate the practice of medicine; genomics will hasten the advance. Within 50 years, we expect comprehensive genomics-based health care to be the norm in the U.S. We will understand the molecular foundation of diseases, be able to prevent them in many cases and design accurate, individualized therapies for illnesses.

In the next decade, genetic tests will routinely predict individual susceptibility to disease. One intention of the Human Genome Project is to identify common genetic variations. Once a catalog of variants is compiled, epidemiological studies will tease out how particular variations correlate with risk for disease. When the genome is completely open to us, such studies will reveal the roles of genes that individually contribute weakly to diseases but interact with other genes and with environmental influences, like diet, infection and prenatal exposures to affect health. By 2010 to 2020, gene therapy should also become a common treatment, at least for a small set of conditions.

Within 20 years, novel drugs will be available that derive from a detailed molecular understanding of common illnesses like diabetes and high blood pressure. The drugs will be designer therapies that target molecules logically and are therefore potent without significant side effects. Drugs like those for cancer will routinely be matched to a patient's likely response, as predicted by molecular fingerprinting. Diagnoses of many conditions will be much more thorough and specific than now. For example, a patient who learns that he has high cholesterol will also know which genes are responsible, what effect the high cholesterol is likely to have, and what diet and pharmacologic measures will work best for him.

By 2050, many potential diseases will be cured at the molecular level before they arise, though large inequities worldwide in access to these advances will continue to stir tensions. When people become sick, gene therapies and drug therapies will home in on individual genes, as they exist in individual people, making for precise, customized medical treatment. The average life span will reach 90 to 95 years, and a detailed understanding of human aging genes will spur efforts to expand the maximum span of human life.

• Will we accurately reconstruct the history of human populations?

Despite what may seem like great diversity in our species, studies from the last decade show that the human species is more homogeneous than many; for example, as a group, we show less variation than chimps do. Among humans, the same genetic variations tend to be found across all population groups, and only a small fraction of the total variation (10-15%) can be related to group differences, leading some population biologists to the conclusion that not so long ago the human species was composed of a small group, perhaps 10,000 individuals, and that populations dispersed over the earth only recently. Most genetic variation predated that time.

Armed with techniques for analyzing DNA, population geneticists have for the past 20 years been able to address anthropological questions with unprecedented clarity. Demographic events like migrations, population bottlenecks and expansions alter gene frequencies and leave a detailed and comprehensive record of events in human history. Genetic data have bolstered the view that modern humans originated relatively recently, perhaps 100,000 to 200,000 years ago, in Africa and dispersed gradually into the rest of the world. Anthropologists have used DNA data to test cultural traditions about the origins of groups, like Gypsies and Jews, to track the migration of people into the south pacific and the americas, and to glean insights about the spread of populations in Europe, among other examples. As DNA sequence data become increasingly easy to accumulate, relationships among groups of people will become clearer, revealing histories of intermingling as well as periods of separation and migration. Race and ethnicity will be proved to be largely social and cultural ideas; sharp, scientifically-based boundaries between groups will be found to be nonexistent.

By 2050, then, we will know much more than we know now about human populations, but an open question is: how much can be known? Human beings have mated with enough abandon and random mutations occur often enough that probably no one family tree will be the unique solution accounting for all human history. In fact, the history of human populations will emerge as a trellis, where lineages often meet and mingle after intervals of separation, not a tree. Still, in 50 years, we will know how much ambiguity remains in our reconstructed history.

• Will we be able to reconstruct the major steps in the evolution of life on Earth?

Molecular sequences have been indispensable tools for drawing taxonomies since the 1960s. To a large extent, DNA sequence data have already exposed the record of 3.5 billion years of evolution, sorting living things into three domains—Archaea (single-celled organisms of ancient origin), Bacteria, and Eukarya (organisms whose cells have a nucleus)—and revealing the branching patterns of hundreds of kingdoms and lower divisions. One aspect of inheritance has complicated the hope of assigning all living things to branches in a single true tree of life. In many cases, different

genes suggest different family histories for the same organisms; this reflects the fact that DNA isn't always inherited in the straightforward way, parent to offspring, with a more-or-less predictable rate of mutation marking the passage of time. Genes sometimes hop across large evolutionary gaps. The most famous examples of this are mitochondria and chloroplasts, the organelles descended from bacteria that were evidently swallowed whole by eukaryotic cells.

This kind of "lateral gene transfer" appears to have been common enough in the history of life that comparing genes among species will not yield a single, universal family tree. As with human lineages, a more apt analogy for the history of life will be, instead of a tree, where branches diverge never to join again, a net or a trellis, where separated lineages often merge again.

In 50 years, we will fill in many details about the history of life, though we may still not understand how the first self-replicating organism came about; we will learn when and how – by inventing, adopting, or adapting genes – various lineages acquired, for example, new sets of biochemical reactions and different body plans. The gene-based perspective of life will have taken hold so deeply among scientists that the basic unit they consider will likely no longer be an organism or a species, but a gene. They will chart which genes have traveled together for how long in which genomes.

Scientists will also address the question that has dogged people since Darwin's day: What makes us human? What distinguishes us as a species? Undoubtedly, many other questions will arise over the next 50 years. As in any fertile scientific field, the data will fuel new hypotheses. Paradoxically, as it grows in importance, genomics may not even be a common concept in 50 years, as it radiates into many other fields and ultimately becomes absorbed as part of the infrastructure of all biomedicine.

• How will individuals, families and society respond to this explosion in knowledge about our genetic heritage?

This social question, unlike the preceding scientific, technological and medical ones, does not come down to a yes-or-no answer. Genetic information and technology will afford great opportunities to improve health and alleviate suffering. But any powerful technology comes with risks, and the more powerful the technology, the greater the risks. In the case of genetics, people of ill will today use genetic arguments to try to justify bigoted views about different racial and ethnic groups. As technology to analyze DNA has become increasingly widespread, insurers and employers have used the information to deny access to health care and work. How we will come to terms with the explosion of genetic information remains an open question.

Finally, will anti-technology and anti-science movements be quieted by all of the stunning revelations of genetic science? After enumerating so many questions where we argue the answer will be yes, this is one question where the answer will probably be no. The tension between scientific advances and the desire to return to a simple and "more natural" lifestyle will likely intensify as genomics seeps into more and more of our daily lives. The challenge will be to maintain a healthy balance and to collectively shoulder the responsibility for ensuring that the advances arising from genomics are not put to ill use. This is, after all, a very old mandate:

> The bravest are surely those who have the clearest vision of what is before them, glory and danger alike, and yet notwithstanding go out to meet it.—Thucydides

FRANCIS COLLINS has been the director of the National Institutes of Health's National Human Genome Research Institute (NHGRI) since 1993. Prior to that, he was a research geneticist at the University of Michigan, where he and his colleagues were the first to clone the gene for cystic fibrosis. A practicing Christian, Collins is particularly interested in the ethical implications of human genetics research. KARIN JEGALIAN, who received her Ph.D. in biology from the Massachusetts Institute of Technology in 1998, is a science writer at NHGRI.

FURTHER READING

NEW GOALS FOR THE U.S. HUMAN GENOME PROJECT: 1998-2003. F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, and the members of the DOE and NIH planning groups in *Science*, Vol. 282, pages 682-689; October 23, 1998.

PRINCIPLES OF MEDICAL GENETICS, Second Edition. T. D. Gelehrter, F. S. Collins, and D. Ginsburg. Williams and Wilkins, 1998.

SHATTUCK LECTURE – MEDICAL AND SOCIETAL CONSEQUENCES OF THE HUMAN GENOME PROJECT. F. S. Collins in *The New England Journal of Medicine*, Vol. 341, pages 28-37; July 1, 1999.

National Human Genome Research Institute: www.nhgri.nih.gov

Department of Energy: www.ornl.gov/hgmis/