GENOMICS : WHAT'S ALL THIS EXCITEMENT ?

PARTHA P. MAJUMDER*

THERE is excitement blowing in the wind. The eye of this storm is *Genomics*, which is really the *study of a genome as a whole,* that is the study of all the genetic material in the chromosomes of a particular organism, as opposed to studying only selected parts of a genome. Studying a genome as a whole permits, among other things, the identification of patterns that have been conserved over the course of evolution and interactions among various regions of a genome.

June 26, 2000, marks a watershed in the annals of science. On this day came the announcement of the completion of the "draft" sequence of the human genome. The "draft" sequence of the human genome comprises an assembly containing overlapping fragments that covers 99 per cent of the genome which have been sequenced, but still contains some gaps and ambiguities. This has been possible through an international effort. The International Human Genome Sequencing Consortium includes scientists at 16 institutions in France, Germany, Japan, China, the UK and USA. The five largest centres are located at : Baylor College of Medicine, Houston, Texas ; Joint Genome Institute in Walnut Creek, California ; Sanger Centre near Cambridge, England; Washington University School of Medicine, St. Louis; and the Whitehead Institute, Cambridge, Massachusetts. In addition a draft sequence was also produced by a private company in the U.S.A., Celera Genomics. Special issues of *Science* (Feb. 16, 2001) and *Nature* (Feb. 15, 2001) contain the working draft of the human genome sequence. *Nature* papers include initial analysis of the descriptions of the sequence generated by the publicly sponsored International Human Genome Sequencing Consortium, while *Science* publications focus on the draft sequence reported by Celera Genomics. These publications have resulted in tremendous scientific excitement. Here are some whiffs of the excitement :

- *"This is a momentous day ... and it may come to be recorded as one of the most significant dates in human history. I say this because knowledge from the Human Genome Project – reading the genetic book of mankind – has the potential to touch the lives of every person on the planet."* (Dr Michael Dexter, Director of the Wellcome Trust, U.K.)
- *"Along with Bach's music, Shakespeare's sonnets and the Apollo Space Program, the Human Genome Project is one of those achievements of the human spirit that makes me proud to be human."* (Professor Richard Dawkins, The Simonyi Professor of the Public Understanding of Science, University of Oxford. U. K.).
- *"The achievement of a fairly complete draft sequence is the celebration of a milestone, rather than the sudden unveiling of previously unsuspected treasures."* (Professor Martin Bobrow, Professor of Medical Genetics at Addenbrooke's Hospital, Cambridge).
- *"The near completion of the Human Genome Project is a remarkable achievement, which will lay a firm base for some remarkable developments in biology and medicine in the new millennium. In essence, it will set us on the long road of learning how our thousands of genes function and in the longer term, how they are orchestrated so that they can control the functions of our cells and organs. In the long term, this information will undoubtedly help us to solve the most important question in biology, that is, how does a single fertilised egg turn into an adult animal or human being."* (Professor Sir David Weatherall, Regius Professor of Medicine at the University of Oxford).

The Human Genome Project ?

The Human Genome Project (HGP) began formally in the U.S. in 1990. The project originally was planned to last 15 years, but rapid technological advances and the formation of an international Consortium have accelerated the expected completion date to 2003. The goals of the project are to :

- *identify* all the approximately 30,000 genes in human DNA,
- *determine* the sequences of the 3 billion chemical bases that make up human DNA,

[*] Anthropology & Human Genetics Unit, Indian Statistical Institute, Kolkata 700 035.

- *store* this information in databases,
- *develop* faster, more efficient sequencing technologies,
- *develop* tools for data analysis, and
- *address* the ethical, legal, and social issues (ELSI) that may arise from the project.

What procedure was used in the HGP to sequence the human genome ?

When the HGP started only a few thousand bases could be sequenced each day. One of the first steps was to produce a framework – a map – of our DNA. This is similar to producing a detailed contents list for a huge encyclopaedia. Another was to generate a set of isolated stretches of DNA, called clones, that could be used as landmarks on the map. That way, the HGP knew where all its DNA pieces are derived from. The third main aim was to fund research into new technology to improve the efficiency of DNA sequencing.

The map has been extremely useful, both to genome sequencers and to medical researchers. The approach in the HGP was to take one of the clones (the BACs) and to break it into fragments of a few thousand bases. Many fragments were produced to ensure that they overlap one another. The ends of the fragments were then sequenced – about 500 bases from each end - and all these sequences were put into a computer. The computer then sorted out the order of the fragments by building up the overlaps between fragments.

This is 'assembly' and is complicated by some of the more notable features of our genomic landscape such as repeated sequences (like seeing a street that looks just like another – you really don't know where you are). It is also complicated by the fact that some DNA sequence does not readily emerge from shotgun sequencing clones. What had to be done then is 'hand finishing' – a highly skilled job to fill in the missing links which can't be done by automation.

The finishing teams worked out strategies to get at the bits that are hard to work with and only when all existing technologies were exhausted did they cease. This is the reason why there are still some 'gaps' and unfinished regions in the available 'draft' human genome sequence.

Some Interesting Facts : When the HGP was initiated in 1990, vital automation tools and high-throughput sequencing technologies had to be developed or improved. The cost of sequencing a single DNA base was about $10 then; today, sequencing costs have fallen about 100-fold to $0.10 to $0.20 a base and still are dropping rapidly.

Since spring 1998, when the private company, Celera Genomics, announced its sequencing goal, other private companies also have declared their intention to sequence or map genomic regions to varying degrees. Some people questioned whether the HGP and the private sector were duplicating work, and they wondered who would win the race to sequence the human genome. Although the HGP and private companies do have overlapping sequencing goals, their finish lines are different because their ultimate goals are not the same. The HGPs commitment from the outset has been to create a scientific standard (an entire reference genome). Most private-sector human genome sequencing projects, however, focus on gathering just enough DNA to meet their customers needs probably in the gene-rich, potentially lucrative regions. Such private data continue to be enriched greatly by accurate free public mapping (location) and sequence information. Celera's shotgun sequencing strategy, for example, creates millions of tiny fragments that must be ordered and oriented computationally using HGP research results. Most data at Celera, Incyte, and other genomics information based companies are proprietary or available only for a fee. In addition, companies are filing numerous patent applications to stake early claims to genes and other potentially important DNA fragments.

What Has Been Learned from Analysis of the Working Draft Sequence of the Human Genome?

In generating the draft sequence, scientists determined the order of base pairs in each chromosomal area at least 4 to 5 times (4x to 5x) to ensure data accuracy and to help with reassembling DNA fragments in their original order. This repeated sequencing is known as genome "depth of coverage." Draft sequence data are mostly in the form of 10,000 basepair-sized fragments whose approximate chromosomal locations are known. To generate high-quality sequence, additional sequencing is needed to close gaps, reduce ambiguities, and allow for only a single error every 10,000 bases, the agreed-upon standard for HGP finished sequence. Investigators believe that a high-quality sequence is critical for recognizing regulatory components of genes that are very important in understanding human biology and such disorders as heart disease, cancer, and diabetes. The finished version will provide an estimated 8x to 9x coverage of each chromosome. Thus far, finished sequences have been generated for only about 35% of the human genome; the goal is to reach 100% by 2003.

The following facts have been collated from the journals *Science* and *Nature*, and the Wellcome Trust, Human Genome News and HGP web sites.

- The human genome contains 3164.7 million nucleotide bases.
- The average gene consists of 6,000 nucleotide bases, though sizes vary greatly.
- The total number of genes is estimated at 30,000 to 35,000—much lower than previous estimates of 80,000 to 100,000 which were based on extrapolations from gene-rich areas as opposed to a composite of gene-rich and gene-poor areas.

- 99.9% of all nucleotides are exactly the same in all people.
- The function is unknown for over 50% of genes that have been discovered.
- The actual part of the genome that codes for proteins makes up less than 2% of the genome.
- Repeated sequences or "junk DNA" make up at least 50% of the genome. Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. They hold important clues about evolutionary events, help chart mutation rates, and, by seeding DNA rearrangements, they can modify genes and create new ones.
- There seems to have been a dramatic decrease in the accumulation of repeats in the human genome over the past 50 million years.
- The human genome's gene-dense regions are predominantly composed of the nucleotides G and C and are called "GC-rich regions." In contrast, the "junk-DNA" regions are AT rich.
- Genes appear to be concentrated in random areas along the genome with vast expanses of noncoding "junk DNA" in between.
- CpG islands —stretches of up to 30,000 letters with only two bases, C and G, repeating over and over—often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." It is believed that these CpG islands help regulate gene function.
- Humans appear to have gotten hundreds of their genes from various bacteria through horizontal transfer.
- Chromosome 1 has the most genes (about 3000); chromosome Y has the fewest (about 250).
- There are ~ 30,000 genes in the human, 18,000 in the nematode worm, 13,000 in the fruitfly, 6,000 in yeast, and 4,000 in the tuberculosis microbe.
- Humans have on average three times as many kinds of proteins as the fly or worm because of "alternative splicing," a process that can yield different protein products from the same gene.
- 50% of the human genome is repeats compared with the mustard weed (11%), the worm (7%), and the fly (3%).
- Unlike the human's seemingly random distribution of gene-rich areas, many other organisms' genomes are more uniform, with evenly spaced genes throughout.

  While humans appear to have stopped accumulating junk DNA over 50 million years ago, there seems to be no such decline in repeats in rodents. Likewise, many extinct or near-extinct repeats in the human genome appear to be alive in the mouse genome, suggesting that these differences may account for some of the fundamental differences between hominids and rodents.
- About 1.4 million single base differences, or single nucleotide polymorphisms (SNPs) — have been identified. This information promises to revolutionize both mapping diseases and tracing human history.
- The ratio of mutations in males versus females is 2:1. Scientists point to several reasons for the higher mutation rate in the male germline, including the fact that there are a greater number of cell divisions involved in the formation of sperm than in the formation of eggs.

What Remains Unknown

- Exact gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content, and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology, including microbial consortia useful for environmental restoration
- Developmental genetics.

Genomics Spawns New Interfaces

The Human Genome Project has resulted in vast masses of data, which are stored in databases spread across the world. The creation, development, and operation of databases and other tools to collect, organize, analyze and interpret data these data has given rise to the new sciences of Bioinformatics and Computational Biology. Efficient interpretation of the functions of human genes and other DNA sequences requires that resources and strategies be developed to enable large-scale investigations across whole genomes. A technically challenging first priority is to generate complete sets of full-length cDNA clones and sequences for human and model-organism genes. Other functional-genomics goals include studies into gene expression and control, creation of mutations that cause loss or alteration of function in nonhuman organism, and development of experimental and computational methods for protein analyses. The functions of human genes and other DNA regions often are revealed by studying their parallels in nonhumans. To enable such comparisons, HGP researchers have obtained complete genomic sequences for the bacterium *Escherichia coli,* the yeast *Saccharomyces cerevisiae,* the roundworm *Caenorhabditis elegans,* the fruitfly *Drosophila melanogaster* and the laboratory mouse. The availability of complete genome sequences generated both inside and outside the HGP is driving a major breakthrough in fundamental biology as scientists compare entire genomes to gain new insights into evolutionary, biochemical, genetic, metabolic, and physiological pathways. HGP planners stress the need for a sustainable sequencing capacity to facilitate future comparisons.

Computers can be used very effectively to indicate the location of genes and of regions that control the expression of genes and to discover relationships between each new sequence and other known sequences from many different organisms. This process is referred to as "sequence annotation." Annotation (the elucidation and description of biologically relevant features in the sequence) is the essential prerequisite before the genome sequence data can become useful, and the quality with which annotation is done will directly affect the value of the sequence. In addition to considerable organizational issues, significant computational challenges must be addressed if DNA sequences that are produced can be successfully annotated. It is clear that new computational methods and a workable process must be implemented for effective and timely analysis and management of these data. The analysis of genomic sequence regions needs to be updated continually through the course of the Genome Project. On any given day, new information relevant to a sequenced gene may show up in any one of many databases, and new links to this information need to be discovered and presented. Additionally, our capabilities for analyzing the sequence need to be vastly improved. There will be a significant advantage in reanalyzing sequences and updating our knowledge of them continually as new sequences appear from many organisms, methods improve, and databases with relevant information grow. Thus, sequence annotation is a living thing that will develop richness and improve in quality over the years.

The recognition of important features in a sequence, such as genes, must be highly automated to eliminate the need for time-consuming manual gene model building. Five distinct types of algorithms (pattern recognition, statistical measurement, sequence comparison, gene modeling, and data mining) must be combined into a coordinated toolkit to synthesize the complete analysis. One of the key types of algorithms needed is pattern recognition. Efficient methods must be designed to detect the subtle statistical patterns characteristic of biologically important sequence features, such as genes or gene regulatory regions. DNA sequences are remarkably difficult to interpret through visual examination. However, when examined in the computer, DNA sequence has proven to be a rich source of interesting patterns, having periodic, stochastic, and chaotic properties that vary in different functional domains. These properties and methods to measure them form the basis for recognizing the parts of the sequence that contain important biological features.

Efficient methods are needed to locate and retrieve information relevant to newly discovered genes. If similar genes or proteins are discovered through sequence comparison, often experiments have been performed on one or more homologs that can provide insight into the newly discovered gene or protein. Relevant information is contained in more than 100 databases scattered throughout the world, including DNA and protein sequence databases, genome mapping databases, metabolic pathway databases, gene expression databases, gene function and phenotype databases, and protein structure data-bases. These data can provide insight into a gene's biochemical or whole organism function, pattern of expression in tissues, protein structure type or class, functional family, metabolic role, and potential relationship to disease phenotypes. Using the Internet, researchers are developing automated methods to retrieve, collate, fuse, and dynamically link such database information to new regions of sequence. The target data resources are very heterogeneous (*i.e.*, structured in a variety of ways), and some are merely text-based and poorly formatted, making the identification of relevant information and its retrieval difficult. One challenge here is that information relevant to an important gene or protein may appear in any database at any time. As a result, systems now being developed dynamically update the descriptions of genes and proteins in data warehouses and continually poll remote data resources for new information.

The sheer volume and complexity of the analyzed information and links to data in many remote databases require advanced data visualization methods to allow user access to the data. Users need to interface with the raw sequence data; the analysis process; and the resulting synthesis of gene models, features, patterns, genome map data, anatomical or disease phenotypes; and other relevant data.

What is the Next Step

Data from the HGP are already helping scientists to identify how a specific type of damage can lead to a defective protein, and a disrupted function that leads to disease. This will help uncover new targets for diagnosis and for therapy.

In addition, the identification of defective genes will improve diagnosis. Early diagnosis may allow treatment to start as early as possible. Genetic tests are already being used to check for a variety of genetic disorders, such as cystic fibrosis and sickle cell anaemia and many more will come from data arising from the HGP. For multifactorial conditions, genetic tests may eventually be able to detect predisposition rather than existence of disease and a battery of tests may be able to tell us what disease we are more likely to suffer from, which may open new avenues for preventive medicine. These tests may also give clues to which drugs are going to be most effective in this individual.

Sophisticated computers are being used to predict the 3D structures of proteins from DNA sequence information alone. Once the 3D structure of a protein is known, drugs can be designed that 'fit' with the protein structures, so speeding up the design of possible therapies.

DNA sequences of other organisms will help scientists for two reasons. First, the genome sequences of organisms such as the nematode worm, *C. elegans,* and brewer's yeast (*Saccharomyces cerevisiae*), which have already been sequenced, can provide clues to the functioning of human genes. These and other organisms provide cheap and user-friendly models for studying several aspects of human development – from cell division and growth to the development of specialized tissues. Many genes are highly conserved in living creatures and related genes carry out related functions even in quite disparate species.

One major area of future research involves the identification of minute genetic differences – known as single nucleotide polymorphisms (SNPs or 'snips') – between individuals. This will help scientists to understand why some individuals are susceptible to specific diseases and some resistant.

The Future is Genomic

Genome sequencing will ultimately change the face of biology. More immediately, it is likely to change the way scientists conduct their research. The genomics revolution is already having a dramatic effect on the way scientists work. With a draft sequence of the genome readily available, scientists have access to a data which provides a more rational approach to gene hunting and determining the function of those genes.

As genes are identified, scientists can work from the gene towards function, but at present most genes – probably tens of thousands – remain a mystery. This presents a further challenge to analyse the data and assign functions to genes.

Researchers are already analysing the DNA sequence to identify variations in our genome (known as single nucleotide polymorphisms – or SNPs) which, with the help of the draft sequence, can easily be spotted. In 1999, the Wellcome Trust and ten pharmaceutical companies formed a joint initiative to identify the common variations in our genes. The SNP Consortium (which now numbers 12 companies and the Trust), one of the first projects set up to exploit data from the human genome project, is pinpointing the subtle genetic differences that predispose some but not others to diseases such as Alzheimer's, cancer and diabetes. Similar projects have also been initiated at smaller scales by various other organizations across the world. These data will provide the blueprint for establishing personalized medicine. Without the knowledge arising from human genome project, tracking the SNPs would be a very difficult task. ❐