

The human genome project

Prapon Wilairat, Ph.D.*

Talk presented at the Faculty of Medicine, Srinakarinwirot University, June 14, 2001

The Human Genome Project is an ambitious but achievable goal of sequencing the entire human genome, which consists of 3,000 million base pairs. This was initiated in 1991, and was planned to be completed in 2005, covering a period of 15 years. In the process genetic and physical maps of all 24 chromosomes (22 autosomal, X and Y chromosome) will be constructed as well as the acquisition of DNA sequences for actively expressed genes. With recent advances in the rate at which the genome can be sequenced, it was possible to announce on June 26, 2000, that 97% of the human genome had been sequenced. Previous to that, the DNA sequences present in two human chromosomes (21 and 22) had been determined. To date, the entire genomes of 51 bacteria and 4 eukaryotic organisms (yeast (*Saccharomyces cerevisiae*), worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*) and plant (*Arabidopsis thaliana*)) have been sequenced and annotated. In February 2001, the data of human genome were published by Celera Genomics, a private company, in *Sciences* (v. 291 (Feb 16) 2001) and by the public Human Genome Project (HGP) consortium in *Nature* (v. 409 (Feb 15) 2001).

In its simplest term, the human genome project involves generating and sequencing cloned DNA molecules, and ordering them along the 24 chromosomes. A fragment of DNA (average size

of 400 kb) is first cloned into a yeast artificial chromosome (YAC, constructed from the telomeric, centromeric and replication origin sequences allowing for replication in yeast cells) (see Figure 1). A smaller fragment (40 kb) is then cloned into a cosmid (plasmid vector containing cos gene of phage lambda allowing it to be packaged into phage particles for infection in *E. coli* in which it replicates as a plasmid). Finally fragments of 5 kb can be cloned into plasmid for sequencing. Once the sequences of these DNA fragments have been determined, the order of these pieces along the chromosome is constructed by aligning overlapping sequences. This was the approach originally taken by the HGP consortium. A simpler strategy pioneered by Celera Genomics is to generate random fragments produced by partial digestion with a restriction enzyme. DNA clones with overlapping sequences are then identified and the fragments aligned enabling the DNA sequence of the chromosome to be assembled (Figure 2).

The human genome consists of 3,296,000,000 base pairs which contains information for about 32,000 genes comprise 1% of the genome, but introns occupy 24% of the genome. Fifty percent of the genome are repeat sequences. Surprisingly, there are some 100 genes that are more similar to bacteria than to yeast, worm, fly or plant.

In the past few years, there has been a shift in emphasis from structural genomics, i.e. the

*Department of Biochemistry, Faculty of Sciences, Mahidol University

construction of high-resolution genetic, physical and transcript map of an organism, to functional genomics, i.e. an understanding of the function of all genes at once, conducted in a systemic fashion, providing new insights into the behavior of biological systems not available from previous studies involving only single genes. This shift is very important for target research, since it moves from sequence information to biological information. It requires knowledge about transcription of genetic messages and expression of the transcripts as proteins.

In order to study the expression of all genes in a cell or tissue sample, it is necessary to identify the mRNAs present. This can now be readily accomplished by means of DNA microarrays (or sometimes called 'DNA chips'). This consists of a very large number (from thousands up to possibly million) of microdots of immobilized DNA, each representing a gene, part of a gene or even random sequences of oligodeoxynucleotides. A chip the size of microscope slide can contain thousands of microdots, thus enabling the entire genome of an organism to be placed on one

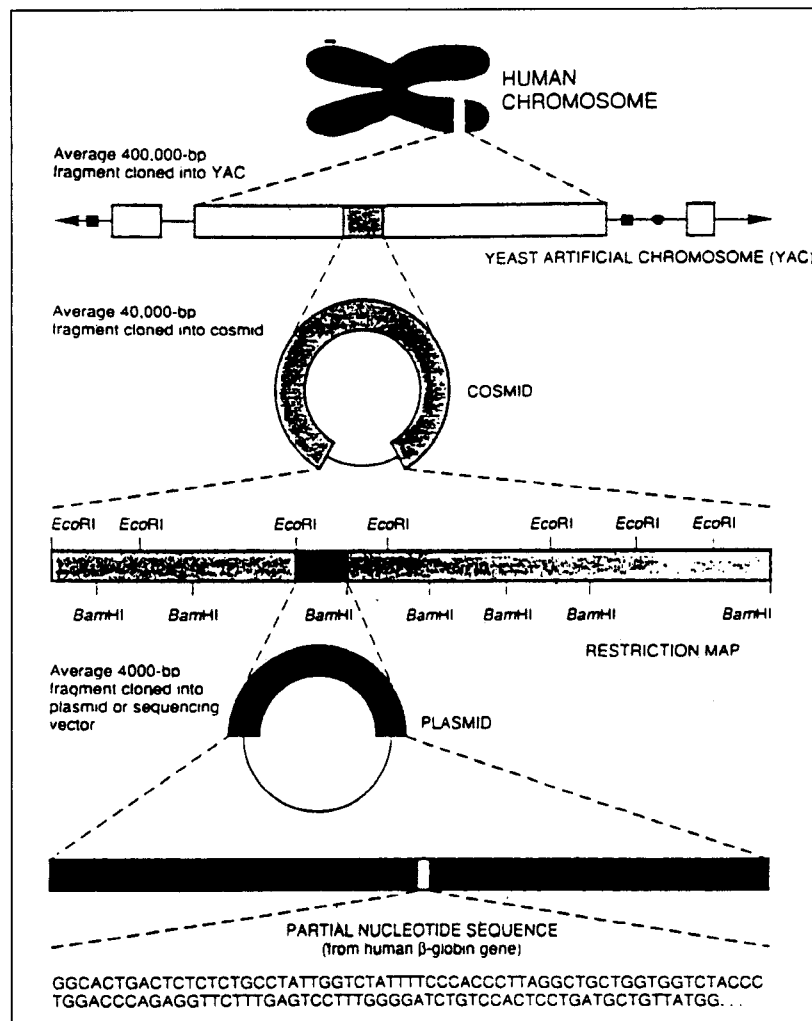


Figure 1 Construction of human DNA clones for sequencing

Cloned DNA sequences must be made progressively smaller in order to be sequenced. The results are then compiled to produce longer stretches.

(Source : United States Development of Energy, "Human Genome 1991-92 Program Report" Office of Health and Environmental Research, Washington, D.C., 1992, P. 207)

microarray. The chip can then be hybridized with samples of RNA or DNA labeled with fluorescent dyes to allow their location in the microarray to be readily visualized. There are two methods for creating the DNA microarray: Affymatrix microphotolithographic technique in which oligomers are built up one nucleotide at a time on the array, and the direct attachment of oligonucleotides synthesized by PCR onto the solid matrix (generally microscope slide). The Affymatrix method is proprietary whereas the latter technique can developed by individual laboratories.

However there is not always a tight correlation between gene expression (mRNA) and the protein components within a cell or tissue. This lack of correlation may be due to such

factors as difference between the time at which the gene and protein acts, difference in the site of gene expression and where the protein acts, difference in the turnover rates of mRNA and protein, and alternate splicing of mRNA to yield various protein products. Therefore, the expression of a gene has no direct bearing on the ultimate expression or amount of a protein product and thus there is a need to catalogue or profile the complete set of protein in order to have a valid comparison between different tissue or between normal and diseased cells. The term proteome has been coined to define the entire protein set expressed by a genome or by a cell or tissue type. Proteomics is the study of the whole protein complement.

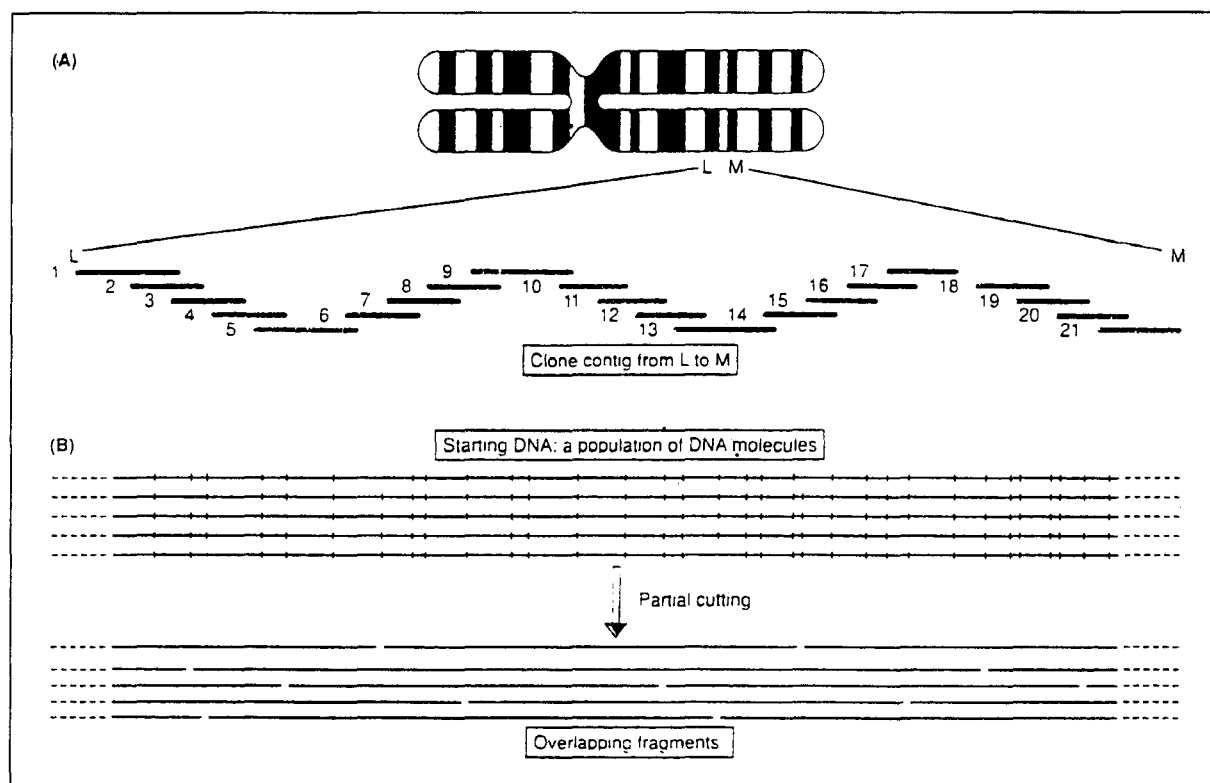


Figure 2 A clone contig

(A) Clone contig between chromosomal sites L and M consists of 21 clones arranged as a series of clones with overlapping sequences.

(B) Overlapping DNA fragments are generated by partial restriction enzyme digestion.

(Source : Tom Strachan and Andrew P. Read. "Human Molecular Genetics", bios Scientific Publishers Ltd.,

Proteomics has the additional advantage of providing information regarding post-translational modification of proteins, which cannot be obtained from knowledge of the gene sequence or its expression. Post-translational modifications include cleavage into smaller fragments, phosphorylation and glycosylation, processes which affect protein structure, activity and compartmentalization. For example, transmembrane signaling involves a complex cascade of protein phosphorylation and

de-phosphorylation, often utilizing pre-existing enzymes with no involvement of gene expression. Only analysis at the protein level will reveal these changes.

The existing operation of proteomics involves the use of two-dimensional polyacrylamide gel-electrophoresis (2D-PAGE) and mass spectrometry (MS). 2D-PAGE provides the most sensitive high-resolution method for displaying proteins and MS is the most sensitive method available to identify proteins (see Figure 3).

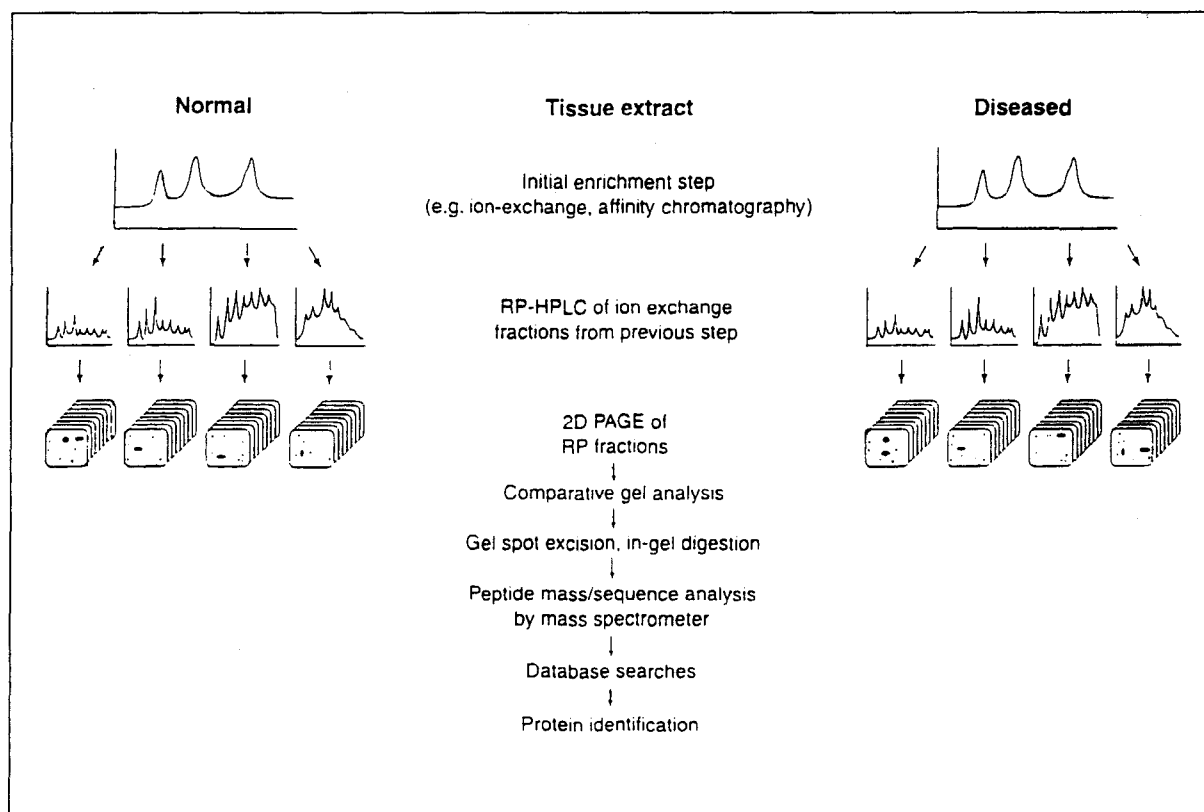


Figure 3 Overview of the sequence of steps involved in a typical proteomics procedure

(Source : Wang JH, Hewick RM. Proteomics in drug discovery. Drug Discovery Today 1999; 4: 129–133.)

Proteomics has become a necessary component to any functional genomics program. It is still a technically demanding procedure requiring relatively large amounts of proteins, but these drawbacks will eventually be overcome as the importance of proteomics is recognized.

Thus a century after the British zoologist William Bateson discovered the findings of Gregor

Mendel and coined the word 'genetics' research on genes has changed beyond recognition. "The world of biomedical research will be divided into the two halves: the gene-by-gene, clone-by-clone approach that marked the first half, and the global, population-based, genome-centered analysis that will dominate the second" (From Kevin Davis Cracking the Genome, The Free Press, NY, 2001)