Bioinformatics in the Post-genomic Era

Sangsoo Kim

Korea Research Institute of Bioscience & Biotechnology
Yu-seong, Taejon, KOREA

The current 7,000 or so drugs target about 500 genes. It is expected that the number of drug targets would reach close to 10,000 when all of the 100,000 - 140,000 human genes are identified by the Human Genome Project. With this hope, many small and big pharmaceutical companies have invested in functional genomics. Now the challenge facing these companies is how to sort out *druggable* targets out of these many genes. Bioinformatics, after compiling and sorting out various experimental and biblical information, is expected to prioritize the genes so that which genes ought to be examined closely.

Functional genomics has to deliver experimental information in three different aspects. First of all, the biochemical nature of a gene must be understood. For a rational drug discovery, it is helpful to know which protein family the genes of interest belong to. For example, a wealth of knowledge has been accumulated for designing drugs targeting proteases or G-protein-coupled receptors. Much of the biochemical information usually comes from motif detection by comparing the amino acid sequence of the candidate gene with those of the known genes. Completion of the Human Genome Project will provide the sequence information of all the human genes and would facilitate to a certain extent the assignment of biochemical functions.

The second issue has to do with the tissue distribution of the gene product. One of the desirable feature of a good drug target is that it is expressed only in certain tissue or disease tissue. So far much of the information has been compiled through the EST sequencing projects. By counting the cluster size of each unique gene in the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene), it is possible to obtain a qualitative measure of gene expression profile over various tissues classes. More quantitative measures can now be obtained rapidly through DNA microarray experiment. Gene Logic, Inc. recently announced the completion of such a database covering 40 different types of human tissues.

The information pertaining biological roles of a gene or its product used to be in the realm of gene-by-gene molecular cell biology. In the post-genomic era, this kind of

information has to be gathered in a high throughput fashion. Surveying genome-wide expression profile using DNA microarray under various genetic or chemical insult would yield information regarding which genes are involved in which metabolic pathway or biological processes.

As genomics deals with massive biological data and generates complex information, a new information technology has been needed. Bioinformatics is a half information technology and a half biological science. With the advent of World-wide Web and internet, it became possible to access the database around the world using relatively easy user interface. The basic database handling and the development of analysis algorithm have been evolved with the progress of the Human Genome Project. During the sequencing phase of the project, the contig assembly, EST clustering, gene prediction and sequence alignment have been well developed. In the functional genomics era, the focus has been moved to the analyses of microarray data and Single Nucleotide Polymorphism.

In this talk, I will review some of the new developments in microarray data analysis and present the data obtained using locally implemented software.