S1-4 *[2000 Oct. 19 (Thur), 11:35~12:05/Hall A]*

Bioinformatics for The Korean Functional Genomics Project

Sangsoo Kim

Korea Research Institute of Bioscience & Biotechnology, 52 Oun-dong, Yusung-gu, Taejon, 305-333 KOREA

Genomic approach produces massive amount of data within a short time period. New high-throughput automatic sequencers can generate over a million nucleotide sequence information overnight. A typical DNA chip experiment produces tens of thousand expression information, not to mention the tens of megabyte image files. These data must be handled automatically by computer and stored in electronic database. The analysis result are usually shown in Web environment. Thus there is a need for systematic approach of data collection, processing, and analysis. DNA sequence information is translated into amino acid sequence and is analyzed for key motif related to its biological and/or biochemical function. Functional genomics will play a significant role in identifying novel drug targets and diagnostic markers for serious diseases. As an enabling technology for functional genomics, bioinformatics is in great need worldwide. In Korea, a new functional genomics project has been recently launched and it focuses on identifying genes associated with cancers prevalent in Korea, namely gastric and hepatic cancers. This involves gene discovery by high throughput sequencing of cancer cDNA libraries, gene expression profiling by DNA microarray and proteomics, and SNP markers in Korea patient population. Our bioinformatics team will support these activities by collecting, processing and analyzing these data.

Key Ideas :

EST sequencing of stomach and liver cDNA libraries will produce many redundant sequences. These will be clustered in order to identify unique clones. The goal is to identify 10-20k unique genes for each of stomach and liver tissues. These clones will be used in the production of stomach- or liver-specific DNA chips. These expression profile data will be cluster over both genes and tissues. The goal is to identify cancer marker genes and potentially to subclassify cancer-types. Patient information will be collected in conjunction with clinical tissues and will be used to draw correlation between gene expression profiles and epidemiological factors. It is also hoped that this expression profile data give clues to identifying novel anti-cancer drug targets.

Methods & Research Contents :

EST sequence data will be collected from our MegaBACE automatics sequencers and transfered to Linux servers for automatics base-calling and vector trimming. Autonomous BLAST servers are being set up for parallel annotation of 1,000 sequences per day. EST sequences are first compared to NCBI's UniGene database, subsequently to non-redundant peptide sequence database via BLASTX. As genomic sequences become available, each EST sequence will be mapped to the chromosome loci.

As an initial attempt, we will produce DNA microarray containing about 10,000 known human genes. Gene annotaion information as well as expression profile data from 6 clinical pathology teams will be collected and stored in a relational database system. The system will be scalable as the amount of data grows and provide easy retrieval. There are several clustering methods available for DNA chip data. Employing these methods, the expression data will be clustered in two ways: one over genes and the other over tissues. Statistical analysis will be implemented in order to identify cancer-specific genes.