Sunitha Kogenaru
Dr. sc. hum.

Translation of large-scale genomics data into functional profiles: a bioinformatics approach

Promotionsfach: DKFZ, (Deutsches Krebsforschungszentrum)
Doktorvater: Prof. Dr. rer. nat. Sándor Suhai

As the complexity of living organisms is getting unveiled, enormous amount of sequence and genome data is being generated. This overwhelming amount of data cannot be efficiently contained in a single biological database, but is distributed over several databases specialized in different domains of interests. Consequently, the biomedical scientists, who intend to design, and interpret their high-throughput experiments, are bound to face difficulties. Because, performing multi-database queries manually is a time-consuming process, especially when it involves large amount of data. This is further compounded by the need for selection of suitable databases that are relevant to the context under consideration. Therefore, there is a need to develop more customized applications and derived databases that can help to process, interpret, and analyze the data from high-throughput experiments in a more automated fashion. This fact has been the motivation for developing TissueDistributionDBs, ExpressMiner, and MapIT.

TissueDistributionDBs is a repository of tissue-distribution profiles based on expressed sequence tags (EST) data extracted from UniGene. This repository provides quantitative expression of genes in tissue-types at four different tissue ontological levels: upper-organ, lower-organ, upper-tissue, and lower-tissue system. Tissue Synonym Library is created, in order to overcome the occurrence of natural language variations in the EST's source tissue-types, and in order to standardize them by cross-referencing to the controlled vocabulary available at BRENDA Tissue Ontology. Finally, the tissue-distribution profiles are generated by calculating the quantitative expression of genes in the tissue-types at different ontological levels. Subsequently, the profiles are integrated into the Sequence Retrieval System (SRS) to help complex querying. TissueDistributionDBs is currently available for twenty different model organisms. This database system is useful for understanding the tissue-specific expression patterns of genes across different organisms, which have implications in the identification of new possible therapeutic drug targets, gene discovery, design and analysis of micro-arrays.

ExpressMiner is a tool to automatically integrate information from several resources to create functional profiles based on Gene Ontology (GO) and biochemical pathway information. These functional profiles provide an overview of the underlying predominant biological theme, at cell-level. ExpressMiner constructs two different types of functional profiles: "gene-specific" and "list-specific", by assigning quantitative scoring at different levels, for GO categories: molecular function, biological process and cellular components. ExpressMiner can also construct profiles using hierarchical biochemical pathway information. Further, ExpressMiner acts as an interface between the tissue-level and chromosome-level by providing functional profiles using tissue-distribution data and cytogenetic locations. ExpressMiner is currently available for six different model organisms. Using ExpressMiner, it was shown that retroviral and lentiviral insertions during gene therapy are non-random and biased to genes having specific molecular functions and biological processes. ExpressMiner

has been also useful in molecular classification of differentially expressed genes in studies related to the adverse prognostic factors of multiple myeloma.

To understand genomes at chromosome-level, MapIT is developed. It systematically maps the genes on to the genome in six different model organisms. MapIT provides gene with or without splice variants information. MapIT uses stringent query criterion, which enhances the sensitivity of MapIT, when compared to other resources. MapIT, has been successfully used to show that promiscuously expressed genes tend to co-localize in clusters in the studies related to epigenetic co-regulations. MapIT has been also used to demonstrate in multiple myeloma, that specific chromosomal deletions are indeed independent adverse prognostic factors. Human diseases like cancers, heart diseases and even personality traits involve several genes that co-localize, therefore MapIT has implications in studies related to these conditions at chromosomal-level.

TissueDistributionDBs, ExpressMiner, and MapIT, assist to dissect the large-scale genomics data into functional profiles at three hierarchical levels: tissue-level, cell-level and chromosome-level, respectively, having several implications in the biomedical research.