Yakhini, Zohar

Analysis of gene expression data: clustering and beyond

Zohar Yakhini[1], Amir Ben-Dor[2], Stuart Kim[3] & Ron Shamir[4]

[1]Hewlett Packard Laboratories, Heifa Israel.
[2]Department of Computer Science & Engineering, University of Washington, Seattle, Washington, USA
[3]Stanford University, Department of Developmental Biology, Stanford, California, USA
[4]Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Data obtained from multiconditional gene expression assays contain information that is instrumental in understanding related biological processes, in a variety of levels and contexts. The nature of studies of multiconditional gene expression patterns may vary widely. Accordingly, it is important to have flexible analysis tools that are useful in as many contexts as possible. Clustering techniques are applicable as they cluster sets of genes that 'behave similarly', under any appropriate definition of similarity. A common theme in the literature of clustering methods is the need to fit the approach to the problem at hand and the necessity to assess the quality of solutions by subjective impression of experts in the field. We have developed an efficient and effective clustering algorithm well suited to the analysis of gene expression data. In contrast to prior approaches to clustering gene expression patterns, which use hierarchical methods (constructing phylogenetic trees) or methods that work for only a single class of similarity metrics (Euclidean distance), our algorithm is based on an abstract graph theoretic approach. There is no need to assume a particular similarity function or number of clusters sought. The cluster structure is produced directly, without involving an intermediate tree stage. On an appropriate stochastic model of the input, a O(n *log*(n))-time version of the algorithm recovers cluster structures with high probability (n is the number of genes). Practical heuristic improvements of the algorithm were implemented in a software package we routinely use to analyse actual data. In the talk we demonstrate the algorithm's performance on simulated data as well as on gene expression data from humans, *Caenorhabditis elegans* and other organisms. The algorithm correctly generated a number of expression clusters with genes that are known to have similar biological function. For example, two known tumour-suppressor genes were present on the pilot *C. elegans* arrays, and both genes appeared in the same cluster with four other genes. Clustering of cancerous and normal human tissues demonstrates the power of gene expression profiles as a base for functional classification. Our package enables further analysis of the clusters such as highlighting the conditions that best characterize them. We also support the analysis of expression data given as distributions rather than as single numbers, thus allowing for more confidence in the implied relationships. It seems clear that this method of analysis will facilitate the generation of new biological hypotheses and insights.

Yang, Liming

Building lymphochip-computational algorithms for selecting clones highly expressed in B cell cDNA libraries

Liming Yang[1], John Powell[1], Ash Alizadeh[2], R. Eric Davis[2], Chi Ma[2], Hajeer Sabet[2], Truc Tran[2] & Louis M. Staudt[2]

[1]Bioinformatics and Molecular Analysis Section, CBEL, CIT, NIH, Bethesda, Maryland, USA
[2]Metabolism Branch, Division of Clinical Sciences, NCI, NIH, Bethesda, Maryland, USA

The human immune system is important in immune responses, cancer biology and several genetic diseases. Studies on the expression levels of genes related to immune system are critical in understanding the mechanisms of immune response and pathology of human diseases. Recent development of cDNA microarray technology makes it possible to study genome-scale gene expression. Lymphochip is designed as a cDNA microarray chip, representing genes in a wide variety of lymphocyte differentiation and activation processes. Screening over 70,000 EST sequences from germinal centre B cell, follicular lymphoma, follicular mixed small and large cell lymphoma, mantle cell lymphoma and chronic lymphocytic leukemia libraries required development of automated computational tools and a structured, robust data management system. We used multiple criteria for selecting clones on the Lymphochip. One is the frequency with which a gene sequence has been sequenced from the select set of B cell and lymphoid libraries, an expanded set of lymphoid libraries versus all other libraries. Clones were classified as being unique to a library, unique to the pool of libraries or mostly lymphoid (i.e. 75% of matching hits were derived from an expanded set of lymphoid libraries). The analysis was performed by tabulating the BLAST results of each candidate sequence against the dbEST database. A second criterion was to select clones matching a specific set of interesting genes. This set of named genes includes genes encoding cytokines, cytokine receptors, adhesion molecules, cell surface differentiation markers, signal transduction and transcription factors, cell cycle and apoptosis proteins, oncogenes, tumour suppresser and human viral genes. Additionally, if they existed, we chose a second clone from a cluster meeting a criterion. Computational algorithms were designed for automating the selection process for the above criteria. To minimise the redundancy of clones included on Lymphochip, we used a modified version of the CLEANUP algorithm to cluster the overlapping sequences. UniGene clustering was also used as a reference and some additional clones from "mostly lymphoid" UniGene clusters were selected. The current version of Lymphochip includes 17835 clones, with 9865 clones unique to the B cell libraries. The clones on Lymphochip represent a total of 6343 UniGene clusters. With the fast accumulation of new EST sequence data and new studies on known genes, iterative analysis with the computational algorithms is necessary to include further novel clones derived from B cell libraries and new genes of interests on Lymphochip. A number of experiments with clinical samples or laboratory cell lines have been performed with Lymphochip.