Eytan Domany

Liat Ein-dor
Ilan Tsafrir
Hilah Gal
Itai Kela
Tal Shay
Noam Shental
Dafna Tsafrir
Or Zuk
Hila B. Rodrig
Shiri Margel
Michal Mashiach
N. Kopelman
Michal Sheffer
Yuval Tabach
Libi Hertzberg
Assif Yitzhaky

Exploratory analysis of biological data

Department of Physics of Complex Systems

Tel. 972 8 934 **3964** **Fax.** 972 8 934 **4109**
E-mail: eytan.domany@weizmann.ac.il
Web page: www.weizmann.ac.il/physics/complex/compphys

We use exploratory methods for data analysis, with emphasis on novel clustering and sorting schemes developed in our group, to study problems in biology (microarray data analysis, promoter motifs, antigen 'chip'); physics (spin glasses) and information technology (machine vision). The newest addition to our tool-box is SPIN (Sorting Points Into Neighborhoods), an unsupervised method for data organization and visualization, which utilizes traits of distance matrices to order data in a way that highlights the underlying structure. In the context of gene array experiments SPIN provides two-way sorting, using the genes to order the samples and vice versa. One of the strengths of the tool is its clear and intuitive graphical output, which includes a two-way sorted expression matrix, sorted distance matrices and PCA images colored according to the order suggested by the algorithm. SPIN's advantage over traditional analysis methods, such as hierarchical clustering, is it's ability to go beyond identification of the clusters into which the data is naturally partitioned, by revealing also the shapes and relative relationships of these clusters.

Another novel tool is aimed at discovering transcription regulation underlying expression signals. The motivation is to gain additional insight on clusters of co-expressed genes, by using the principle that co-expression may indicate co-regulation. The hints for the regulation program of genes are believed to be lying in the genes' upstream regions – the promoter sequences. We first developed a probabilistic algorithm for searching for putative binding sites in the promoter sequence, using a given PSSM (Position Specific Score Matrix) to find the p-value corresponding to one (or more) positions with high score. Next, we extended the concept to create POC, a program that searches for known enriched Transcription Factors' binding motifs in a cluster of genes, compared to a background set (which can be either all the known genes or other clusters - see figure 2). Knowing that the location within the promoter is often important for the function of the motifs, we search for a further enrichment in "windows", e.g. short segments in the promoter regions, and check for strand-specific enrichment. We analyze the changes in the PSSMs of the actual motifs found in given clusters and at specific strand and location restrictions. This way we are able to find PSSMs which are significantly different (or more 'conserved') than the original PSSM we have started from.

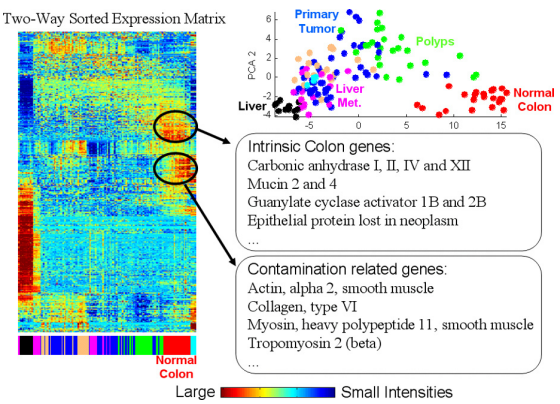

Fig. 1 *Left: two-way sorted expression matrix for 1,000 highest variance genes over 144 samples. Colored bar below indicates sample type. Top right: PCA image of the samples in the space of the 'intrinsic colon' gene-cluster (selected genes are listed below). We see a gradual decrease in expression, correlated with tumor progression. The second highlighted gene-cluster is also high in some normal colon samples, but is related to tissue composition rather then disease.*

The central topic of our work in biology is global gene expression analysis, mostly on cancer related problems. A selection from our current projects is briefly outlined below. As part of an international

colon cancer consortium we used the Affymetrix U133A GeneChip to study 144 microdissected samples from patients with colon cancer. The samples included 47 primary carcinomas; 23 polyps; 22 normal colon epithelium; 16 liver metastasis; 19 lung metastasis; 11 normal liver; and 5 normal lung. The expression matrix for the genes with highest variance was sorted by SPIN, as shown in figure 1. The reordered expression matrix clearly separates the samples according to tissue types (colon, lung, liver), and also tumors from normal tissue and adenomas from adenocarcinomas. Genes partition into several clusters; some are tissue specific and others are associated with clinical and pathological disease state. Samples of colon tissue that were isolated from liver metastases arrayed into an elongated, ellipsoid cluster extending from colon samples towards liver samples (a similar picture is seen for lung metastasis). The genes that induced the elongation were characteristic of liver (or lung), suggesting that this pattern reflects contamination of the metastasis samples with cells from the liver or lung. Once we identified the 'cleanest' samples, it became possible to focus on expression profiles associated with tumor progression.

We are involved in several studies of leukemia. In collaboration with E. Canaani's group we identified the target genes of a translocated MLL oncogen. We also discovered that among the ALL patients with the MLL translocation there is a subgroup of early differentiators. In another study we found evidence for a role of viral infection in early childhood leukemia.

Together with a group at the University Hospital at Lausanne we are studying Glioblastoma (GBM). The motivation is to study genetic alterations and biological mechanisms involved in disease initiation and progression, and uncover uncharacterized subtypes. Gene expression profiles of more than 60 GBM tumor samples were obtained. The search of inherent sub-divisions will be performed using unsupervised methods, such as CTWC clustering algorithm, and the SPIN sorting algorithm. The ability to differentiate response to therapy and outcome by means of molecular diagnostics would have great clinical impact since it would allow identification of subgroups of patients who are most likely to benefit from currently available therapies.

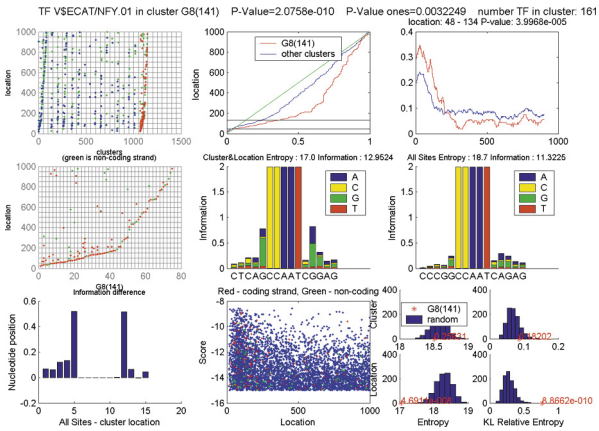

Fig. 2 *A typical output of POC, for NFY transcription factor, found to be significant in the "proliferation cluster" found in an experiment done in V. Rotter's group. The output includes information regarding significance (p-value) for location, enrichment in the cluster, strand and conserved PSSM.*

Selected Publications

Blatt, et. al. (1996) Super-paramagnetic Clustering of Data, Phys. Rev. Lett. 76, 3251-3254.

Getz, et. al. (2000) Coupled two-way clustering analysis of gene microarray data. PNAS 97,12079-12084.

Godard et. al. (2003) Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. Cancer Res. 2003 Oct 15;63(20):6613-25.

Rozovskaia, et. al. (2003), Expression profiles of acute lymphoblastic and myeloblastic leukemias with ALL-1 rearrangements, PNAS 100:7853.

Lotem, et. al. (2003)
Inhibition ofp53-induced apoptosis without affecting expression of p53-regulated genes, PNAS 100: 6718.

Quintana, et. al. (2003) Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: a bio-informatic approach to immune complexity. J Autoimmun. 2003 Aug;21(1): 65-75.

Tsafrir, et. al. (2004) Sorting Points Into Neighborhoods (SPIN), (manuscript in preparation)

Hertzberg L., Zuk O., Getz G. and Domany E. (2004) Finding Motifs in Promoter Regions (submitted).

Acknowledgements:

E. Domany is the incumbent of the H. J. Leir Professorial Chair. Our work was supported by grants from the NIH, GIF, Minerva, the ISF, the EC and the Ridgefield Foundation.