Eytan Domany

Himanshu Agrawal
Liat Ben-tovim
Hilah Gal
Gaddy Getz
Itai Kela
Osnat Ravid
Noam Shental
Daniel Volk

Cluster analysis of biological data - gene expression and antigen reactivity

Department of Physics of Complex Systems

Tel. 972 8 934 3964 **Fax.** 972 8 934 4109
E-mail: eytan.domany@weizmann.ac.il

We use advanced methods of data analysis, with emphasis on clustering methods developed in our group, to study problems in biology (DNA chip data analysis, antigen 'chip'); biochemistry (protein structure classification); physics (spin glasses) and information technology (document classification). Work done in collaboration with D. Givol, on identifying the primary targets of P53, is presented in detail in his poster.

The central topic of our work in biology is global gene expression analysis, mostly on cancer related problems. This work is done either in direct collaboration with the groups that perform the measurements, or by analyzing publicly available data.

Together with a group at the University Hospital at Lausanne we studied gene expression in 33 tissues obtained from human brain tumors (glioblastoma, GBM), and in 3 cell lines. Of the 33 tumors 17 were primary GBM (i.e. large tumors , fully developed at the time of diagnosis and biopsy), 12 - low grade astrocytoma and 4 - secondary GBM. For each tumor the results of several genetic tests (e.g. P53 mutation status, EGFR amplification, etc) were recorded, together with clinical information (age, survival, etc). The aim of the research was to identify (i) groups of genes that differentiate between primary and non-primary GBM (ii) genes whose expression levels partition tumors either according to genetic labels, or (iii) clinical information, or (iv) in a new, unanticipated way.

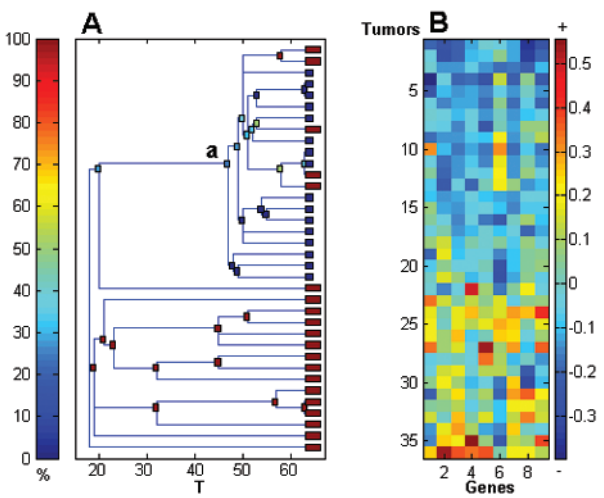

Fig. 1 *Clustering 36 glioblastoma, using the expression levels of 9 genes*

An example of our findings is identification of a cluster of 9 genes, some of which (such as VEGF and VEGFR) are related to angiogenesis. Fig. 1 presents clustering of the 36 GBM, using the expression levels (Fig. 1B) of these 9 genes. It is not surprising that such genes have considerably higher expression (red color in Fig 1B) in the primary GBM . In the dendrogram (Fig 1A) one sees a very stable cluster of samples, 'a', which contains mainly non-primary GBM. The boxes, representing clusters, are colored according to the % of primary GBM in each.

There were, however, also genes of unknown function in this cluster, whose high correlation with the angiogenesis related genes came as a surprise and has been tested and verified in an independent experiment. Another finding of interest concerns a cluster whose genes are related to the immune response and inflamation. These genes partition the tumors in an unexpected way; the low-grade and secondary cluster together with a group of primary GBM, taken from patients with relatively long survival.

In a different study, on breast cancer, we used publicly available expression data, of the Stanford group. Out of 65 tumors, 40 were paired, with samples taken before and after chemotherapy, to which 3 (out of 20) subjects responded positively. Using our methodology, we were able to extract from the data partitions that have eluded previous analysis. For example, we found a cluster of 33 genes, whose expression levels correlate well with the cells' proliferation rates. Using these genes to cluster the 65 tumors we found the dendrogram shown in Fig. 2A. The cluster marked 'a' contains samples with low proliferation rates (blue on 2B) - these are 'normal breast - like'. The 'before treatment' samples, taken from all three tumors for which chemotherapy succeeded, were in cluster 'b', of samples with intermediate proliferation rates; the corresponding 'after treatment' samples are in the 'normal breast' cluster. Hence the expression levels of

these 33 genes may have predictive power with respect to the success rate of the particular therapy that was used. The boxes in Fig 2A are colored according to the % of ER- samples.

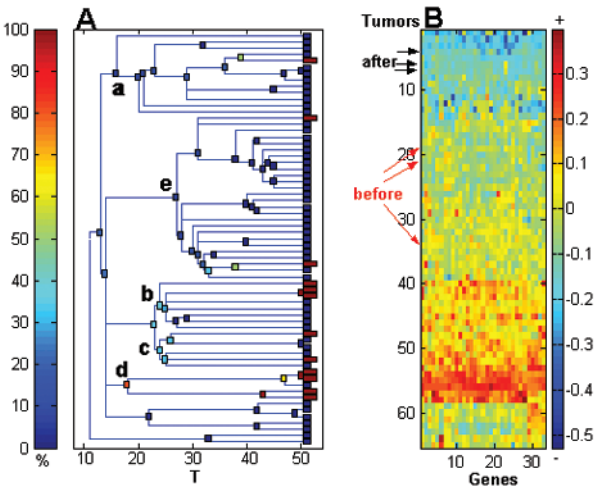

***Fig. 2** Clustering 65 breast cancer samples, using expression levels of 33 genes*

In collaboration with I. Cohen's group we analyzed data obtained by an 'antigen array'. In this experiment the reactivities of the antibodies contained in human serum are measured, in parallel, for 87 selected antigens. This measurement is repeated for 40 samples (sera collected from 20 healthy subjects and from 20 with diabetes. We identified clusters of antigens that have correlated reactivities over the sera. When some of these antigen clusters were used (one cluster at a time), to characterize the sera, we observed statistically significant separation into healthy subjects versus diabetes.

Selected Publications

Blatt, M., Wiseman, S. and Domany, E. (1996) Super-paramagnetic Clustering of Data, Phys. Rev. Lett. 76, 3251-3254.

Blatt, M., Wiseman, S. and Domany, E. (1997) Data Clustering Using a Model Granular Magnet. Neural Comp. 9, 1805-1842.

Getz, G., Levine, E. and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. PNAS 97, 12079-12084

Getz, G., Levine, E., Domany, E. and Zhang, M. Q. (2000) Super-paramagnetic clustering of yeast gene expression profiles, Physica A279, 457 - 464.

Kannan, K., Amariglio, N., Rechavi, G., Jakob-Hirsch, J., Kela, I., Kaminski, N., Getz, G. Domany, E. and Givol, D. (2001) DNA microarrays identification of primary and secondary target genes regulated by p53, Oncogene 20, 2225 - 2234.

Quintana, F. J., Getz, G., Hed, G., Domany, E. and Cohen, I. R. (2002) Cluster analysis of human autoantibody reactivities in health and in type 1 diabetes mellitus: A bio-informatic approach to immune complexity. (submitted)

Getz, G, Vendruscolo, M., Sachs D. and Domany, E. (2001) Automated assignment of SCOP and CATH protein structure classification from FSSP scores. Proteins (in press).

Hed, G., Hartmann, A. K., Stauffer, D. and Domany, E. (2001) Spin domains generate hierarchical ground state structure in J = +/- 1 spin glasses. Phys. Rev. Lett. 86, 3148-3141.

Volk, D. and Stepanov M. G. (2002) Resampling methods for document clustering (submitted).

Goddard, S., Getz, G., Kobayashi, H., Nozaki, M.A., Diserens, A. C., Hamou, M. F., Stupp, R., Janzer, R.C.,Bucher, P., de Tribolet, N., Domany E. and Hegi, M.E. (2002) Gene expression profiling divides gliomas into distinct subgroups, (manuscript in preparation).

Kela, I., Getz G. and Domany E (2002) Coupled two-way cluster analysis of breast cancer data, (manuscript in preparation).

Acknowledgements

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

For additional information see:

www.weizmann.ac.il/physics/complex/ compphys