

Simple Discriminant functions identify numerous small sets of genes that stratify breast cancer grades and stages

Gul S. Dalgin¹, Charles DeLisi^{2,3}

Keywords: breast cancer, gene expression, diagnostic genes, Bayesian discriminant analysis

Breast cancer, even at an early stage, is a very heterogeneous disease in which the alterations in molecular mechanisms affecting tumor growth, progression and metastatic potential vary among individual tumors. High-throughput gene expression profiling can identify sets of genes that are over or underexpressed in one or another phenotype; e.g. sets of over or underexpressed genes stratify closely related diseases, but the sets produced to date are too large to be economically viable diagnostics. We use a hybrid decision tree-discriminant analysis to identify pairs of genes whose joint expression distribution separates tissue samples in different stages and grades of malignancy.

The analysis was applied to the gene expression data generated by Ma and colleagues [2]. The method employs supervised classification technique which divides the samples randomly into training and test sets in each simulation. The results of testing single genes and gene pairs that correctly partition all samples in the training set are summarized in Table 1. The genes were ranked according to their minimum Euclidean distance between the expression values of the genes in tumor and normal samples.

	Number of single classifier genes	Number of pairs (genes involved)
Normal-ADH	10	2136 (336 genes)
Normal-DCIS I	11	8515 (502 genes)
Normal-DCIS II	18	8087 (455 genes)
Normal-DCIS III	24	12520 (670 genes)
Normal-IDC I	56	12836 (564 genes)
Normal-IDC II	23	15948 (649 genes)
Normal-IDC III	26	20823 (743 genes)

Table 1: Summary of number of single genes and gene-pairs identified for each normal and tumor stage comparison.

In particular we were able to distinguish normal from premalignant stage (ADH), three different grades of preinvasive (DCIS) and invasive (IDC) breast cancers using no more than two genes in each instance. Hierarchical clustering of the samples using the expression profiles of single genes revealed that each set of genes distinguish normal from the specific tumor samples. Some of these genes include previously identified cancer related genes such as Angiopoitein-like 4, which is known to be important in sustained angiogenesis; Matrix metalloproteinase 7, which was found to

¹ Molecular Biology, Cell Biology and Biochemistry Program, Boston University, 2 Cummington Street, Boston, MA 02215, E-mail: sdalgin@bu.edu

² Dept. of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, E-mail: delisi@bu.edu

³ Bioinformatics Graduate Program, Boston University, 44 Cummington Street, Boston, MA 02215

be up-regulated in colorectal carcinomas [3] and Glutamine synthase, which is also up-regulated in tumor and important in tumor progression [1]. Grade specific genes also agree with previous findings. As an example, BIRC5 (survivin) gene, which is known to be overexpressed in common human cancers and was found to be correlated with Grade III (poor prognosis) tumors [2], was identified only in Grade III tumors in this study.

Although a diagnostic signature is obtained using only a pair of genes, the total sum of all such pairs includes a large number of genes (Table 1), and that provides an entrée into the search for targets. We briefly discuss the identification of biological processes that are enriched in subsets of such gene groups.

[1] Dang, C.V. and Semenza, G.L. 1999. Oncogenic alterations of metabolism. *TIBS Reviews* 24:68-72.

[2] Ma, X.J., Salunga, R., Tuggle, J.T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B.M., Zhou, Y.X., Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T.M., Erlander, M.G. and Sgroi, D.C. 2003. Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences USA* 100:5974-5979.

[3] Masaki, T., Matsuoka, H., Sugiyama, M., Abe, N., Goto, A., Sakamoto, A. and Atomi, Y. 2001. Matrilysin (MMP-7) as a significant determinant of malignant potential of early invasive colorectal carcinomas. *British Journal of Cancer* 84:1317-1321.