Robust meta-analysis of genomic data for cancer diagnosis

Gyan Bhanot[1], Gabriela Alexe[2], Babu Vengataraghavan[3], Ramakrishna Ramaswamy[4], Jorge Lepre[5], Arnold J Levine[6], Gustavo Stolovitzky[7]

Keywords: patterns, combinatorial biomarkers, meta-analysis, meta-classifiers, cancer diagnosis

1 Introduction.

The rapid development of microarray technologies allows the analysis of gene expression patterns to identify subsets of genes which are differentially expressed between different phenotypes (e.g., different types of cancer), and to integrate data into personalized models capable of providing diagnosis and predicting prognosis. There is a lot of ongoing research in developing tools and methodologies to extract information from biomedical data (e.g., [1]). However, there remains a need to integrate the results of these tools with existing biological knowledge to extract information valuable for medical diagnosis.

In this presentation we describe an approach we have recently developed ([2], [3]) and which addresses these issues.

2 Methods

Our approach integrates several machine learning techniques and robust noise analysis on data obtained from different platforms to identify phenotypes and biomarkers from gene array and mass spectrometry data. An important ingredient of our technique is the use of patterns extracted from data as synthetic variables which define boundaries on gene expression values for separating the phenotypes. The abstract space of these synthetic variables provides additional structural information about the phenotype and it is used for training the machine learning tools. The predictions provided by the individual machine learning tools are integrated by weighed voting into a single predictor. If necessary, a principal component analysis can be applied before combining the different predictors.

We show how the use of our tool along with biological information (about the p53 pathway) results in finding groups of genes that are good predictors of the cancer phenotype. This is done in the

[1] IBM Computational Biology Center, IBM Research, Yorktown Heights, New York, NY, USA. E-mail: gyan@us.ibm.com
[2] Institute for Advanced Study, Einstein Drive, Princeton, NJ, USA. E-mail: galexe@ias.edu
[3] Institute for Advanced Study, Einstein Drive, Princeton, NJ, USA. E-mail: babu@ias.edu
[4] Institute for Advanced Study, Einstein Drive, Princeton, NJ, USA. E-mail: rama@ias.edu
[5] IBM Computational Biology Center, IBM Research, Yorktown Heights, New York, NY, USA. E-mail: jorge@us.ibm.com
[6] Institute for Advanced Study, Einstein Drive, Princeton, NJ, USA. E-mail: ajlevine@ias.edu
[7] IBM Computational Biology Center, IBM Research, Yorktown Heights, New York, NY, USA. E-mail: gustavo@us.ibm.com

context of studying the progression of follicular lymphoma into diffuse large B-cell lymphoma, and also in identifying robust clusters of genes which can separate different breast cancer subtypes.

We demonstrate the effectiveness of our approach in diagnosis on several cancer gene expression data, including the oligonucleotide microarray gene expression data of Shipp et al. ([4]), the Affymetrix gene expression data produced by DallaFavera laboratory at Columbia University ([5]), and the cDNA micrarray data of Botstein et al. ([6]).

Results

Our pattern-based meta-classification technique achieves higher predictive accuracies than each of the individual classifiers trained on the same dataset, is robust against various data perturbations and provides subsets of predictive genes. For example, figure 1 presents the error distributions on the test lymphoma datasets ([4], [5]) of the of the meta-classifier and of the individual classifiers trained on raw and on pattern data, respectively (a dot corresponds to an error). Notice that the predictions of the meta-classifier are better than the predictions of any individual classifier.

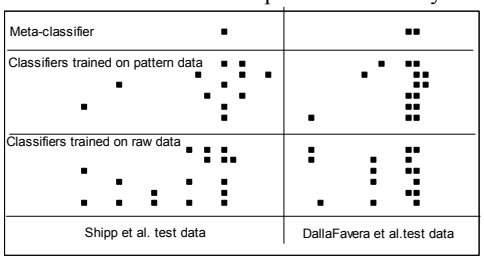

Figure 1. Error distribution of the meta-classifier and of the individual classifiers trained on raw and pattern data

We also find that combinations of p53 responsive genes are highly predictive of phenotype. For example, we find that in diffuse large B cell lymphoma cases, the mRNA level of at least one of the three genes p53, PLK1 and CDK2 is elevated, while in follicular lymphoma cases, the mRNA level of at most one of them is elevated.

4 References

[1] Califano, A., Stolovitzky, G., and Tu, Y. 2000. Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the International Conference on Intelligent Systems in Molecular Biology* 8:75-85.

[2] Bhanot, G., Alexe, G., Venkataraghavan, B., and Levine, A.J. 2004. A robust meta-classification strategy for cancer detection from mass spectrometry data. Submitted.

[3] Alexe, G., Bhanot, G., Levine, A.J. and Stolovitzky, G. 2005. Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories Submitted.

[4] Shipp, M.A., Ross, K.N., Tamayo, P., Weng, A.P., Kutok, J.L., Aguiar, R.C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G.S., Ray, T.S., Koval, M.A., Last, K.W., Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E.S., Aster, J.C., Golub, T.R. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8: 68-74.

[5] Stolovitzky G. (2005) Gene selection strategies in microarray expression data: applications to case-control studies. In Deisboeck T.S., Kresh J.Y., and Kepler T.B. editors, *Complex Systems Science in BioMedicine*. Kluwer/Plenum Publishers, NY in press (preprint: http://www.wkap.nl/prod/a/Stolovitzky.pdf).

[6] Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., et al. 2003. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences USA*. 100: 8418-23.