A Robust Meta-Classification Strategy for Cancer Diagnosis from Gene Expression Data

Gabriela Alexe
IBM Computational Biology Center, IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA
galexe@us.ibm.com

Gyan Bhanot
IBM Computational Biology Center, IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA, and Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, USA
gyan@us.ibm.com

Babu Venkataraghavan
Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, USA
babu@ias.edu

Ramakrishna Ramaswamy
Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, USA
rama@ias.edu

Jorge Lepre
IBM Computational Biology Center, IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA,
leprej@us.ibm.com

Arnold J Levine
Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, USA, and Robert Wood Johnson School of Medicine and Dentistry, Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA
alevine@ias.edu

Gustavo Stolovitzky
IBM Computational Biology Center, IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA
gustavo@us.ibm.com

Abstract

One of the major challenges in cancer diagnosis from microarray data is to develop robust classification models which are independent of the analysis techniques used and can combine data from different laboratories. We propose a meta-classification scheme which uses a robust multivariate gene selection procedure and integrates the results of several machine learning tools trained on raw and pattern data. We validate our method by applying it to distinguish diffuse large B-cell lymphoma (DLBCL) from follicular lymphoma (FL) on two independent datasets: the HuGeneFL Affmetrixy dataset of Shipp et al. (www. genome.wi.mit.du/MPR /lymphoma) and the Hu95Av2 Affymetrix dataset (DallaFavera's laboratory, Columbia University). Our meta-classification technique achieves higher predictive accuracies than each of the individual classifiers trained on the same dataset and is robust against various data perturbations. We also find that combinations of p53 responsive genes (e.g., p53, PLK1 and CDK2) are highly predictive of the phenotype.

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], [2]). However, there remains a need to integrate the results of these tools with existing biological knowledge to extract information valuable for medical diagnosis. The aim of this study is to present such a tool, recently developed for cancer detection from mass spectrometry data ([3]), and to adapt it for cancer diagnosis from gene expression data.

We demonstrate our approach by creating a diagnosis model to accurately distinguish between follicular lymphoma (FL) and diffuse large B-cell lymphoma (DLBCL). We use the oligonucleotide microarray gene expression data of Shipp et al. ([4], WI data), and validate our findings on a separate Affymetrix gene expression data produced by DallaFavera laboratory at Columbia University (CU data, see [5]). The WI and CU datasets report gene expression data for DLBCL and FL cases which were obtained by using different Affymetrix chips (HuGeneFL chip for WI dataset and Hu95Av2 for the CU dataset). We also show that one can combine the two datasets into a single meta-dataset, while maintaining the accuracy of predictions.

Using our meta-classification method on a training subset of the WI data, we identified a robust subset of 30 predictive genes and constructed a meta-classifier which misclassified only one FL case when validated on the test set of the WI data and misclassified only two FL cases when validated on the external CU data. We obtained further biological insight by focusing on the subset of p53 responsive genes and extracted relevant patterns characteristic of FL and DLBCL. In particular, we showed that the combination of the gene expression of p53, PLK1 and CDK2 is an accurate biomarker for distinguishing FL from DLBCL. Currently our research is oriented on integrating the meta-classification tool with unsupervised consensus clustering techniques and applying it to discriminate between different breast cancer subtypes from various microarray platforms (e.g., Affymetrix, cDNA, Agilent).

2. Methods

Our approach integrates several machine learning techniques and robust noise analysis on data obtained from different platforms to identify phenotypes and robust 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. Each pattern can be interpreted as a synthetic 0-1 variable associated with the samples in the dataset, the value 1 being assigned when the corresponding sample satisfies the defining conditions of the pattern, and the value 0 otherwise. Each sample is then represented by a vector with 0-1 entries, where each entry corresponds to a pattern. In this way, the original data can be represented in an abstract space which we call "pattern data". The abstract pattern data provides additional structural information about the phenotype and it is used in our approach for training various individual machine learning tools (e.g., support vector machines, artificial neural networks, decision trees, random forests, weighted voting and *k*-nearest neighborhood systems, etc). We have recently developed ([6]) an efficient algorithm for exhaustive pattern extraction from biomedical data.

Our method starts by applying a pattern-based multivariate approach (see e.g., [2]) to identify a subset of predictive genes out of a pool of genes by requiring them to satisfy stringent filtering criteria. Next, we combine the predictions of several machine learning tools trained on the subset of predictive genes and on pattern data with the aim of producing an accurate predictor. It is well-known (e.g., [7]) that combining individual classifiers with independent error distributions into a meta-classifier has the effect of improving the error rate. In our method this effect is further boosted by using pattern data.

We showed how the use of our tool along with biological information (about the p53 pathway) results in finding combinations of genes that are good predictors of the cancer phenotype. This was done in the context of studying the progression of follicular lymphoma into diffuse large B-cell lymphoma; currently, we follow the idea of [8] and apply consensus clustering and meta-classification techniques to microarray data from various platforms to identify robust clusters of genes which can separate between different breast cancer subtypes.

Figure 1 presents the flow chart of our meta-classification approach.

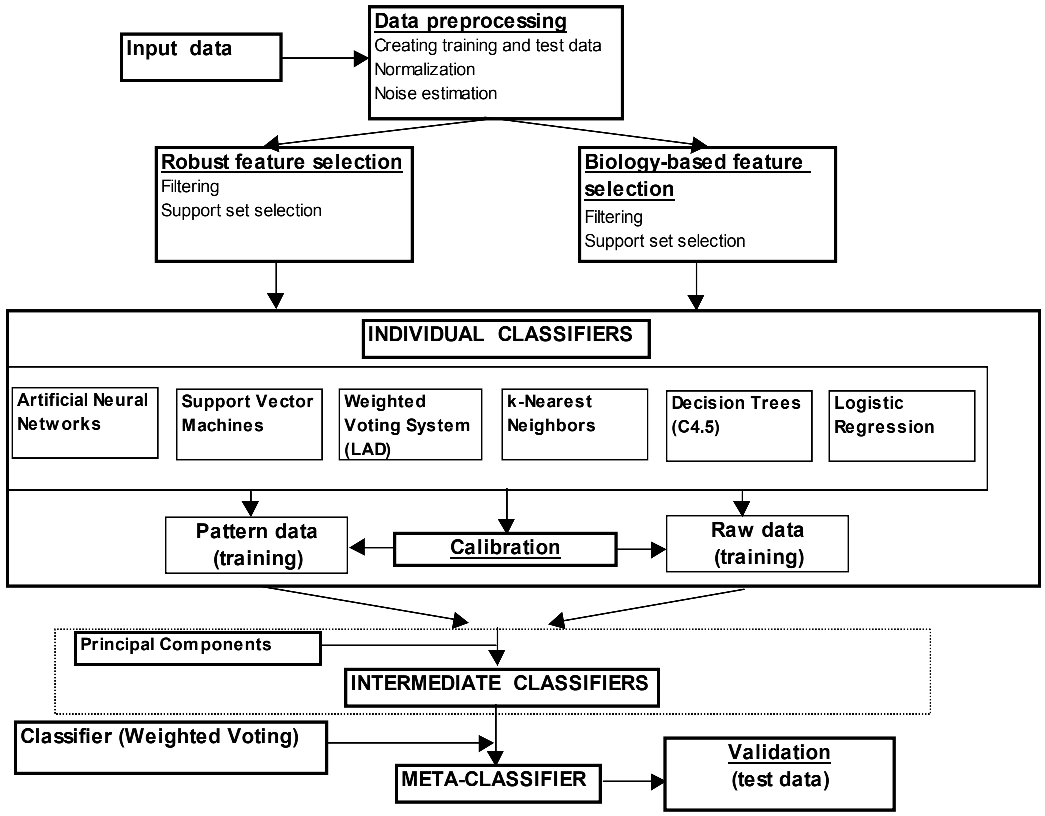

Figure 1. Flow chart of the meta-classifier approach.

3. 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 2 presents the error distributions on the test lymphoma datasets ([4], [5]) of the meta-classifier and of the individual classifiers trained on raw and on pattern data, respectively (a dot represents an error). Notice that the predictions of the meta-classifier are better than the predictions of any individual classifier.

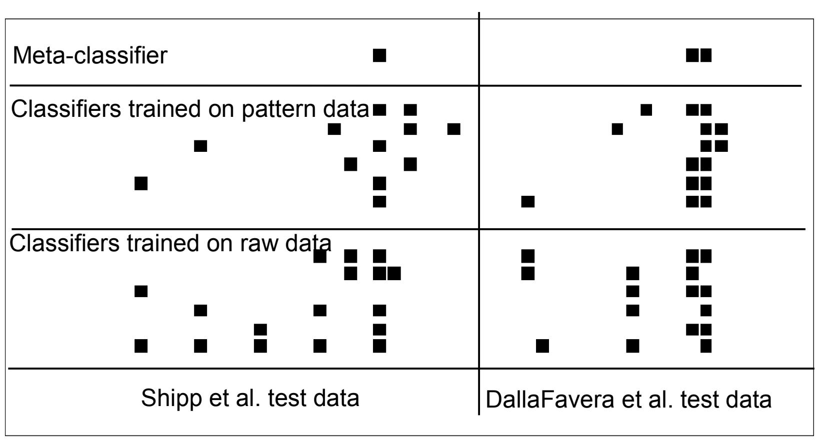

Figure 2. Error distribution of the meta-classifier and of the individual classifiers trained on raw and pattern data

Numerous studies e.g., [9], [10], associated a correlation between overexpression of p53 and FL progression to DLBCL, and also showed that mutations of p53 are associated with histologic transformation in approximately 25% to 30% of FL cases. Other studies e.g., [11], [12] suggested that over-expression of MDM2 (and p53) identifies DLBCL and poor prognosis for FL cases, while altering the feedback loop p53-MDM2. We also found that combinations of p53 responsive genes are highly predictive of phenotype. For example, we found that in 80% of the 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 80% of the follicular lymphoma cases, the mRNA level of at most one of them is elevated.

4. References

[1] Califano A., G. Stolovitzky, and Y. Tu, 2000. Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the International Conference on Intelligent Systems in Molecular Biology* ISMB 2000, pp. 5-85.

[2] Alexe, G., Alexe, S., Axelrod, E.D., Hammer, P.L. and Weissmann, D. Logical analysis of diffuse large B-cell lymphomas. *Artificial Intelligence in Medicine*, in press.

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

[4] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok,, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, and T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine* 8, 2000, pp. 68-74.

[5] Stolovitzky, G. 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, New York, in press (preprint: http://www.wkap.nl/prod/a/Stolovitzky.pdf).

[6] G. Alexe and P.L. Hammer, Spanned patterns in Logical Analysis of Data, *Discrete Applied Mathematics,* in press.

[7] Merz, C. Classification and Regression by Combining Models. Disertation, .University of California at Irvine, 1998.

[8] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J.S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C.M. Perou, P.E. Lonning, P.O. Brown, A.L. Borresen-Dale, and D.Botstein, Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences USA*. 100, 2003, pp. 8418-8423.

[9] C.A. Sander, T. Yano, H.M. Clark, C. Harris, D.L. Longo, E.S. Jaffe, and M. Raffeld, p53 mutation is associated with progression in follicular lymphomas, *Blood*, 82(7), 1993, pp. 1994-2004.

[10] F. Lo Coco, G. Gaidano, G., D.C. Louie, K. Offit, R.S. Chaganti, and R. Dalla-Favera, p53 mutations are associated with histologic transformation of follicular lymphoma, *Blood*, 82(8), 1993, pp. 2289-2295.

[11] M.B. Moller, O. Nielsen, and N.T. Pedersen, Oncoprotein MDM2 overexpression is associated with poor prognosis in distinct non-Hodgkin's lymphoma entities, *Mod. Pathol.*, 12(11), 1999, pp. 1010-1016.

[12] K.S. Elenitoba-Johnson, S.D. Jenson, R.T. Abbott, R.A. Palais, S.D. Bohling, Z. Lin, S. Tripp, P.J. Shami, L.Y. Wang, R.W. Coupland, R. Buckstein, B. Perez-Ordonez, S.L. Perkins, I.D. Dube, and M.S. Lim, Involvement of multiple signaling pathways in follicular lymphoma transformation: p38-mitogen-activated protein kinase as a target for therapy, *Proc. Natl. Acad. Sci. U.S.A.*, 100(12), 2003, pp. 7259-7264.