European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Exploring Classifiability Metrics for Selecting Informative Genes

James Wu[a], and Ioannis P. Androulakis [a,b,*]
[a]Chemical & Biochemical Engineering Department and
[b]Biomedical Engineering Department
Rutgers, The State University of New Jersey
Piscataway, NJ 08854

Abstract

Microarray experiments are emerging as one of the main driving forces in modern biology. By allowing the simultaneous monitoring of the expression of the entire genome for a given organism, array experiments provide tremendous insight into the fundamental biological processes that translate genetic information. One of the major challenges is to identify computationally efficient and biologically meaningful analysis approaches to extract the most informative and unbiased components of the microarray data. In this paper we introduce a framework that integrates machine learning and optimization techniques for the selection of maximally informative genes in microarray expression experiments. The proposed approach provides tremendous reduction in the number of informative genes, compared to similar analyses by generating biologically relevant minimal subsets of genes.

Keywords: microarray experiments, feature selection, machine learning

1. Microarray experiments: Brief introduction and major limitations

The goal of modern biology is to bridge the gap between the genetic information at its most elementary level and the collective expression of behaviour. The successive steps for translating sequence to structure to function are highly complex and cannot be modelled strictly from first principles. Experimental techniques have been developed that have revolutionized the way we look at complex biological systems since they allow to monitor changes during the process of transforming genetic information.

The genetic information is stored in the DNA. In order for the genome to direct, or affect, changes in the cell a transcriptional program must be activated dictating all biological transformations. This program is regulated temporarily according to an intrinsic program or in response to changes in the environment. The expression of the genetic information, stored in DNA, takes places in two stages: transcription, during which DNA is transcribed into mRNA, and translation, during which mRNA provides

* Author to whom correspondence should be addressed: yannis@rci.rutgers.edu

the blue-print for the production of specific proteins. Measuring the level of production of mRNA, thus measuring the expression levels of the associated genes, provides a quantitative assessment of the levels of production of the corresponding proteins. Innovative approaches such as cDNA and oligonucleotide microarrays were recently developed to extract genome-wide information related to gene expression (Brown and Botsein, 1999; Cheung et al., 1999; Dudoit et al., 2000). A number of experiments can thus be designed to address a variety of issues. For instance:

1. Diversion from normal physiology is frequently accompanied by changes in gene expression patterns. Therefore, genes inappropriately transcribed cause diseases like cancer. Comparison of the expression profiles of such cells provides the basis for the understanding of the genetic causes of a disease.
2. By monitoring the changes in the expression levels of a genome in the presence of environmental changes provides the beginning for a fundamental understanding of the causes of the response in the presence of an environmental stimulus.

One of the major challenges is to extract in a systematic and rigorous way the biologically relevant components from the array experiments in order to establish meaningful connections linking genetic information and cellular function. Because of the significant amount of experimental information that is generated (expression levels of thousands of genes) computer-assisted knowledge extraction processes are the only realistic alternative for managing such an information deluge. Array experiments are characterized by a number of inherent limitations, which have to be well understood before any analysis attempts are made. Given the complexities of biological functions, it is not necessary that co-expressed genes be also co-regulated. In array experiments the generated data are interpretations of measurements rather than hard data. Furthermore, tremendous variability and uncertainty exists not only because of biological fluctuations, but also as a result of the processing of the experiment itself. Finally, although a large number of genes are monitored during the experiment we must also realize that, in general, we have a very limited number of cells that are analyzed and an even smaller, if any, number of repeats to statistically validate the robustness of the measurement. Simply put, we have a much larger number of independent (input) variables that we measure compared with the number of experiments (output variables) that we generate. In principle, when the ratio of experiments/variables is very small it is highly unlikely that we can correctly capture the inherent non-linear structure of the experiments and the relationship between input and output variables.

A number of excellent publications have focused on different aspects of gene expression experiments, primarily for clustering of cells and genes (Alizadeh et al., 2000; Alon et al., 2000; Golub et al., 1999). The development of novel computational approaches that exploit large warehouses of gene expression data have been identified as major enablers for realizing fully the potential of this technology (Basset et al., 1999). Even though these approaches were very successful at, implicitly, reducing the number of putative genes with significant signature characteristics the resulting models still involve a significant number of genes, often in the hundreds. In our prior work (Androulakis, 2004) we presented a framework that identifies the minimum subset of informative genes while imposing the maximum possible simplicity of the model describing the

data. The approach will be briefly summarized in the following section. In this paper we extend the proposed model to describe the topology of the distribution of data that (1.1) results from the selection of the maximally informative genes. We demonstrate how the incorporation of explicit topological complexity measures identifies not only robust but also highly biologically relevant solutions.

2. An integrated Optimization Machine Learning approach for the selection of informative features

In a recent publication (Androulakis, 2004) a model was presented which is based on classification trees (Breinman et al., 1984, Quinlan, 1993) as well as a thorough review of the basic principles characterizing the feature selection problem and its complexities in the context of machine learning literature in general and in selecting informative genes in particular. The fundamental assumption of the approach is that the minimum set of maximally informative genes is the one that produces the least complex decision tree. The complexity of the decision tree is determined according to the number of genes used for the classifier and the number of rules that comprise the tree. The issue of simplicity in classification trees has long been advocated as a rule for building robust classifiers (Fayyad, 1990). The framework is put together in a large non-linear combinatorial optimization problem as follows:

$$
\begin{aligned}
&\min \left\| C - C' \right\| \\
&\quad \text{subject to:} \\
&\qquad C = T(\lambda_i, i = 1, ..., N) \\
&\qquad \min \text{Complexity} = \text{Number of Rules} \\
&\qquad \min \sum_{i=1}^{N} \lambda_i, \quad \lambda_i = \begin{cases} 1, g_i \in I_G \subseteq G \\ 0, g_i \notin I_G \subseteq G \end{cases}
\end{aligned}
\tag{1}
$$

The objective in (1) measures the accuracy of the classifier. C and C' are vectors containing the class assignment of the samples. C' denotes the actual assignment whereas C is the assignment derived based on the classifier. The latter depends of the number of features and the particular decision tree that is derived and is implicitly defined via the use of the classifier, denoted in our formulation as C=T (λ_i, i=1...N). The "norm" ||C-C'|| can be defined in a number of different ways: count of the number of erroneous predictions, percent of erroneous prediction, etc. Once the classifier has been applied to a given set of features its complexity, in terms of the number of classification rules required, is identified and is denoted as "Complexity" in (1). This defines another level of optimization as we search for the minimum possible complexity in the classifier. Finally, feature selections are modelled through the use of appropriate binary variables. The value of the binary variable is 1 if the particular gene is to be incorporated in the classifier, 0 otherwise. The set of informative genes, I_G, is a subset of the original set of genes. The details of the solution methodology are presented in (Androulakis, 2004). It should be pointed out that the fundamental hypothesis defining

the relationship between informative features and the complexity of the classification rules does not dependent on the classification algorithm.

3. Incorporating the complexity of the classification problem

Formulation (1) deals only with the complexity of the classification model, i.e., number of features used and number of rules in the model. However, these conditions do not capture the geometric characteristics of the space partitioning achieved by the distribution of the feature values. What we would like to incorporate is also a "method-independent" metric of complexity. For that reason we have explored the concept of "separating boundaries" (Ho, 2002). A measure of the complexity of the boundary separating the classes is given by the "boundary length" defined as the percent points on an edge connecting two opposite classes in the minimum spanning tree (MST) connecting all samples (Friedman and Rafksy, 1979). For multi-class problems we introduce as our geometric complexity metric (GCM) the following quantity: $\mathrm{GCM} = \sum_{i} \sum_{j} \frac{\mathrm{MST}_{i \to j}}{\min\left\{ d_{i \to j} \right\}}$. According to this definition, we evaluate the MST of our data, based on the particular selection of genes, and for each pair of classes (i,j) we determine the separating boundary (arcs of the tree connecting i and j) and normalize with respect to the minimum arc between the two classes. The case where no arc between two classes exists is appropriately taken into account. The goal is to achieve a spatial partitioning of the data that generates the least number of inter-class MST arcs whereas the class separability is maximized. Thus, our formulation incorporates one extra level in order to minimize the geometry complexity metric as well.

3.1 Computational Results

The aforementioned formulation has been successfully applied to a variety of problems described in *Table 1*. Our proposed methodology outperforms any other method in its ability to significantly reduce the number of "informative" genes. The key points will be illustrated using the "Small Round Blue Cell Tumours" (SRBCT) as our motivating example. It will be demonstrated how the "computationally" informative features are also biologically relevant. SRBCt is a descriptive category encompassing a large number of malignant tumours that tend to occur in childhood. They are united by their similar histo-pathological appearance. However, subtle clues may be present to distinguish between the tumours. For proper characterization pathologists often employ immunohistochemistry, electron microscopy, and molecular analysis for chromosomal abnormalities. The SRBCTs include neuroblastoma (NB), rhabdomyosarcoma (RMS), non Hodgkin lymphoma (NHL/Burkitt Lymphoma), and the Ewing family of tumours (EWS). Currently no single biological or chemical test exists that can detect SRBCTs. Khan et al. (2001) presented a comprehensive study in which a large number of genes were monitored. Their analysis includes a training set of 63 samples and a blind set of 20 samples for testing. Overall they identified a sub-set containing 96 most informative genes by performing an exhaustive sensitivity analysis with a model combining PCA and artificial neural networks. Our approach however has the ability to explicitly incorporate the various complexity metrics thus leading to an optimal solution that has

capture the essence of the experimental data. We identified that the minimum numbed of informative genes is indeed 3. However, the analysis has to be parametric in terms of geometric complexity. The parametric analysis is easily incorporated by replacing the inner optimizations by equality constraints, i.e., setting the desired levels of complexity and number of genes. We have performed numerous parametric analyses. For illustration purposes we discuss the two "simpler" models. The optimal solution with 9 rules gives a more complex separating boundary (GCM = 12.0) than the optimal solution with 11 rules (CM = 5.4), both solution use 3 genes. The reason is that the second solution selects variables that have a wider range of distribution values. What is, however, far more significant is the analysis of the solutions that we generated. Specifically:

1. *Solution 1*: 3 genes, 9 rules, complexity metric 12.0, active genes: MIC2, FGF4, CTNNA1. No misclassified samples in the training test, 1/20 misclassified testing samples.
2. *Solution 2*: 3 genes, 11 rules, complexity metric 5.4, active genes: MIC2, IGF2, MAP1B. No misclassified samples in the training test, 2/20 misclassified testing samples.

Table 1. Data sets employed in the analyses

Data Set	*Data Source*	*Genes*	*Samples*	*Samples/Genes*	*Classes*	*Informative genes*
SRBCT[1]	Nat. Med, **7**:673(2001)	2303	63(20)[2]	0.04	4	3
Colon Cancer	PNAS, **96**:6745(1999)	2000	62	0.03	2	3
CMM[3]	Nature, **406**:536(2000)	8067	31	0.04	2	1
GIST[4]	Can. Res., **61**:8624(2001)	1987	18	0.01	2	1
Leukemia	Science, **286**:531(1999)	2000	38(34)[5]	0.04	2	2
Breast Cancer	NEJM, **8**:539(2001)	3226	22	0.01	3	2
HPC[6]	Can. Res., **61**:4663(2001)	6500	25	0.01	2	1

(1) Small Round Blue Cell Tumours, (2) 63 training samples, 20 testing samples, (3) Cutaneous Malignant Melanoma, (4) Gastrointestinal Stromal Tumour, (5) 38 training samples, 34 testing samples, (6) Human Prostate Cancer

According to the rule sets we have generated, FGF4 is mostly responsible for classifying EWS samples, whereas MIC2 for EWS samples. Even though this is a purely computational observation, MIC2 is indeed used to diagnoses EWS, whereas FGF4 is known to be actively related to myogenesis (Khan et al., 2001). Furthermore, IGF2 has been reported in RMS is various studies. Our computational results, in terms of variables that discriminate between RMS and EWS are very consistent with the leading biological hypotheses regarding these two classes. The two remaining genes in

our respective models, CTNNA1and MAP1 are known to be abnormally expressed in cancerous cells. Both solutions are depicted in *Figure 1*.

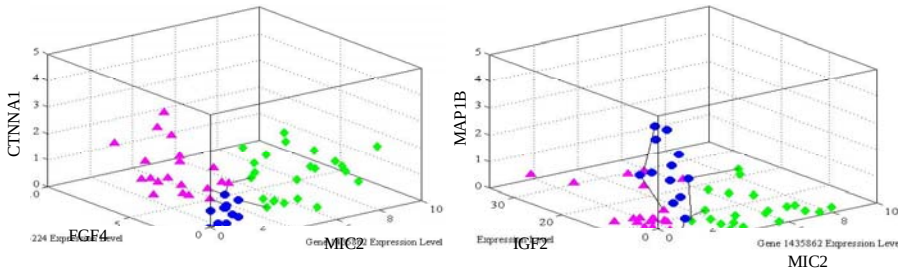

Figure 1: Distribution of expression data points for solutions 1 (left) and 2 (right). Also depicted are the intra-class minimum spanning tress connections.

4. Conclusions and future work

We have demonstrated how our novel methodology for informative gene selection has not only produced minimal sets of maximally informative genes that build accurate classifiers, but also we proved that the genes that were selected are also biologically relevant and can be used as potential targets. Currently we are exploring more efficient solution methodologies as well as extensions of the approach to temporal gene expression data.

References

Androulakis, I.P., 2004, Comp. Chem. Eng., accepted for publication

Alizadeh, A.A., M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitrage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Bodstein, P.O. Brown, and L.M. Staudt, 2000, Nature, 403, 503

Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, 1999, PNAS, 96,6745

Basset, D.E., M.B. Eisen, and M.S. Boguski, 1999, Nature Genetics, 21, 51

Breiman L, J.H., Friedman, R.A. Olshen, and C.J. Stone, 1984, Classification and Regression Trees. Chapman & Hall (Wadsworth, Inc.), New York.

Brown, P.P., and D. Botstein, 1999, Exploring the new world of the genome with DNA microarrays, Nature Genetics, 21, 33

Cheung, V.G., M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati and C. Childs, 1999, Nature Genetics, 21, 15

Dudoit, A., Y.H. Wang, M.J. Callow, and T.P. Speed, 2000, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Report 578, Stanford University

Friedman, J.H., and L.C Rafsky, 1979, Annals Stat., 7, 697

Golub, T.R., D.K. Slomin, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, 1999, Science, 286:531

Ho, T.K., 2002, Pattern Analysis and Applications, 5, 102

Khan, J., J.S. Wei, M. Ringer, L.H. Saal, M. Landanyi, F. Westerman, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, and P. S. Meltzer, 1991, Nature Medicine, 7, 673

Quinlan, R.J., 1993, C4.5: Programs for machine learning, Morgan Kaufmann Publishers