Selecting Maximally Informative Genes: the Interplay between Accuracy and Complexity

James Wu[1] and Ioannis P. Androulakis[1,2,#]
Department of [1]Chemical and Biochemical Engineering and [2]Biomedical Engineering
Rutgers, The State University of New Jersey
[#]E-mail: yannis@rci.rutgers.edu

Abstract

Microarray experiments are emerging as one of the main driving forces in modern biology. Via 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. We explore the relationship between computational complexity, robustness, and biological relevance. We formulate the problem of identifying maximally informative genes as a combinatorial optimization problem and demonstrate how the combination of integer optimization and machine learning approaches produces biologically interpretable sets of informative genes with strong biological implications. We suggest how to analyze the complexity of the model and how to incorporate complexity issues in the selection process. We demonstrate our methodology using numerous publicly available microarray datasets. Finally, we comment on the computational complexity of our approach and on necessary algorithmic and computational developments for achieving optimal efficiency.

1. Microarray experiments: a brief introduction and major limitations

The goal of modern biology is to bridge the gap between genetic information at its most elementary level and the collective expression of behavior. For the genome to direct changes in the cell, activated transcriptional programs intrinsically dictate all biological transformations. Expression of the genetic information, stored in DNA, takes places in two stages: transcription and translation. Measuring mRNA production level, i.e., measuring gene expression levels provides a possible quantitative assessment of corresponding levels of protein production. Innovative approaches were recently developed to quantify genome-wide gene expression [1]. One major challenge is to extract in a systematic and rigorous way the biologically relevant components needed to establish meaningful connections linking genetic information and cellular function. Because of the amount of experimental information generated, computer-assisted knowledge extraction processes are the only realistic alternative for managing such an information deluge. However, a number of inherent limitations need to be analyzed before attempting any analysis. In array experiments, the generated data are interpretations of measurements rather than hard data. Tremendous variability and uncertainty exist, not only because of biological fluctuations but also because of the processing of the experiment. Although a large number of genes are monitored during the experiment, we must also realize we have a very limited number of analyzed cells, and an even smaller, if any, number of repeats to statistically validate the robustness of the measurements. We measure a much larger number of independent (input) variables compared to the number of experiments (output variables) we generate. In principle, with a small ratio of experiments to variables, it is highly unlikely that

we can correctly capture the inherent non-linear structure of the relationship between input and output variables.

Publications have focused on different aspects of gene expression experiments, primarily clustering of cells and genes [2, 3]. Development of novel computational approaches that exploit large warehouses of gene expression data significantly enables the realization of the full potential of this technology. Even though previous approaches were successful at implicitly reducing the number of putative genes with significant signature characteristics, the resulting models still involve a significant number of genes, often hundreds. We will present a framework that identifies the minimum subset of informative genes while imposing maximum possible simplicity of the model describing the data. We summarize the approach in the following section. In this paper, we extend the model to describe the topology of the data generated from the maximally informative genes. We demonstrate how by incorporating topological complexity measures we identify robust and biologically relevant solutions.

2. Integrating optimization and machine learning to select informative genes

Our proposed methodology addresses simultaneously a number of design issues. In a recent publication [4], we presented a model based on classification trees [5] and a thorough review of the basic principles characterizing the feature selection problem and its complexities in the context of machine learning and selection of informative genes. A fundamental assumption is that the minimum set of maximally informative genes is the set of genes producing the least complex decision tree. The decision tree's complexity is determined according to the number of genes used for the classifier and the number of rules comprising the tree. The issue of simplicity in classification trees has been advocated as a way to build robust classifiers [6]. Formulation 1 shows our framework as a large non-linear combinatorial optimization problem. C and C' are vectors containing classifier-derived class assignments and actual class assignments, respectively, of the samples. The former depends on number of features and is implicitly defined via the classifier, denoted $C=T(\lambda_i, i=1...N)$. The "error" term resulting from applying the specific classifier model used, or $\|C-C'\|$, can be defined in a number of different ways. Binary variables model the features selected. The binary variable equals 1 if a particular gene is incorporated in the classifier, and 0 otherwise. The set of informative genes, I_G, is a subset of the original set of genes.

$$
\begin{aligned}
\min \Bigg\{ & \omega_{CA} \underbrace{\|C - C'\|}_{\substack{\text{Classifier} \\ \text{Accuracy} \\ \text{(CA)}}} + \omega_{GCM} \underbrace{\sum_{i=1}^{N_{classes}} \sum_{j \neq i} \frac{\sum_{arc_{i,j} \in MST} arc_{i,j}}{\min_{arc_{i,j} \in MST} \left\{ d_{arc_{i,j}} \right\}}}_{\substack{\text{Geometric Complexity Metric} \\ \text{(CGM)}}} \Bigg\} \\
\text{subject to:} & \\
& C = T(\lambda_i, i = 1, \ldots, N) \\
& CCM = CCM_{target} \qquad (1) \\
& \sum_{i=1}^{N} \lambda_i = N_G \\
& \lambda_i = \begin{cases} 1, g_i \in I_G \subseteq G \\ 0, g_i \notin I_G \subseteq G \end{cases}
\end{aligned}
$$

Minimizing the number of features (genes) in the model. The obvious way to simplify the complexity of our classifier is to minimize the number of degrees of freedom used for building the model. A widely used approach in microarray analysis is feed-forward or backward feature selection process [3]. However, the problem is synergistic effects are not properly captured, and it is often difficult to conclude the actual number of informative features. Therefore, a search algorithm must explicitly account for the actual number of features used. For linear discriminant models, concepts such as Akaike and Bayesian Information Criteria (AIC, BIC) have been used, albeit in a stepwise fashion. In either case, the maximum likelihood estimation

is augmented to account for the number of features (variables) used in the model. We also treat the total number of features used in the model explicitly as one of our complexity criteria.

Minimizing model complexity (Classifier Complexity Metric, CCM). The definition of complexity in a model is not an easy task. When the decision boundary is a hyperplane, it is rather straightforward to require a minimum number of non-zero coefficients (e.g., AIC and BIC). In many other cases outside of a hyperplane, it is not obvious how to decide which model is "simpler" since simplicity is ill defined. We choose to adopt "axis parallel decision trees" as our classifier because they (i) have been shown to be very robust models and (ii) provide a simple way for describing the "complexity" of the classifier by monitoring the number of rules (terminal nodes) in the decision tree. Therefore, we consider explicitly the number of terminal nodes as part of our optimization objectives. A general introduction for decision trees and details of the specific implementation of C4.5 decision tree used in this study are accessible [5]. Our fundamental hypothesis defining the relationship between informative features and the complexity of the classification does not depend on classification algorithm.

3. Incorporating the complexity of the classification problem

So far, we have discussed only the complexity of the classification model. This does not capture the geometric characteristics of the space partitioning achieved by the data's distribution in the reduced space defined by the selected subset of features. We would like to incorporate a "method-independent" metric of complexity characterizing the data's geometric "layout"; in other words, identify features rendering the classification problem simpler. Thus, we have explored the concept of "separating boundaries" [7]. A measure of the complexity of the boundary separating the classes is the "boundary length," the percentage of edges connecting two different classes in the minimum-spanning tree (MST) built from all samples. This measure is an extension of the independence test of samples from two univariate distributions, F_X and F_Y. In this case, standard methods, the Wald-Wolfowitz and Smirnov non-parametric two-sample tests, evaluate the null hypothesis (H_0: $F_X = F_Y$) by sorting the data and collecting statistics on the total number of *runs*. For multivariate case, these tests are extended by using the concept of the MST in order to "sort" the data. In the multivariate case, runs are defined as the sub-graphs of the MST containing points from the same distribution. The concept was extended to characterize inherent separability of two class problems and proved the correlation between complexity metric and ability to build accurate classifiers [7].

For multi-class problems, Equation 2 shows our geometric complexity metric (GCM).

$$GCM = \sum_{i=1}^{N_{classes}} \sum_{j \neq i} \frac{\sum_{arc_{i,j} \in MST} arc_{i,j}}{\min/avg_{arc_{i,j} \in MST} \{d_{arc_{i,j}}\}} \quad (2)$$

According to Equation 2, we determine the MST of the data, based on a particular selection of genes, and for each pair of classes (i,j), we determine the separating boundary (arcs connecting class i and class j), and normalize by the minimum or average arc between the two classes. We appropriately take into account when no arc between two classes exists. Thus, our formulation incorporates one extra level in order to minimize the geometry complexity metric. We desire to minimize complexity either by minimizing MST_{ij}, the MST boundary length or by minimizing GCM. This complexity metric determines how elaborate the expected classifier will be or how challenging it will be to classify the data in the reduced space. The GCM determines the spacing that occurs between multiple classes. By including the GCM, we can characterize various solutions individually and determine the best solution from multiple solutions found. We discovered a direct linear relationship between MST boundary length and the ratio between average intra-class-nearest neighbor (NN) and average inter-class-NN.

4. Computational results

The aforementioned formulation has been successfully applied to the variety of problems in Table 1. SRBCT stands for small round blue cell tumours. Melanoma stands for cutaneous malignant melanoma. GIST denotes gastrointestinal stromal tumor. NCI60 represents 60 samples from the National Cancer Institute's cancer cell lines. Samples in parentheses denote number of testing samples; otherwise, samples denote training samples. Our methodology significantly outperforms any other method in the reduction to a small number of "informative" genes.

Table 1: Data sets used in the analyses

Data Set	Genes/Samples/Classes	I_G
SRBCT [8]	2303/63(20)/4	3
Colon [2]	2000/62/2	3
Melanoma [9]	8067/31/2	1
GIST [10]	1987/18/2	1
Leukemia [3]	2000/38(34)/2	2
Breast Cancer [11]	3226/22/3	2
Prostate [12]	6500/25/2	1
NCI60 [13]	7129/60/9	6

SRBCT Data Set: SRBCT is a category encompassing a large number of malignant tumors that tend to occur in childhood. SRBCTs include neuroblastoma (NB), Burkitt Lymphoma (BL), the Ewing family of tumours (EWS), and rhabdomyosarcoma (RMS). Currently, no single biological or chemical test can detect SRBCTs. The original analysis included a training set of 63 samples and a blind set of 20 samples for testing. A sub-set of 96 most informative genes was identified by performing sensitivity analysis by combining principal component analysis (PCA) and artificial neural networks [8]. Our approach, however, has the ability to explicitly incorporate various complexity metrics leading to a solution capturing the essence of the experimental data. We identified 18 solutions with three genes, the least number of informative genes. More important however is the observation that the vast majority of these 18 solutions contain the same pair of genes: MIC2 and FGFR4, and thus define a conserved pattern of informative genes.

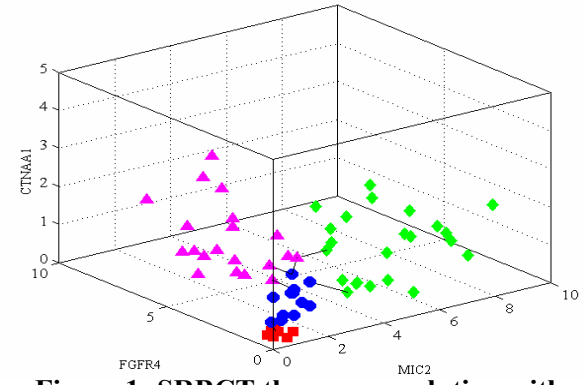

Figure 1: SRBCT three-gene solution with minimum CCM with intra-class connections in MST depicted with lines

Figure 1 shows a solution with RMS (triangles), NB (circles), EWS (squares), and BL (diamonds). ***NCI60 Data Set***: NCI60 contains were nine classes of cancers: central nervous system (CNS), renal (RE), ovarian (OV), leukemia (LE), colon (CO), breast (BR), melanoma (ME), prostate (PR), and non-small-lung (NS). Our analysis found the minimum number of genes was six. A significant portion of six-gene rule sets generated contain a quartet of genes: IL11, GNAI2, CD24, and GLG1. The fifth gene involved has a GenBank accession number of X99393. These five genes are commonly present together in six-gene rule sets.The five genes form a conserved pattern much like MIC2 and FGFR4 did for the SRBCT data set. We found 83 six-gene solutions with these five genes. The decision trees formed by the six-gene combinations are quite similar to one another; hence, we believe further analysis is necessary into this emerging pattern. According to the rule sets we generated, IL11 segregates NS from CO, CD24 distinguishes CN from RE and LE from OV, and GLGL1 discriminates between BR and ME. In many solutions, GNAI2 separates NS from RE and CO from PR.

We have shown our proposed methodology finds (i) a set of maximally informative genes accurately classifying the data and exhibiting minimum classifier and geometric complexity and (ii) multiple solutions, useable as valid biological hypotheses, providing accurate classification and minimum model complexity in terms of CCM and the number of genes used. Our framework found NCI60 and SRBCT data sets have a conserved pattern of five and two genes, respectively, in multiple solutions, which demonstrates the robustness of our solutions. Two key genes found in our solutions to the SRBCT data set: FGFR4 and MIC2, also known as CD99. FGFR4 is a tyrosine kinase inhibitor expressed during myogenesis and prevents terminal differentiation in myocytes [14]. It is currently a therapeutic target due to its high expression in RMS. This is highly consistent with the associated rule identified stating high values of FGFR4 primarily characterize RMS. The second most frequent gene, MIC2, is currently used to diagnose EWS [15]. This is consistent with our decision tree which identified high expression values of MIC2 characterize EWS. For the NCI60 data, IL11, GNAI2, CD24, and GLG1 were four of the five genes we found in a set of solutions forming a six-gene emerging pattern. Previous literature supports our selection of these genes as potentially being involved in cancer development. Breast cancer cells stimulate the production of IL11 [16]. Deletions in the region of the GNAI2 gene have been associated with lung cancer [17]. CD24 is predictive of prostate-specific antigen relapse and was over-expressed in 38.5% of patients with prostate carcinomas [18] and has been linked with ovarian carcinomas [19]. Atypical expression of GLG1 has been associated with pancreatic adencarcinoma [20]. Genes involved in a significant majority of our minimal-gene rule sets have been previously associated with abnormal expression patterns linked to cancer.

6. Conclusions and future work

We have demonstrated that biologically relevant interpretations of large-scale gene expression experiments are plausible if we look at the interpretation problem not from an accuracy point of view but also from a complexity point of view. We formulated the problem of informative gene selection as a large-scale combinatorial optimization problem demonstrated the development of a number of complexity criteria that need to be optimized. We have demonstrated further that novel combinations of machine learning and optimization algorithms provide insightful leads for addressing the pressing problems of modern biology. Using our methodology, we found multiple solutions for use as biological hypothesie and which could be further distinguished by additional analyses. These multiple solutions had a conserved pattern of genes, providing robustness in our methodology and our proposed models. The computational tasks from our novel combination of machine learning and optimization algorithms are daunting. Solution of nonlinear combinatorial optimization problems requires not only the development of novel algorithms but also the support of advanced computer architectures. One main advantage of a mathematical programming formalism, like the one we presented, is the ability to incorporate seamlessly additional biological knowledge as constraints and to integrate diverse sources of biological information.

References

[1] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA microarrays," *Nat Genet*, vol. 21, pp. 33-7, 1999.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc Natl Acad Sci U S A*, vol. 96, pp. 6745-50, 1999.

[3] T. R. Golub, D. K. Slonim, 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, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-7, 1999.
[4] I. P. Androulakis, "Selecting maximally informative genes," *Computers & Chemical Engineering*, vol. 29, pp. 535-546, 2005.
[5] R. Quinlan, *C4.5 Programs for machine learning*: Morgan Kaufmann Publishers, 1993.
[6] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63-91, 1993.
[7] T. Ho, "A data complexity analysis of comparative advantages of decision forest constructirs," *Pattern Analysis Applications*, vol. 5, pp. 102-112, 2002.
[8] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat Med*, vol. 7, pp. 673-9, 2001.
[9] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, "Molecular classification of cutaneous malignant melanoma by gene expression profiling," *Nature*, vol. 406, pp. 536-40, 2000.
[10] S. V. Allander, N. N. Nupponen, M. Ringner, G. Hostetter, G. W. Maher, N. Goldberger, Y. Chen, J. Carpten, A. G. Elkahloun, and P. S. Meltzer, "Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile," *Cancer Res*, vol. 61, pp. 8624-8, 2001.
[11] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent, "Gene-expression profiles in hereditary breast cancer," *N Engl J Med*, vol. 344, pp. 539-48, 2001.
[12] J. Luo, D. J. Duggan, Y. Chen, J. Sauvageot, C. M. Ewing, M. L. Bittner, J. M. Trent, and W. B. Isaacs, "Human prostate cancer and benign prostatic hyperplasia: molecular dissection by gene expression profiling," *Cancer Res*, vol. 61, pp. 4683-8, 2001.
[13] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat Genet*, vol. 24, pp. 227-35, 2000.
[14] B. T. Greer and J. Khan, "Diagnostic classification of cancer using DNA microarrays and artificial intelligence," *Ann N Y Acad Sci*, vol. 1020, pp. 49-66, 2004.
[15] R. C. Shamberger, M. P. LaQuaglia, M. C. Gebhardt, J. R. Neff, N. J. Tarbell, K. C. Marcus, S. L. Sailer, R. B. Womer, J. S. Miser, P. S. Dickman, E. J. Perlman, M. Devidas, S. B. Linda, M. D. Krailo, H. E. Grier, and L. Granowetter, "Ewing sarcoma/primitive neuroectodermal tumor of the chest wall: impact of initial versus delayed resection on tumor margins, survival, and use of radiation therapy," *Ann Surg*, vol. 238, pp. 563-7; discussion 567-8, 2003.
[16] H. Morgan, A. Tumber, and P. A. Hill, "Breast cancer cells induce osteoclast formation by stimulating host IL-11 production and downregulating granulocyte/macrophage colony-stimulating factor," *Int J Cancer*, vol. 109, pp. 653-60, 2004.
[17] J. Roche, F. Boldog, M. Robinson, L. Robinson, M. Varella-Garcia, M. Swanton, B. Waggoner, R. Fishel, W. Franklin, R. Gemmill, and H. Drabkin, "Distinct 3p21.3 deletions in lung cancer and identification of a new human semaphorin," *Oncogene*, vol. 12, pp. 1289-97, 1996.
[18] G. Kristiansen, C. Pilarsky, C. Wissmann, S. Kaiser, T. Bruemmendorf, S. Roepcke, E. Dahl, B. Hinzmann, T. Specht, J. Pervan, C. Stephan, S. Loening, M. Dietel, and A. Rosenthal, "Expression profiling of microdissected matched prostate cancer samples reveals CD166/MEMD and CD24 as new prognostic markers for patient survival," *J Pathol*, vol. 205, pp. 359-76, 2005.
[19] A. D. Santin, F. Zhan, S. Bellone, M. Palmieri, S. Cane, E. Bignotti, S. Anfossi, M. Gokden, D. Dunn, J. J. Roman, T. J. O'Brien, E. Tian, M. J. Cannon, J. Shaughnessy, Jr., and S. Pecorelli, "Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: identification of candidate molecular markers for ovarian cancer diagnosis and therapy," *Int J Cancer*, vol. 112, pp. 14-25, 2004.
[20] T. Crnogorac-Jurcevic, E. Efthimiou, P. Capelli, E. Blaveri, A. Baron, B. Terris, M. Jones, K. Tyson, C. Bassi, A. Scarpa, and N. R. Lemoine, "Gene expression profiles of pancreatic cancer and stromal desmoplasia," *Oncogene*, vol. 20, pp. 7437-46, 2001.