

Informative gene selection and design of regulatory networks using integer optimization

E. Yang^a, T. Maguire^a, M.L. Yarmush^a, I.P. Androulakis^{a,b,*}

^a Biomedical Engineering Department, Rutgers University, 617 Bowser Road, Piscataway, NJ 08854, United States

^b Chemical & Biochemical Engineering Department, Rutgers University, Piscataway, NJ 08854, United States

Received 16 June 2006; received in revised form 21 January 2007; accepted 22 January 2007

Available online 12 February 2007

Abstract

A central problem in bioinformatics and systems biology is the selection of appropriate models in a rational and systematic way. This fundamentally combinatorial problem can be readily formulated and addressed within an integer optimization framework. In this paper we examine two such applications related to the identification of informative genes and the quantification of regulatory networks. We demonstrate how multiple alternatives can be systematically derived and assess the information content of the proposed solutions.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Gene expression; Regulatory networks; Gene selection; Classification; Integer optimization

1. Introduction

The recent advances in high throughput gene expression analysis have sparked an ongoing revolution in modern biology (Bowtell, 1999; Cheung et al., 1999; Kafatos, 2002; Lipshutz et al., 1999; Quandt et al., 1995; Schena et al., 1995). The capability of recording the coordinated response of a large amount of interacting genes offers the tantalizing possibility of unraveling, and quantifying, the elements guiding the temporal dynamic response of an organism. The implications can, potentially, be profound as the identification of the structure and the quantification of the interactions underlying the biological mechanism which controls the expression dynamic in response to external signals will provide important clues towards the understanding of the process of activation and deactivation of key regulators.

Extracting useful information from large scale genomic studies remains a daunting task. The combinatorial nature of the selection process (which components and what structure) makes simple enumeration techniques irrelevant. A multitude

of approaches are motivated by the realization that model reduction renders interpretation meaningful avoiding point-less over fitting of experimental data. In genomic studies where the main goal is the generation of alternative testable hypotheses, having the ability to generate multiple realizations is extremely important. Two critical issues arise that require further investigation: (i) how to effectively generate multiple simplified model realizations; (ii) how to control the complexity of those models. The inherent nature of the combinatorial character of these problems and the need for the systematic determination of alternatives, makes mathematical programming, and in particular integer optimization, a useful approach.

In this paper we will discuss potential applications of integer optimization in two important problems in bioinformatics and systems biology. Namely, the selection of informative genes and the construction of transcriptional regulatory networks. We will demonstrate how these tasks can be effectively formulated as mathematical programming models and how reformulations and linearization techniques have to be invoked to improve the solution efficiency. Specifically, we will test the hypothesis that model building can be formulated as mixed-integer (non)-linear optimization problems by analyzing two case studies: (i) the selection of informative genes in a cancer case study, and (ii) the quantification of transcriptional regulatory networks in a rat burn injury animal model.

* Corresponding author at: Biomedical Engineering Department, Rutgers University, 617 Bowser Road, Piscataway, NJ 08854, United States.
Tel.: +1 732 445 0099.

E-mail address: yannis@rci.rutgers.edu (I.P. Androulakis).

2. Informative gene selection

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 (Bassett et al., 1999). A number of excellent publications have focused on different aspects of the analysis of gene expression experiments (Alizadeh et al., 2000; Allander et al., 2001; Alon et al., 1999; Bittner et al., 2000; Chilingaryan et al., 2002; Dettling & Buhlmann, 2004; Dudoit et al., 2002; Golub et al., 1999; Khan et al., 2001; Luo et al., 2001; Perou et al., 1999; Pollack et al., 1999; Ross et al., 2000; Szabo et al., 2002). The main focus of these studies is to derive a single interpretation of the data by selecting a reduced set of genes that exhibit coherent expression patterns. Further analyses of the computational results assign a certain level of significance to smaller subsets of genes whose expression patterns could potentially indicate a more direct involvement in the biological process under study.

Machine learning algorithms, like the ones usually employed for gene selection, are known to be prone to deteriorating performance when faced with many irrelevant or correlated features (Kohavi & John, 1997). A universal, therefore, problem is to decide on which aspects, i.e., features, of a problem are relevant. Narendra and Fukunaga (1977) were among the first to present a formal approach based on a branch and bound scheme for addressing the very same problem. A recent review by Kohavi and John (1997) examines a number of issues associated with the problem of feature selection. More recently, Liu and Motoda (2000) also present ideas related to the coupling of information theory and feature selection. A fundamental problem in machine learning is the development of accurate classifiers in sparsely populated datasets, i.e., *almost empty spaces* (Duin, 2000). A key complexity of microarray experiments is the essential lack of observables (cell lines or tissue samples) to support the large number of probes monitored. The consequences of the small ratio of features to samples were extensively discussed in Jain and Zongker (1997). The inability of sparse data to properly capture the complexity of a classification problem was also analyzed by Ho (2002). A nice discussion of the impact of the small sample size problem in array expression data is presented in Dougherty (2001). The implications of the ratio of features to samples is critical as sparsely populated datasets can very easily lead to random features appearing to be informative (i.e., able to classify data) when in reality no structure exists in the data whatsoever. It should be expected that simple minimization of the number of features (genes) in a model need not necessarily provide the best possible answer. Therefore, additional complexity restrictions will have to be proposed to balance the lack of available data although no definite answer can be provided as no analysis can replace accurate and adequate data.

We recently examined the interplay between accuracy and complexity in the context of extracting informative genes (Androulakis, 2005) under the assumption that simplicity of model representation should be invoked when dealing with sparse data. However, modeling the complexity of a model is not a trivial task, with the exception of the straightforward require-

ment that the least number of features are used in a model. However, the inherent model complexity, irrespective of number of degrees of freedom, is a more intricate concept. Mathematical programming approaches offer the advantage that a number of structural constraints, such as model complexity, can be explicitly incorporated. In the context of classification the idea of transforming the problem to a non-linear optimization formulation is a very well established one. The pioneering work of Mangasarian (1965, 1968) demonstrated how to formulate the problem of constructing planes to separate linearly separable sets of points. In addition, early work by Freed and Glover (1981a, 1981b, 1986), Gehrlein (1986), Glover et al. (1988) and Glover (1990) skillfully discussed various aspects of discriminant analysis from the point of view optimization. A more recent excellent review was presented in Stam (1997) highlighting numerous developments that defined the field of applications of mathematical programming to statistical classification. The more recent work of Shioda (2003) and Uney and Turkay (in press) identified opportunities for successful reformulations of various data mining tasks in the context of integer optimization while Busygin et al. (2005) present some more recent ideas for addressing the bi-clustering problem as a fractional 0–1 optimization problems. One the main challenges, however, remains the issue of multiple classes. Multiclass feature selection was recently formulated as a mixed-integer linear problem in Iannarilli and Rubin (2003) whereas Sun and Xiong (2002) discussed a linear formulation for the selection of informative genes. However, none of these approaches consider the complexity of the classifier nor do they address its control.

An interesting idea was recently introduced (Street, 2005) in the context of oblique multicategory classification trees, whereby the class assignment is modeled through the use of the concept of purity of a partition. Ideally, one wishes to construct a multivariate classifier in such a way that each “partition” is occupied by elements of a single class (orthogonal partitions). However, this formulation is faced with a number of complexities. First, it is highly non-linear, second it does not perform feature selection and finally it builds classifiers sequentially, in the sense of the one-against-all concept. However, the purity concept introduces a very intelligent way of quantifying the ability of a classifier to partition the data. Furthermore, with proper modifications to be discussed shortly, it allows the quantification of the complexity of the classifier.

We will discuss the basic elements of a proposed approach which can be effectively generalized in the context of a mixed-integer optimization to develop a general framework of oblique multi-category trees to address the question of how to build simple, yet informative, classifiers that simultaneously perform informative feature selection.

2.1. Mixed-integer reformulation of the oblique-multi-category feature selection problem

We assume that we are given the ensemble of gene expression data in the form of n f -dimensional vectors belonging to k distinct classes. The question is to identify how many, and which, of these f features are critical for the construction of a simplified,

yet informative, classifier. An oblique multi-category classifier is defined by the intersection of a number of planes. We term these intersections “partitions”, $\pi = 2^p$, where p is the number of planes. We define complexity as the number of occupied partitions that are required to properly classify the data. $q_{n,\pi}$ is a binary variable indicating whether point “ n ” belongs in partitions π . The total number of points of class k in partition π is termed $\sigma_{k,\pi}$ whereas $v_{k,\pi}$ denotes the fraction of points of class k in said partition. Finally, $y_{k,\pi}$ is a binary variable which is 1 if that partition contains even a single point from class k , and 0 otherwise. The location of a point relative a particular plane is defined according to the binary variable $z_{n,p}$ which is 1 if the point is below the plane, and 0 otherwise.

This variable is a critical one since it basically defines all the auxiliary variables in the formulation. Finally, s_f is a binary variable indicating whether feature “ f ” is used in the construction of the classifier. Given that p planes created 2^p partitions we wish to identify the partitions, and the corresponding spatial distribution of points in the reduced space, defined by s_f , which will create the “purest” possible partitions. We model this by analyzing the product $y_{k,\pi}v_{k',\pi}$, $k \neq k'$. In order to account for non-linearly classifiable problems we are basically looking for partitions that satisfy:

$$y_{k,\pi} \times v_{k',\pi} \leq E \Rightarrow \begin{cases} y_{k,\pi} = 0 \wedge v_{k',\pi} \leq 1 \\ y_{k,\pi} = 1 \wedge v_{k',\pi} \leq E \\ y_{k,\pi} = 0 \wedge v_{k',\pi} \leq E \end{cases}$$

or

$$y_{k',\pi} \times v_{k,\pi} \leq E \Rightarrow \begin{cases} y_{k',\pi} = 0 \wedge v_{k,\pi} \leq 1 \\ y_{k',\pi} = 1 \wedge v_{k,\pi} \leq E \\ y_{k',\pi} = 0 \wedge v_{k,\pi} \leq E \end{cases}$$

The modeling idea is that (i) the partition contains no point of class “ k ” and the maximum numbers of “ k' ” (empty partition), (ii) the partition contains points of class “ k ” and contains the minimum number of points of class “ k' ”, and (iii) if no points of class “ k ” are present, then it may contain an arbitrary number of points of class “ k' ”. Obviously, the point is to maximize the number of type (i) partitions while satisfying the “purity” requirements.

The objective thus becomes to minimize the “slack” variable E . The detailed formulation is summarized in Fig. 1. The novelty of our approach is that it allows the complete control of the complexity of the model by treating explicitly the number of features, and number of occupied partitions. Specifically, the formulation optimizes simultaneously the following criteria:

- (i) the feature selection,
- (ii) the construction of a multivariate, multi-class classifiers,
- (iii) the creation of multiple structurally alternative solutions via the introduction of integer cuts (Biegler et al., 1997).

The overall framework remains linear and the solution is done parametrically for a given number of occupied partitions and number of features. This is a simple way for decoupling the objectives, however, more elaborate multi-objective schemes

can, and will be, explored. We will discuss a number of computational studies to illustrate the method.

2.1.1. Motivating example

This is a simple motivating example to illustrate the effect of constructing classifiers of increasing complexity. Since in our formulation the numbers of planes as well as the number of “occupied” partitions are both parameters we can construct classifiers of arbitrary complexity. This motivating example is two variable problem with two classes symmetrically partitions in 4 domains. The optimal planes for various numbers of partitions and planes are depicted in Fig. 2. The main message we wish to convey with this trivial example is the possibility of creating models of arbitrary complexity and the benefits of having the ability to explicitly control the structure of the model.

2.1.2. Iris problem

This is a benchmark classification problem dating back to 1936 (Fisher, 1936). Its originator, R.A. Fisher, developed the problem to test clustering analysis and other types of classification programs prior to the development of computerized decision tree generation programs. The dataset is small consisting of 150 records and the raw data readily available from the UCI repository (<http://kdd.ics.uci.edu/>). The target variable is categorical specifying the species 3 of iris, namely virginica, versicolor and setosa. This test case is used to illustrate the concept of multiple “cuts”, i.e., structurally distinct models. The four characteristics of the iris flower are treated as the features of the problem. The additions of the cuts for the identification of subsequent classifiers results in a significant deterioration of the model. The important thing to realize is that the “most informative” variables (features 3 and 4) results in the best possible classification, i.e., min error. Once again, this motivating example is used to illustrate the power of the ability to rationally generate multiple realizations of the classification model (Fig. 3).

2.1.3. Small round blue cell tumor case study

“Small Round Blue Cell Tumors” (SRBCT) is a descriptive category encompassing a large number of malignant tumors that tend to occur in childhood. They are united by their similar histopathological appearance. However, subtle clues may be present to distinguish between the tumors. For proper characterization pathologists often employ immunohistochemistry, electron microscopy, and molecular analysis for chromosomal abnormalities. The SRBCTs of childhood include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and the Ewing family of tumors (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. The data were reduced by SVD decomposition and the leading factors were used to train an Artificial Neural Network to build a predictive diagnostic device. This study constitutes a milestone since it was the first attempt to use microarray experiments in a predictive to explore the potential application of using such methods for tumor diagnosis. Computationally it is a very interesting problem. It is of relatively modest size, containing the expression levels of

$\begin{aligned} & \min E \\ & \text{st.} \\ & v_{k',\pi} \leq E + (1 - y_{k,\pi}) \\ & \quad \text{or} \\ & v_{k,\pi} \leq E + (1 - y_{k',\pi}) \\ & \sum_{k,\pi} y_{k,\pi} = N_{\pi} \\ & q_{n,\pi} \leq \frac{\sum c_1(p,\pi)z_{n,p} + c_0(\pi)}{N_p} \\ & q_{n,\pi} \geq \frac{\sum c_1(p,\pi)z_{n,p} + c_0(\pi) - (N_p - 1)}{p} \\ & \sum_n \sum_{\pi} q_{n,\pi} = N \\ & v_{k,\pi} = \frac{\sigma_{k,\pi}}{N_k} \\ & \sigma_{k,\pi} = \sum_n B(n,k)q_{n,\pi} \\ & \sum_{\pi} v_{k,\pi} = 1 \\ & \sum_f D(n,f)w_{f,p} + Uz_{n,p} \leq U + \vartheta_{\bar{p}}z_{n,p}\varepsilon \\ & \sum_f D(n,f)w_{f,p} + (u - \varepsilon)z_{n,p} \geq \varepsilon + \vartheta_p \\ & y_{k,\pi} \leq N_k v_{k,\pi} \\ & y_{k,\pi} \geq v_{k,\pi} \\ & w_{f,p} \geq \mu s_f \\ & w_{f,p} \leq M s_f \\ & \sum_f s_f = N_f \end{aligned}$	$z_{n,p} = \begin{cases} 1 & \text{if } \sum_f D(n,f)w_{f,p} \leq \vartheta_p \\ 0 & \text{if } \sum_f D(n,f)w_{f,p} \geq \vartheta_p \end{cases}$ $q_{n,\pi} = \begin{cases} 1 & \text{if point } n \text{ belongs to partition } \pi \\ 0 & \text{otherwise} \end{cases}$ $\sigma_{k,\pi} = \text{number of points of class } k \text{ in partition } \pi$ $v_{k,\pi} = \text{fraction of points of class } k \text{ in partition } \pi$ $y_{k,\pi} = \begin{cases} 1 & \text{if partition } \pi \text{ contains points of class } k \\ 0 & \text{otherwise} \end{cases}$ $s_f = \begin{cases} 1 & \text{if feature } f \text{ is used} \\ 0 & \text{otherwise} \end{cases}$ $y_{k,\pi} v_{k',\pi} \leq E \Leftrightarrow \begin{cases} y_{k,\pi} = 0 \wedge v_{k',\pi} \leq 1 \\ y_{k,\pi} = 1 \wedge v_{k',\pi} \leq E \\ y_{k,\pi} = 0 \wedge v_{k',\pi} \leq E \end{cases} \Leftrightarrow v_{k',\pi} \leq E + (1 - y_{k,\pi})$ <p>or</p> $y_{k',\pi} v_{k,\pi} \leq E \Leftrightarrow \begin{cases} y_{k',\pi} = 0 \wedge v_{k,\pi} \leq 1 \\ y_{k',\pi} = 1 \wedge v_{k,\pi} \leq E \\ y_{k',\pi} = 0 \wedge v_{k,\pi} \leq E \end{cases} \Leftrightarrow v_{k,\pi} \leq E + (1 - y_{k',\pi})$ <p>$w_{f,p}$ and ϑ_p = plane parameters, $Dw \leq \vartheta$ (variables)</p> <p>N = number of samples N_{π} = desired number of occupied partitions (parameter) N_p = number of planes (parameter) N_f = number of features to be selected (parameter) N_k = number of samples in class k (known) u, U, μ, M = big-M parameters c_0, c_1 = model parameters (known) $D(n, f), B(n, k)$ = data p = planes π = partitions = 2^p f = features k = classes n = samples</p>
---	---

Fig. 1. Mixed-integer reformulation for the design of oblique multicategory decision trees.

2303 genes, with 63 cells, belonging to 4 cancer types, used for training purposes. Khan et al. (2001) construct their ANN by using as inputs the projection of the expression measurements onto the first 10 principal directions are determined by a Principal Component Analysis of the raw expression data identified a sub-set containing 96 of most informative genes by performing an exhaustive sensitivity analysis. The genes are ranked based on their ability to discriminate the four classes. The raw data are pre-processed used an extension of the signal-to-ratio approach introduced by Golub et al. (1999). The original method was extended for multiclass-class problems in order to assist in the elimination of irrelevant features. This step reduced the initial number of features to 500. Multiple cuts were generated for a multitude of features/plane combinations and we discuss representative results to illustrate the extracted information. The maximally informative model has 3 features and 2 planes (4 occupied partitions). As a standard validation, we did verify that scrambling the data (systematic error) does significantly

deteriorate the performance of the classifier which is a further validation that the underlying structure in the data in the result of a random process (Fig. 4).

The main goal of our methodology, beyond simply building a classifier, is to test the hypothesis that informative features (genes) should exhibit robustness with respect to the derived models. In other words we wish to identify potentially conserved subsets of genes that are statistically overrepresented in the multiple solutions we have generated. Initial evidence with axis-parallel decision trees (Androulakis, 2005) pointed indeed to this direction. Through the aforementioned analysis we have identified several conserved key genes across multiple solutions, all of which are integral in tumor progression. The first of these genes, caveolin-1 (CAV-1), has been documented, when its expression patterns are altered, to be a key component in the formation of a variety of tumors, such as prostate (Williams et al., 2005), bladder (Rajjayabun et al., 2001), esophageal, and mammary (Lee, Park, et al., 2002; Williams et al., 2004).

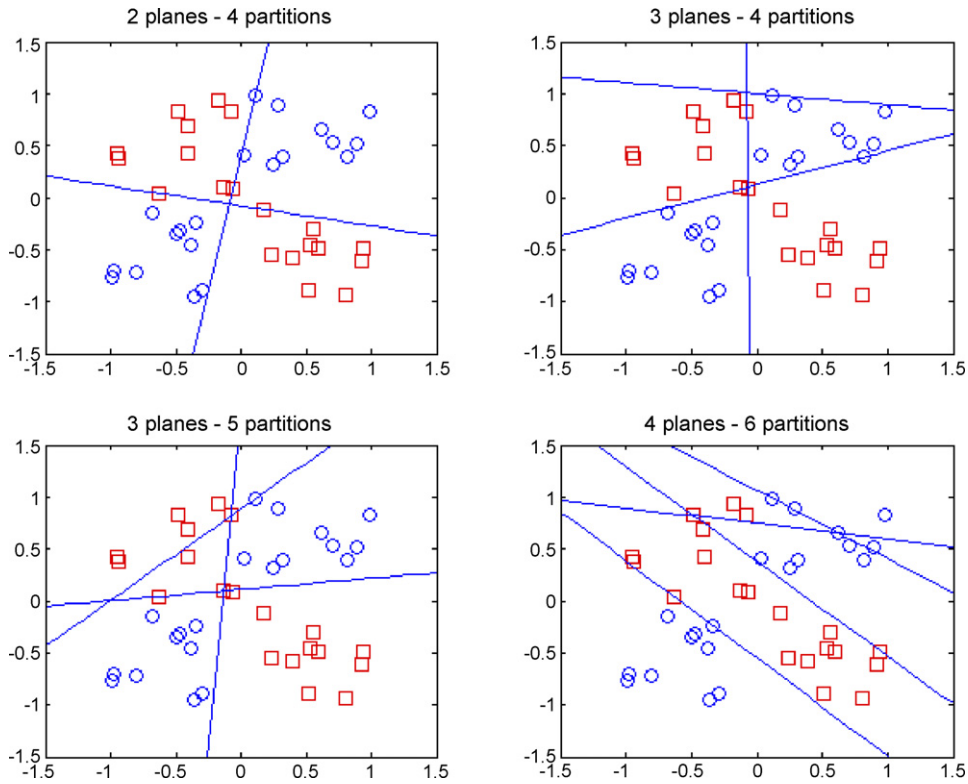


Fig. 2. Partitions of increased complexity.

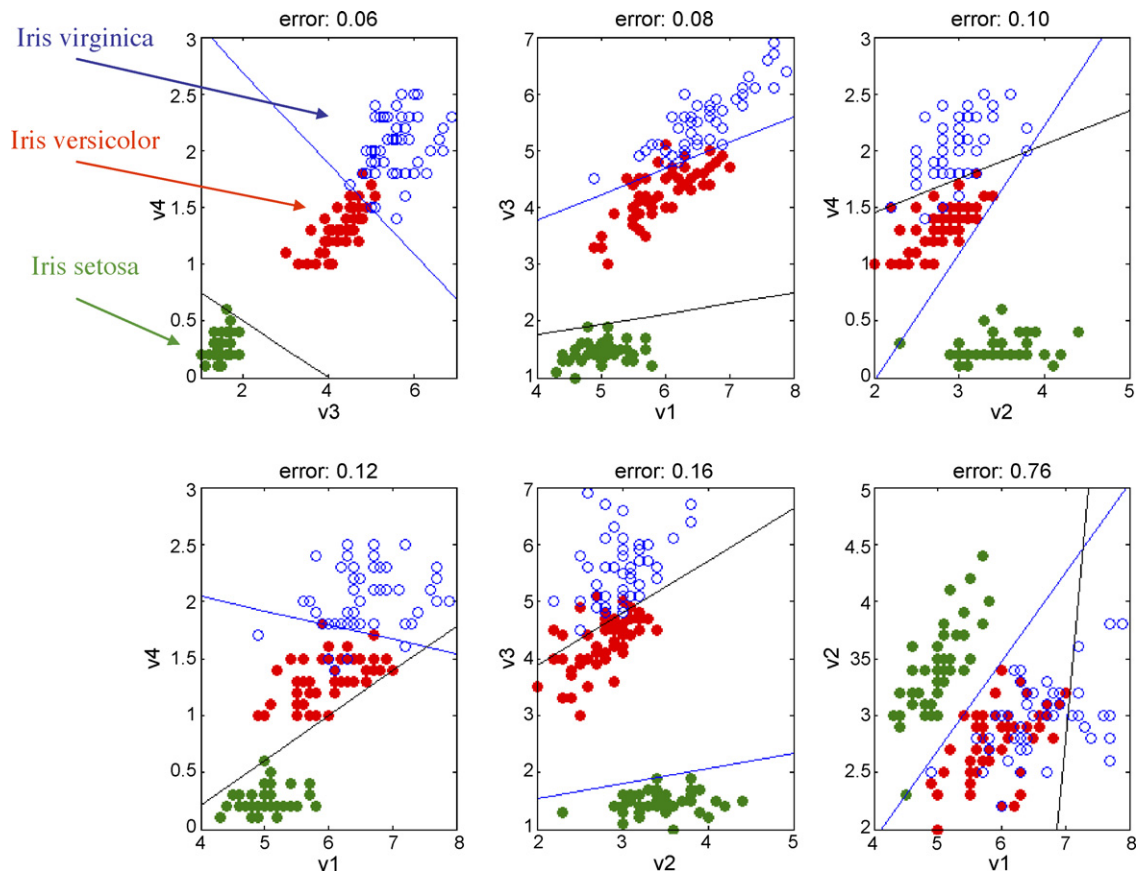


Fig. 3. Multiple solutions to the iris problem.

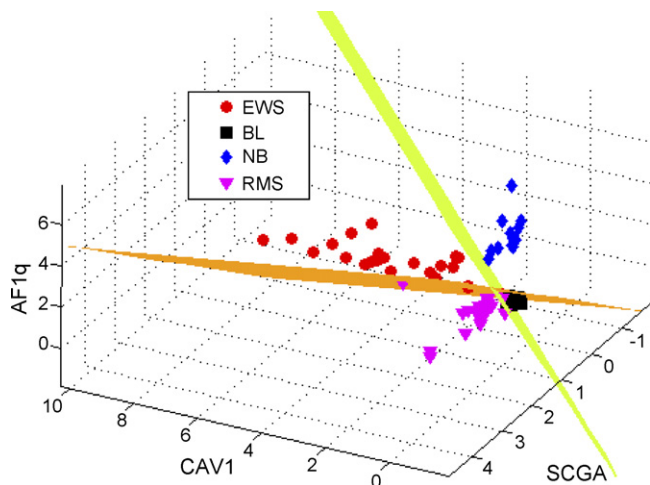


Fig. 4. Two plane, 4 partitions, 3 gene solution to the SRBCT problem.

The effects of CAV-1 in tumorigenesis fall into three major categories: (1) deregulation of cell cycle control (Williams et al., 2003); (2) metalloproteinase production (Williams et al., 2004); (3) induction of angiogenesis (Sonveaux et al., 2004). Tumorigenesis caused by lack of cell cycle control is established through a process involving the hyperactivation of cellular proliferation mediated through cyclin D1. D-cyclins are involved in controlling cell cycle progression by activating their associated kinases cdk4 and cdk6. These cyclin-dependent kinases phosphorylate the retinoblastoma pRB protein, leading to transition through the G1 phase of the cell cycle. In cases where altered expression profiles exist for CAV-1, the aforementioned signaling system is modified, leading to a lack of cell cycle control, and subsequent rapid cellular growth. The second component of CAV-1's involvement in tumorigenesis, is the effect of CAV-1 on metalloproteinase production, and the subsequent production of extracellular matrix production. Specifically, when CAV-1 exhibits aberrant control of metalloproteinases, there is an increase in matrix protein production, which allows for increased tumor migration, invasion, and metastasis. The final effect that CAV-1 may have on tumor formation is its effect on angiogenesis. Specifically, CAV-1 controls the activity of nitric oxide synthase, and vascular endothelial growth factor, two key mediators of angiogenesis. Angiogenesis itself is necessary for the growth of tumors, being that the large cell masses comprising metastatic tumors need high levels of metabolite and oxygen transport in order to live-levels only acquired in highly vascularized environments. Thus, in cases where loss of control of angiogenesis mediators is experienced, such as when CAV-1 is altered, angiogenesis occurs, supporting tumor growth. The second gene identified, neurofibromatosis 2 (NF2), has also been shown to be involved in cancer formation, in neural based tumors such as schwannomas and meningiomas (Fraenzer et al., 2003; Lomas et al., 2005; Ryu et al., 2005; Xiao et al., 2003). NF2 exerts its effect through its gene product, merlin, which is involved in the regulation of cell motility and cell proliferation. Recent studies have highlighted the involvement of merlin in tumor suppression through the inhibition of Rac signaling. In cases where the production of NF2 is

diminished, RAC signaling becomes activated, and the tumor suppression capabilities of NF2 are lost. The third major gene identified is a myeloid/lymphoid or mixed-lineage leukemia marker (AF1Q). While the literature on this gene is limited, it is known that this gene is necessary for neuronal differentiation (Lin et al., 2004). Thus, it may be possible that uncontrolled regulation of this gene may lead to neuronal-based tumors. The fourth gene, sarcoglycan, alpha (SGCA), is a component of the dystrophin–glycoprotein complex, and has been linked to the onset of mammary tumorigenesis (Weir & Muschler, 2003). Tumorigenesis is induced when mutations inhibit the production of SGCA. This leads to a subsequent loss of control over a variety of functions such as growth control, cell survival, cytoskeletal organization, basement membrane assembly, branching morphogenesis, polarity, and tumor suppression in epithelial cells. As mentioned previously in the case of CAV-1, when these key components of cellular function become aberrant, tumorigenesis ensues. The final gene identified through our analysis is the CD99 antigen (CD99), which has been determined to be a marker of lung carcinomas as well as mammary tumors. It is thought that CD99 might play an integral role in the aggregation of breast cancer cells, the initiating step of tumorigenesis. CD99 also assists in the invasive processes characteristic of metastatic tumors. Finally, the gene for receptor, IgG, alpha chain transporter (FCGRT, FCRN). This gene, as well as a few others, has been detected through the use of cDNA microarrays, in studies involved in elucidating the underlying genomic profile of astrocytomas (Huang et al., 2005). Specifically FCRN is known to mediate immune defence in response to the onset of pilocytic astrocytomas, possibly keeping the astrocytoma in a benign state. In addition, FCRN is known to be expressed by dendritic cells (Zhu et al., 2001) and may serve as a basal mechanism of immune function.

Taken together, these genes may work in a concerted effort to induce tumorigenesis, initialized through the signaling capabilities of CD99. CD99 has been shown to initiate cytoskeletal reorganization (Cerisano et al., 2004), a dominant cellular process which is also influenced by NF2, as well as SCGA. However, further studies will be needed to elucidate the interactions of these proteins. Finally, it is very important to realize that all the most frequent genes have already been directly implicated with the various types of SRBCT type of cancers in the literature (Baer et al., 2004; Khan et al., 2001).

2.2. Discussion

Several issues remain to be addressed within this framework. By far the most critical is to augment the design objective with an additional component to model maximum separability of classes, along the support vector machine principles (Guyon et al., 2002). Furthermore, the systematic analysis of the multiple criteria needs to be addressed in an integrated framework as well as alternative decomposition methods for the solution of the linear, albeit significant in size, integer optimization problem. However, the current results have clearly supported the hypothesis that simplicity does improve the information content, through the selection of reduced sets of informative features, as

well as the realization that indeed conserved elements in the feature space can be identified and provide excellent leads for further investigation.

3. Design of regulatory networks

A number of approaches have been proposed in recent years to decipher the hidden complexities of modeling regulatory networks. Statistical methods, such as principal component analysis (Raychaudhuri et al., 2000), independent component analysis (Liebermeister, 2002), or singular value decomposition (Holter et al., 2000) have been applied successfully in order to extract meaningful information from large scale genomic studies by developing low(er)-dimensional representation of the original expression data. Furthermore, other techniques take a more “comprehensive” approach and attempt to build dynamic models by appropriately fitting linear models and selecting among the possible alternatives the ones that generate the best approximation of the available data (Dasika et al., 2004).

Recently a new method, Network Component Analysis (NCA), was proposed Liao et al. (2003) and Kao et al. (2004). NCA is driven conceptually by two key realizations. First, a critical concept behind the proposed equivalence between co-expression and co-regulation is that a limited number of key transcription factors control the observed dynamic of an ensemble of genes that exhibit similar expression temporal dynamics. While even though the number of observables (genes) might be large, the actual number of controlling degrees of freedom (transcription factors) should be significantly smaller. The second conceptual observation is based on the realization that relevant information regarding the nature of transcription factors already exists (and more is accumulated). This has rarely been taken into account when regulatory networks are being quantified. In other works, instead of looking at the entire space of all possible connections, such as through analysis and identification of key time-lagged correlations as proposed by Schmitt et al. (2004) one should limit the search on the space of biologically plausible interactions. The latter is beginning to be assembled either as a result of extensive experimentation or computations (Wei et al., 2004).

The key innovation provided by NCA was the realization that transcription regulatory networks are not in fact a black box system in the traditional sense of the word. This is because outside information about the structure of the regulatory network exists, namely the connectivity of the transcription factors in relation to the genes being investigated. Therefore, instead of making mathematical assumptions in order to make the decomposition feasible, what NCA does is make an assumption with a basis in biology in order to facilitate decomposition. NCA has already been used to decipher regulatory architectures in systems such as *E. coli* (Liao et al., 2003).

NCA is based on the fundamental premise that a unique decomposition, once a set of transcription factors has been identified, can be derived provided that a set of very specific mathematical conditions are satisfied. One of the limitations however, is that in its present form the methodology cannot test the validity of the assumptions a priori and therefore these

conditions have to be verified after the model has been built. In addition, these decompositions, and hence the regulatory structures, are not structurally unique (Boscolo et al., 2004). This potential multiplicity (redundancy) is a key concept often invoked to justify the apparent robust of biological systems (Kitano, 2004). However, generating the alternative structures in a systematic manner is also an open question. The latter becomes quite critical as it has already been hypothesized that biological systems can be potentially be characterized by a very high degree of interconnectivity between the different genes (Barabasi & Oltvai, 2004; Herrgard et al., 2004; Qian et al., 2003). Hence systematic identification and evaluation of alternative structures becomes an important issue.

The gene expression measurements are assumed to have been assembled in the form of a matrix $[E]$ of size N (rows) \times M (columns) where the expression levels (ratio) of N transcripts are recorded at M time points. Assuming that the transcriptional response is controlled by L regulatory signals, then we wish to reconstruct a model describing the expression levels as follows:

$$[E] = [A][P] \quad (1)$$

Such that the matrix $[P]$ (size: $L \times M$) consists of samples of the L regulatory signals at each time point. A meaningful reduction is achieved provided that $L < N$, that is the number of controlling regulators is less than the number of expressed genes. Additionally the number of transcription factors must also be less than the number of time points. The matrix $[A]$ (size: $N \times L$) quantifies the connectivity strength of an active interaction between an expressed genes and a regulator. Eq. (1) is in essence to be interpreted as a function expansion of the original signal $[E]$ along the set of basis functions defined by $[P]$. The essential complexity of the method however results from the fact that the proposed decomposition of $[E]$ into matrices $[A]$ and $[P]$ is ill-defined and in general non-unique. However, Liao et al. (2003) and Kao et al. (2004) demonstrate that the introduction of additional assumptions then the solutions of (1) are such that

$$\begin{aligned} [\bar{A}] &= [A][X] \\ [\bar{P}] &= [X^{-1}][P] \end{aligned} \quad (2)$$

However, the formulation of NCA restricts the solution space, where all of the solutions derived from the same initial connectivity matrix with the same error differ by a diagonal scaling matrix. In order to achieve this, the following criteria are imposed on the structure of the decomposition matrices $[A]$ and $[P]$:

1. The connectivity matrix $[A]$ must be full rank.
2. When a node in the regulatory layer is removed along with all the output nodes connected to it, the resulting networks must be characterized by a connectivity matrix that is still full-column rank. This condition implies that each column of $[A]$ must have at least $L-1$ zeros.
3. $[P]$ must be full row rank.

When an identifiable network, in terms of $[A]$, has been identified then the solution to the following estimation problem yields

an acceptable solution

$$\begin{aligned} \min \quad & \| [E] - [A][P] \|^2 \\ \text{s.t.} \quad & A \in Z_0 \end{aligned} \quad (3)$$

where Z_0 is a topology induced by the network connectivity pattern, i.e., forcing appropriate elements of the $[A]$ matrix to zero. In addition to that, if a particular entry of $[A]$ can be biologically justified, that is there is no documented interaction between a regulator and a gene, also the corresponding entry is forced to zero. NCA Criterion III cannot, in general, be tested a priori. Therefore, the solution may have to be further analyzed to verify that indeed it satisfied all the identifiable properties. The constraints upon $[A]$ and $[P]$ are common to all of the component analysis techniques, which strive to decompose the original observational matrix into a two linearly independent matrices. This is done to satisfy the mathematical notion that the solution space is not being fitted with more dimensions than is necessary. The second criterion in NCA effectively implies that, “No transcription factor can be regulating only a subset of another transcription factor.” This forces the solution space to be a set of solutions which differ only by a diagonal scaling matrix. What this implies is that each column must have at least $L-1$ zeros. However, one must be careful to note, that the solutions will only differ by a diagonal scaling matrix if and only if the residue is identical.

One the major complications however stems from the fact that given an initial binary connectivity matrix obtained through transcription factor analysis (Wasserman & Sandelin, 2004), it is very rare that the initial A matrix will be NCA compliant. Therefore, manipulations must be done in order to allow NCA to operate on this initial matrix. The approach taken by Liao et al. (2003) and Kao et al. (2004), is to go through the A matrix and remove any transcription factors and the genes that break the NCA criteria of full column rank, and the full rank of the reduced form. The result of this is the network can be decomposed into many independent cliques. Furthermore, most biological systems are densely connected, and breaking the problem up into independent cliques removes a lot of the inherent richness in the response of an organism to a simple input.

3.1. Mixed-integer network component analysis (miNCA) reformulation

We are proposing an extension to NCA, mixed-integer network component analysis which attempts to address a number of issues. First, as mentioned earlier, NCA proceeds in two, possibly three steps: identification of Z_0 , solution of the estimation problem, validation of rank requirements on $[P]$. Furthermore, for dense interactions it is expected that potentially multiple structures could be identified. Therefore, the question arises whether an integrated framework can be developed that would, systematically, address both these issues. One of the main goals of this analysis is to determine whether by pruning an initially dense connectivity matrix which we obtain through transcription factor analysis, and determine if NCA in an optimization framework will provide information allowing for the selection

of highly conserved connections. What we are trying to find essentially is the connectivity structure that yields the best fit to the data. It is our belief that these highly conserved connections will, perhaps, represent the primary response pathways of an organism to various stimuli.

In order to make this evaluation, we must determine the optimality of the generated connectivity subsets in terms of their ability to reproduce the given expression profile. Since this makes assumptions upon the $[A]$ matrix, it is important that we apply the NCA formulation instead of the other component analysis algorithms discussed previously. Given the large combinatorial space of the solutions space, it would be infeasible to enumerate all of the possibilities for the pruned subset. Therefore, in order to solve the problem in a reasonable amount of time, we turn to an optimization approach, namely mixed-integer non-linear programming. That mixed-integer non-linear programming theoretically allows is the evaluation of which subset of connections yields the best fit as evaluated by the closeness of fit compared to the original measured response without having to evaluate every possible subset of the initial connectivity matrix. However, in order to do this, the constraints given by NCA must be appropriately modeled in a closed form.

The reconstruction error at each time point is modeled as the difference of between positive slack variables.

$$\begin{aligned} \min \quad & \sum_i \sum_t eP(i, t) + eN(i, t) \\ \text{s.t.} \quad & eP(i, t) \geq 0, \quad \forall i, t \\ & eN(i, t) \geq 0, \quad \forall i, t \\ & E(i, t) - \sum_j A(i, j)P(j, t) = eP(i, t) - eN(i, t), \quad \forall i, t \end{aligned} \quad (4)$$

$E(i, t)$ is the expression profile log normalized to the expression at time 0. The set “ i ” denotes the genes, $i=1, \dots, M$; “ j ” denotes the set of transcription factors, $j=1, \dots, L$; the set “ t ” denotes the number of time points. $eP(i, t)$ is the positive slack variable and $eN(i, t)$ is the negative slack variable.

The existence of a particular regulatory interaction is modeled by using a binary variable $y(i, j)$ such that

$$y(i, j) = \begin{cases} 1, & \text{if TF “} j \text{” affects gene “} i \text{”} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

In order to control the complexity of the approximation the following set of constraints has to be incorporated.

The first two equations guarantee that if a connection exists the corresponding weight is at least ϵ , whereas if the connection does not exist the corresponding weight should be

$$\begin{aligned} \frac{A(i, j) - A_{\min}}{A_{\max} - A_{\min}} &\geq \frac{-A_{\min}}{A_{\max} - A_{\min}} + \left[\epsilon + \frac{A_{\min}}{A_{\max} - A_{\min}} \right] y(i, j) \\ \frac{A(i, j) - A_{\min}}{A_{\max} - A_{\min}} &\leq \frac{-A_{\min}}{A_{\max} - A_{\min}} + \left[1 + \frac{A_{\min}}{A_{\max} - A_{\min}} \right] y(i, j) \end{aligned}$$

$$\sum_i \sum_j y(i, j) = N$$

$$\sum_j 1 - y(i, j) \geq L - 1, \quad \forall i$$

$$\sum_i y(i, j) \geq 1, \quad \forall j$$

$$(i, j) \notin Y^{\text{SuperSet}} \Rightarrow y(i, j) = 0, \quad \forall i, j$$

set to zero. The third equation is a critical one as it controls the expected complexity of the approximation that is the number of active gene–transcription factor interactions that are allowed in the model. Hence, the addition of this constraint guarantees that an approximation of minimum complexity will be identified. It must be emphasized that no such provisions were possible in the original formulation. The fourth equation is the requirement for a bound on the number of zeros for each column in $[A]$, the fifth constraint that each gene is regulated by at least one TF, whereas the final equation captures the known biology about the system, whereby we only allow for known regulatory interactions to be quantified.

The next part of the model tests for linear independence. We elected only to test $[A]$ because of the following notion. If E is of full row rank, and A is of full column rank, provided that the number of transcription factors is less than the number of genes and the number of time points, then the P matrix will also be full row rank. Therefore, we do not have to check for the row rank of $[P]$. The linear independence check utilizes the Cholesky-Infinity decomposition in order to determine if $A^T A$ will have a non-zero determinant. This is done by simply checking of the Cholesky-Infinity decomposition of $A^T A$ has any zero elements on the diagonal. The Cholesky-Infinity decomposition in an optimizations framework is essentially the same as the Cholesky factorization. However, there exists a small difference in that the Cholesky-Infinity decomposition handles positive semi-definite matrices whereas the Cholesky decomposition handles only positive definite matrices.

$$\begin{aligned} \mathbf{M} &= \mathbf{Y}^T \mathbf{Y} \\ \mathbf{M} &= \mathbf{C}^T \mathbf{C} \\ C(i, i) &\geq 0, \quad \forall i \end{aligned} \quad (6)$$

C is the Cholesky decomposition matrix of Y . To check for linear independence, we make sure that the diagonal entries of the Cholesky matrix are non-zero. The first bilinear constraint, $\mathbf{M} = \mathbf{Y}^T \mathbf{Y}$, involves the product of binary variables since $M(i, j) = \sum_k y(i, k)y(k, j)$. Following the reformulation proposed by Glover (1975) these non-convexities can be eliminated by introducing a new set of continuous variables $w(i, j, k) = y(i, k)y(k, j)$ and the following set of constraints for new variable:

$$\begin{aligned} 0 &\leq w(i, j, k) \leq y(i, k) \\ y(k, j) - [1 - y(i, k)] &\leq w(i, j, k) \leq y(k, j) \end{aligned} \quad (7)$$

The new continuous variables (w) and the associated constraints (7) eliminate a major source of non-convexities in our formulation (Fig. 5).

The next set of constraints has not been activated thus far in our development due to the lack of available experimental data, however, it is being discussed to demonstrate the versatility of the method as well as the potential extensions and advantages that an optimization-based framework can provide. It is conceivable that surrogate measurements for the strength of transcription factor activities could be estimated (Kao et al., 2004). In that case, we can explicitly bound the predictions for that transcription factor activity profile appropriately in order to account for additional information about the process and it is associated biology.

$$|P(j, t) - P^{\text{exp}}(j, t)| < \varepsilon \quad (8)$$

It should be pointed out that recent work on NCA extensions also points to the possibility of incorporating certain types of such constraints into the original formalism (Tran et al., 2005). In order to systematically generate structurally independent solutions that approximate the expression dynamics, the final set of constraints needs to be implemented. Once an matrix $[A]$ has been identified that meets all the NCA criteria and is of an acceptable error, the corresponding active connections are defined by the matrix $Y^{\text{active}} = \{(i, j) | y(i, j) = 1\}$. Addition of these “cuts” guarantees that resolving the problem with these additional constraints will generate a structurally distinct solution, i.e., one that does use the exact combination of $y(i, j)$ values (Biegler et al., 1997). This is probably one of the strongest elements of our formulation as we now have the ability to generate, systematically sets of optimal regulatory structures.

Given the existence of transcriptionally related signaling cascades which are associated with specific responses (Scott et al., 2002; Ogata et al., 1997), it seems plausible that only a subset of transcription factors will be directly contributing to an organism’s response. It is our hypothesis that the set of transcription–factor gene connections that are active under an experimental protocol is a subset of possible transcription factor–gene connections. The goal of miNCA is to identify the required set of active connections given both the expression data as well as the connectivity structure between the given transcription factors and the genes in question. However, to do so, we need to obtain the set of possible transcription factor–gene connections in the set of identified genes. Unlike in yeast however, where a large set of transcription factor–gene connections has been identified (Teixeira et al., 2006), analogous information in mammalian systems is much sparser. In order to obtain the necessary information then, we turn computational techniques for the prediction of transcription factor binding sites. The approach which we are utilizing focuses upon the use of position weight matrices (Wasserman & Sandelin, 2004), with the primary difference being that the output of our algorithm will not be a binary 1–0 matrix in which the transcription factor binds, or the transcription factor does not bind, but rather a probability matrix in which under a certain threshold, the probability of the transcription factor binding is zero, but over the threshold, the transcription factor binds with a certain non-zero probability.

$\min \sum_i \sum_t eP(i,t) + eN(i,t)$ <p>s.t.</p> $eP(i,t) \geq 0 \quad \forall i,t$ $eN(i,t) \geq 0 \quad \forall i,t$ $E(i,t) - \sum_j A(i,j)P(j,t) = eP(i,t) - eN(i,t) \quad \forall i,t$ $y(i,j) = \begin{cases} 1 & \text{if TF "j" affects gene "i"} \\ 0 & \text{otherwise} \end{cases}$ $\frac{A(i,j) - A_{\min}}{A_{\max} - A_{\min}} \geq \frac{-A_{\min}}{A_{\max} - A_{\min}} + \left[\varepsilon + \frac{A_{\min}}{A_{\max} - A_{\min}} \right] y(i,j)$ $\frac{A(i,j) - A_{\min}}{A_{\max} - A_{\min}} \leq \frac{-A_{\min}}{A_{\max} - A_{\min}} + \left[1 + \frac{A_{\min}}{A_{\max} - A_{\min}} \right] y(i,j)$ $\sum_j y(i,j) = N$ $\sum_j 1 - y(i,j) \geq L - 1 \quad \forall i$ $\sum_i y(i,j) \geq 1 \quad \forall j$ $(i,j) \notin Y^{\text{SuperSet}} \Rightarrow y(i,j) = 0 \quad \forall i,j$ $\mathbf{M} = \mathbf{Y}^T \mathbf{Y}$ $\mathbf{M} = \mathbf{C}^T \mathbf{C}$ $C(i,i) \geq 0 \quad \forall i$ $0 \leq w(i,j,k) \leq y(i,k)$ $y(k,j) - [1 - y(i,k)] \leq w(i,j,k) \leq y(k,j)$ $ P(j,t) - P^{\text{exp}}(j,t) < \varepsilon$ $IC^k = \left\{ \left(\sum_{(i,j) \in B^k} y(i,j) - \sum_{(i,j) \in N^k} y(i,j) \right) \leq N - 1 \right\}$ $B^k = \{(i,j) y^k(i,j) = 1\}, N^k = \{(i,j) y^k(i,j) = 0\}$ $\sum_{i,j} y(i,j) \pi_{ij} = \Pi$	<p>$eP, eN =$ positive slack variables</p> <p>$E(i,t) =$ mRNA transcripts of gene "i" at time "t"</p> <p>$A(i,j) =$ strength connectivity between gene "i" and transcription factor "j"</p> <p>$P(j,t) =$ transcription factor "j" activity at time "t"</p> <p>$y(i,j) =$ binary variable defining connectivity matrix</p> <p>$N =$ number of active regulatory connections</p> <p>$L =$ number of transcription factors</p> <p>$\mathbf{M}, \mathbf{C} =$ matrices defining Cholesky decomposition of \mathbf{Y} matrix</p> <p>$\pi_{ij} =$ probability of transcription factor "j" regulating gene "i"</p>
---	--

Fig. 5. Mixed-integer reformulation of NCA.

This was done since miNCA attempts to find the subset of active connections by pruning the initial set of TF–gene connections, we need to differentiate the transcription factors by their binding specificities, under the notion that genes which bind more tightly to their associated transcription factors are more likely to be affected by them.

The transcription factor binding matrices were obtained from TRANSFAC (Matys et al., 2003). We conducted 500 random trails and found that the distribution of scores is reasonably Gaussian as determined via the Lilliefors test (NIST, 1998). This allows us to define the probability of a score as in (10), under the assumption that a sequence with the maximum possible score must bind 100% of the time, and a sequence which is right above the cutoff, will have a slight possibility in binding.

$$p_{\text{gene}} = \frac{\text{erf}(x/\sqrt{2}) - \text{erf}(2.5/\sqrt{2})}{1 - \text{erf}(2.5/\sqrt{2})} \quad (9)$$

P is the probability of a transcription factor being active and x is the number of standard deviations away the score is.

This formula compares the respective cumulative distribution of the cutoff which in this case is $\sigma = 2.5$ and the calculated score (X) for the transcription factor and promoter region. The probability is then calculated so that if X equals the cutoff, the probability is zero, and if X is much greater than 2.5σ then the binding probability is essentially one.

The equation was formulated this way so that the cutoff where binding is impossible is preserved, and so the probability function is still continuous.

3.2. Results

Experimental DNA microarray data were obtained from the GEO database available at <http://www.ncbi.nlm.nih.gov/geo> under the accession number GSE802. In this previously published study, male Sprague–Dawley rats were subjected to a

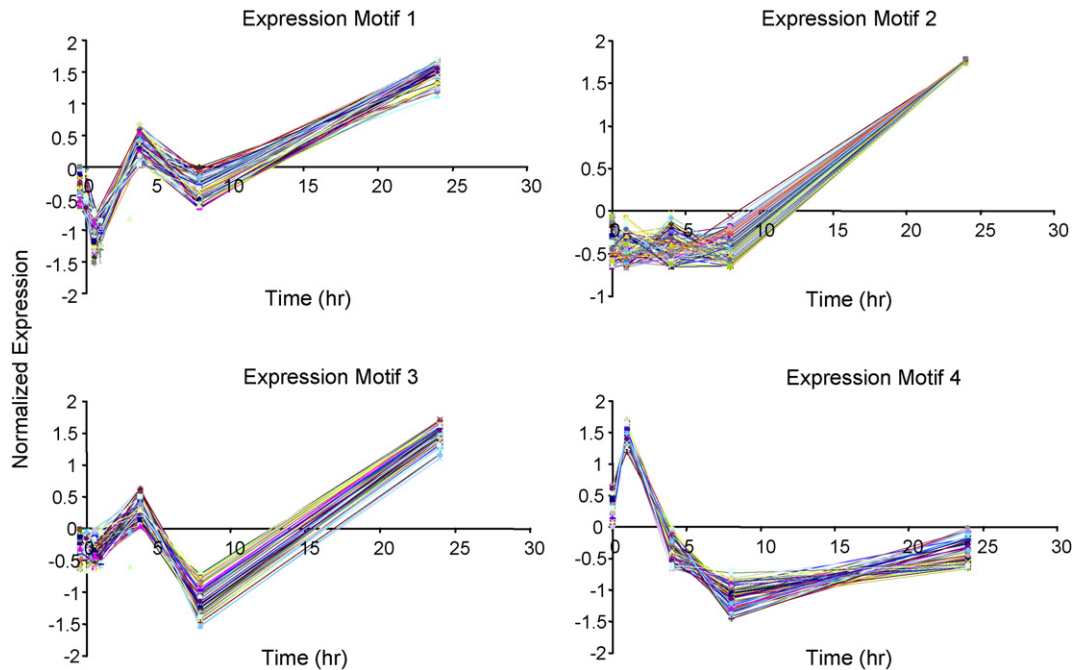


Fig. 6. Informative expression motifs.

cutaneous third degree burn injury consisting of a full skin thickness scald burn of the dorsum, calculated to be $\sim 20\%$ of the rat's total body surface area (Vemula et al., 2004). Liver samples were obtained at 5 time points, 1, 4, 8 and 24 h post-burn. mRNA extracted from the livers was isolated and subsequently hybridized to a U34A GeneChip that had 8799 probes represented on each chip. The control for this experiment was obtained at time 0, which was prior to the injury. It has been previously shown that time had so significant effect upon the response of rats to the sham treatment (Lee et al., 2003).

In order to build a model which accurately reflects the underlying phenomenon being investigated, it is first important to select the relevant expression profiles with which to fit. These expression profiles are extracted from the overall set of 8799 genes via a novel gene selection algorithm that we have developed (Yang et al., 2006; Yang, Roth, et al., 2005) which utilizes hashing in order to perform a fine grained clustering which clusters the expression profiles into specific motifs. The motifs are then selected based upon their ability to reproduce the dynamic response of an experiment. This was done by utilizing the Kolmogorov–Smirnov Statistic in order to quantify the differences in the distributions of gene expressions at any specific time point. Genes were selected based on the ability of their motifs to reproduce the maximum deviations from the baseline distribution (Fig. 6).

An initial list of transcription factors was extracted via Trafac (Jegga et al., 2002) with a promoter region of 200 base pairs upstream of the start codon of the informative genes. We selected 200 base pairs upstream for our transcription factor analysis given the results from Taylor et al. (2006), which suggested that the region of promoter sequence homology between species was at 200 base pairs or less upstream of the start codon. The selection of Trafac and the associated use of BLASTZ is not opti-

mal and it is well understood that the prediction of regulatory elements remains a wonderful and open problem and various assessments have been provided in the open literature (Tompa et al., 2005). In total 139 transcription factors families were identified in the Trafac database, however, what is most interesting is the fact that we identify a significant enrichment for each class of informative motifs and further verified the identification of known critical inflammation-specific TFs. The relative enrichment in specific TFs for each cluster can be calculated and subsequently each cluster can be evaluated.

The purpose of our analysis was to identify regulatory structures of varying complexity able to reconstruct the transcriptional signals that were measured. Out of the 154 possible connections (Gene–TF combinations with non-zero probabilities) we seek to identify the least number of possible connections that would accurately reproduce the expression dynamics. The miNCA formulation overcomes the difficulty of searching via heuristic method for a connectivity matrix which satisfies the NCA requirements. A key advantages of miNCA, is also the ability to systematically identify multiple solutions from a given super-set of possible interactions. Therefore, after verifying the ability to generate a solution that was able to fit the expression data reasonably well, the second task was to determine whether or not extending the mixed-integer optimization step would allow for the identification of significant connections. By activating the cuts constraints, we generated a series of solutions which would fit the data with similar accuracy and we attempted to determine if certain connections were in fact conserved. Solving the problem parametrically in the number of active connections, optimization for the likelihood of the regulatory structure and estimation error, we obtain an ensemble of solutions with diverse average likelihood and approximation error (Fig. 7).

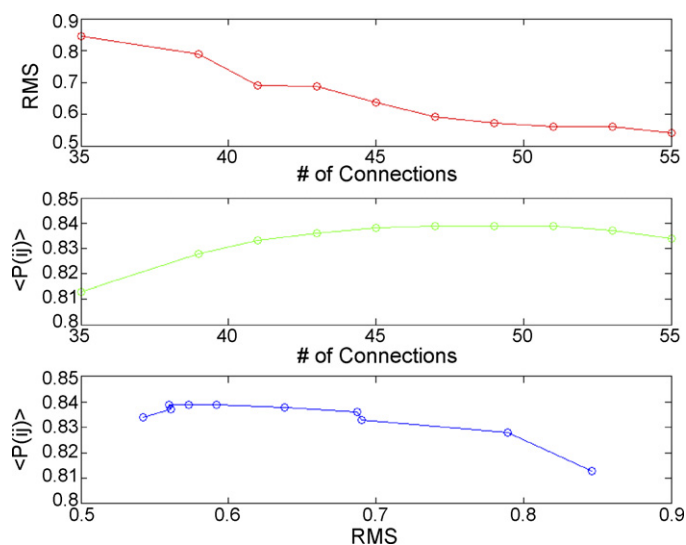


Fig. 7. Multiple miNCA solutions.

Examination of the transcription factor binding suggests that perhaps by taking the transcription factor with the greatest probability for each row, one would be able to obtain a A_0 for use in NCA decomposition. And while this can lead to a valid NCA

compliant A_0 , it however does not return the solution with the lowest error. While the optimal solution does include the transcription factors and genes with the highest probability (from Fig. 7 (bottom panel)), it becomes evident that the solution with the highest average probability does not lead to the lowest error, but more importantly that the optimal solution at 35 connections (the fewest possible connections given a naïve strategy for selecting A_0) is significantly worse than the solution reached with 49 connections, signifying the need to include the less probable connections as well. It is the need to identify these less probably connections that arrant the use of miNCA. Figs. 8 and 9 show the results of the decomposition with Fig. 8 showing the predicted transcription factor activity. These profiles are normalized per the procedure suggested in Tran et al. (2005) and Fig. 9 depicts the results of the reconstructions, i.e., the expression profiles obtained when the predicted $[A]$ and $[P]$ matrices were multiplied together. The results of the decomposition appear to fit the raw data well.

Given the linear nature of the model in question, the anti-correlated solutions are to be expected. Essentially there exists a $[X]$ diagonal scaling matrix with a different sign that would end up flipping the predicted transcription factor activity. While it is known that the gene expression levels of transcription factors are

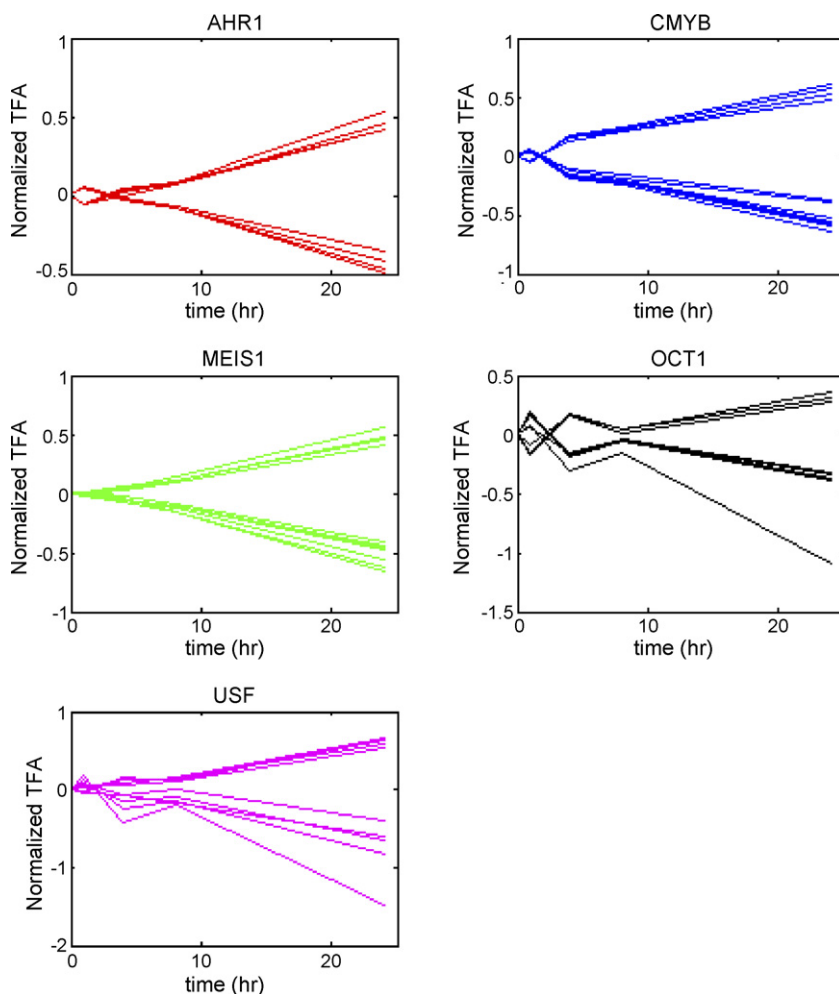


Fig. 8. Transcription factor activity reconstruction.

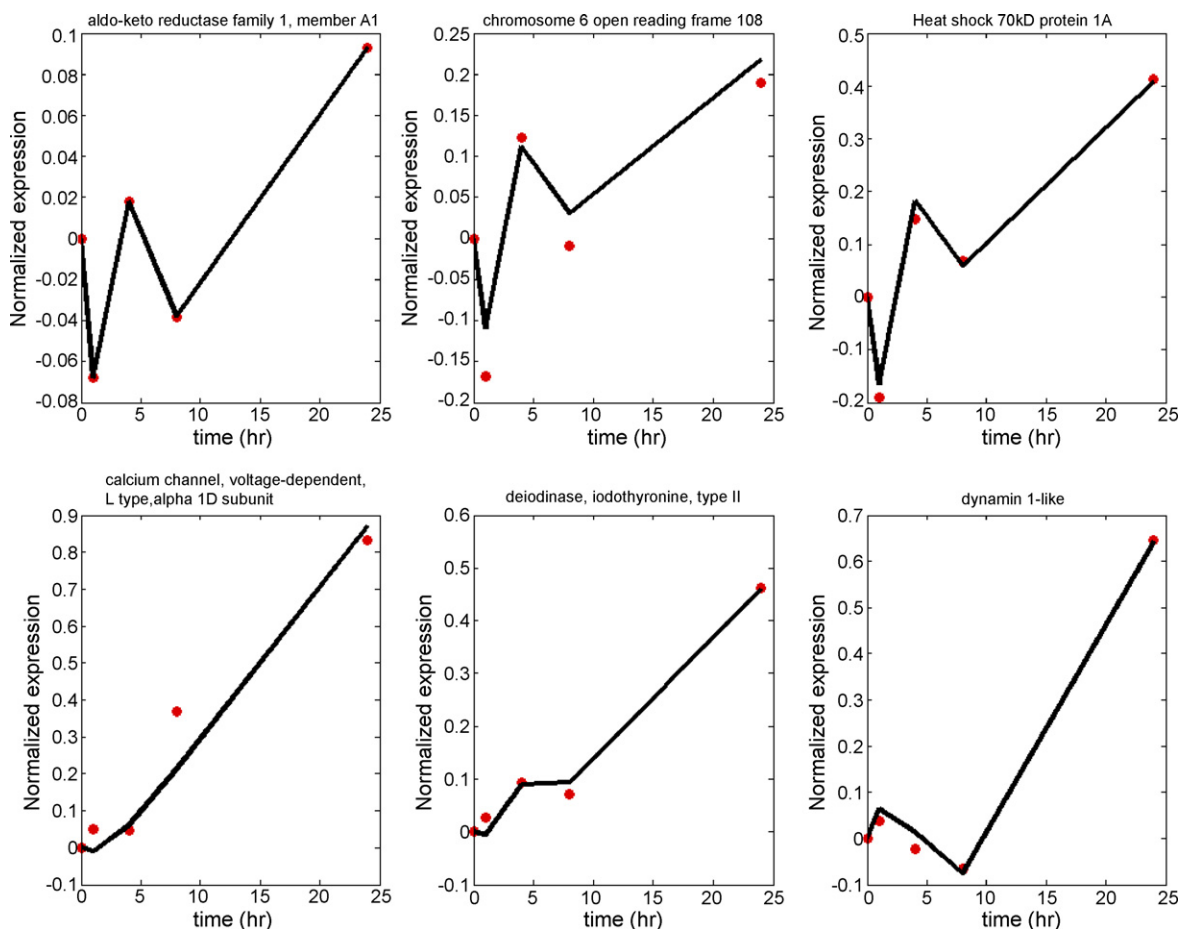


Fig. 9. Gene expression reconstruction (symbols: raw data, line: reconstruction).

not good analogues for transcription factor activity (Liao et al., 2003), due to complications in phosphorylation, dimerization, or transport kinetics we believe that the expression data of these transcription factors can at least provide a hint as to the overall up or down-regulation of that particular factor. Therefore, after we have solved for the overall shape, we can determine either the up or down-regulation of the transcription factor by looking at the overall expression profiles of the transcription factor. In Fig. 10, we can see the expression profiles for one of the transcription factors. While it does not track the predicted transcription factor activity, it does however provide one important piece of information, namely that the positive activity should be selected.

Due to the high degree of correlation Fig. 11 we believe that the generalized shape has been correctly predicted. What we can do at this point to determine whether or not the transcription factor activity is being up-regulated or down-regulated is to utilize information from the original microarray. Therefore, the important character which we wish to observe is the correlation of the solution. By looking at the correlation, we are able to evaluate how conserved the solution is over different combinations of transcription factor–gene connections. Additionally, looking over the space of multiple solutions, we find that for the highly correlated genes, their conservation also tracks closely with that of the gene enrichments. Calculating gene enrichments we find

that the enrichment for AHR was greatest in Cluster 1 and the enrichment for MEIS was greatest in Cluster 2. Likewise looking at the number of possible connections, AHR was found to be more likely to interact with Cluster 1 than any other cluster. MEIS on the other hand was more likely to interact with Cluster 2.

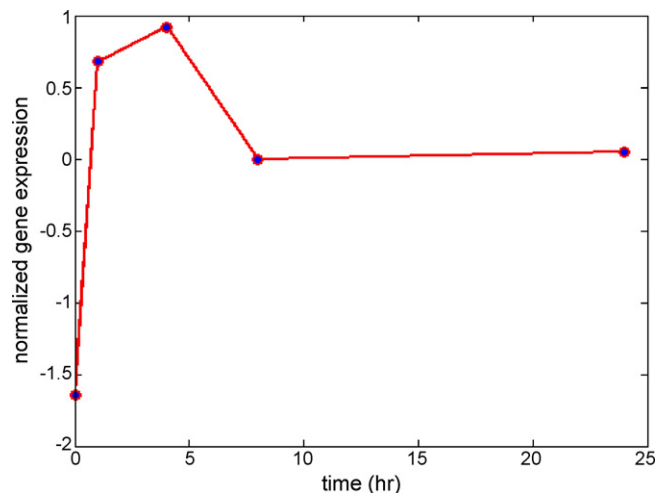


Fig. 10. mRNA data for the AHR transcription factor.

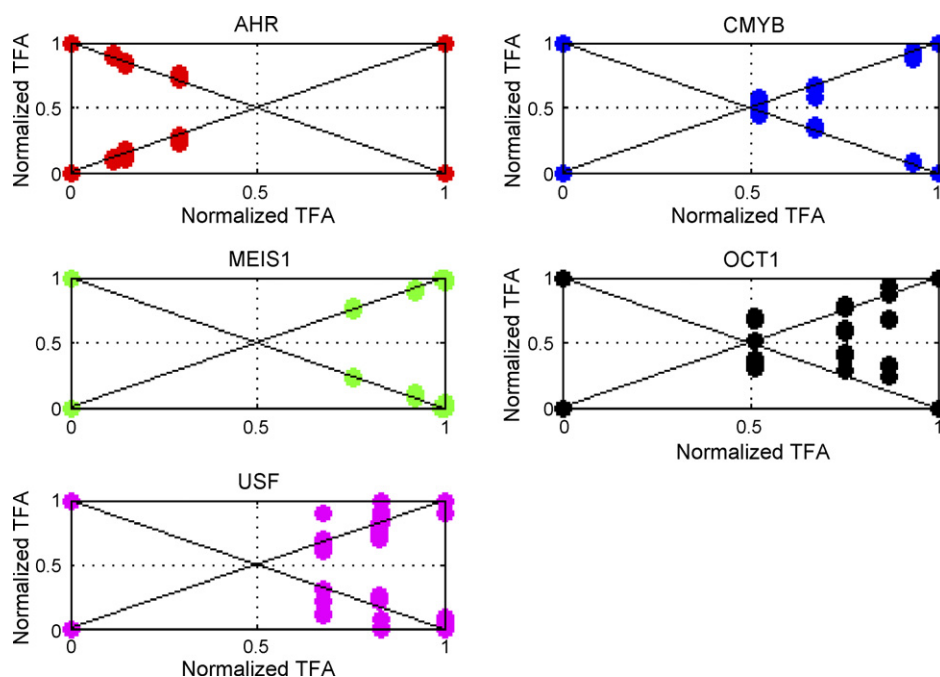


Fig. 11. Correlation among multiple TFA profiles.

From these results, we make the hypothesis that the transcription factors that show the greatest degree of correlation between solutions, i.e., exhibit unique profiles across multiple structurally different model realizations, are the most biologically significant ones under the experimental regimes. By clustering gene expression data, we obtain a set of genes which are responding to a given stimulus. By conducting miNCA upon this group of genes, the transcription factors which show the greatest conservation are the ones which are least affected by the inherent randomness we have inserted into the system to find multiple solutions, and would henceforth be the least affected by inherent biological noise. That information can be useful if we wish to determine which genes to silence. For instance to eliminate a response in its entirety, one would target transcription factors which show responses which are less correlated with themselves and do not reflect any of the activities of a given cluster. If one wanted to eliminate the response of a given cluster, then one would silence the transcription factors which show both high correlation in their solutions as well as those which seem to trace the overall activity of the genes they regulate.

In our motivating study we examined the reconstructed activity profiles of the following transcription factors. Aryl hydrocarbon receptor 1 (AHR1), transcription factor CMYB (CMYB), myeloid ecotropic viral integration site 1 (MEIS1), octamer-binding transcription factor 1 (OCT1), and upstream stimulatory factor 1 (USF) have all been demonstrated as integral in inflammatory response processes (Lee, Lee, et al., 2002; Schroer et al., 2002; Sollars et al., 2002; Vrzal et al., 2004). Of these transcription factors, AHR1 exhibits the highest level of robustness, and also is the most documented in terms of regulating key phases of the inflammatory response process, specifically through the control of cytochrome P450s such as Cyp4501A1. An interesting finding, upon analyzing the kinetic

profile of AHR1 mRNA expression, AHR1 activity, and the expression profile of motif 1, in which AHR1 is the most enriched, is that mRNA levels for AHR1 are up-regulated at 2 h post-burn, followed by the initiation of AHR1 activity, at 4 h post-burn, which is coupled with the initiation of up-regulation of mRNA for genes contained within motif 1. This makes sense biologically, being that the paradigm in biology is the up-regulation of mRNA for a TF, followed by the translation of the mRNA to the protein product for the TF, and then the subsequent initiation of transcription initiated by the binding of the TF to its specific sites in the promoter regions of its related genes. In addition, these kinetics for AHR1, i.e., a 2 h separation between the quantified up-regulation in TF mRNA levels and its subsequent transcriptional activity, have been experimentally validated, and response times as short as 30 min have even been detected in other models of cellular stress response (Wang et al., 2004). Due to the robust activity AHR1 displays, in conjunction with its expected temporal profile of up-regulation at the mRNA level and at the level of transcriptional regulation, as aforementioned, AHR1 may serve as a beneficial point of clinical intervention. For example, it has been demonstrated that the use of small interfering RNA (siRNA) is capable of inhibiting transcriptional activation by AHR, resulting in the down-regulation of AHR responsive genes (Abdelrahim et al., 2003; Chen et al., 2006; Miao et al., 2005). Thus, new age therapies may introduce the use of siRNA, specific for AHR1, in order to modulate the level of AHR1 activity, and hence the levels of production and activity of AHR1 responsive genes. Network component analysis offers the possibility of predicting actual transcription factor activities which may not necessarily coincide with mRNA levels of the transcription factors (Yang, Suen, et al., 2005). However, they do capture the expected overall dynamic response of the activity of the transcription factors. What is quite remarkable is that

despite the fact that the activity reconstructions were performed on different regulatory structures (i.e., different regulatory architectures) significant consistencies do exist between the various profiles corresponding to the multiple solutions.

3.3. Discussion

The versatility of the formulation does not come without any cost. Two main issues remain to be resolved and we are currently pursuing them aggressively. First, the complexity of the problem is increased significantly as we need to introduce $M \times N$ binary variables which significantly increase the complexity of the problem but to its integer optimization character (Floudas, 1995). Second, the formulation is non-linear and non-convex with the primary source of non-convexity be the bilinear terms in Eqs. (4) and (7). As a result, global optimality of the corresponding problem cannot be guaranteed unless appropriate global optimization algorithms are implemented (Adjiman et al., 2000; Sahinidis, 1996).

4. Concluding remarks

Optimization has been greeted with renewed enthusiasm in “non-traditional” areas of process systems engineering (Floudas, 2005). We have discussed opportunities and challenges of integer optimization formulation in problems in bioinformatics and systems biology. Mathematical programming offers the possibility of extending available models and maximizing the information upgrade from experimental data by allowing the rational inclusion of biological constraints and the systematic generation of alternatives. Among the main challenges we identify primarily the combinatorial complexity and the non-linear, non-convex nature of the resulting models, and also the general issues associated with the computational cost for the solution of large scale (linear and non-linear) integer optimization problems. However, all are active areas of scientific research.

Acknowledgments

This work was partially supported by grants from the Shriners Hospitals for Children. IPA and EY acknowledge financial support from NSF-BES 0519563 and EPA GAD R 832721-010.

References

- Abdelrahim, M., Smith, R., 3rd, et al. (2003). Aryl hydrocarbon receptor gene silencing with small inhibitory RNA differentially modulates Ah-responsiveness in MCF-7 and HepG2 cancer cells. *Molecular Pharmacology*, 63(6), 1373–1381.
- Adjiman, C. S., Androulakis, I. P., et al. (2000). Global optimization of mixed-integer nonlinear problems. *Aiche Journal*, 46(9), 1769–1797.
- Alizadeh, A. A., Eisen, M. B., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769), 503–511.
- Allander, S. V., Nupponen, N. N., et al. (2001). Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Research*, 61(24), 8624–8628.
- Alon, U., Barkai, N., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceeding of the National Academy of Sciences of the United States of America*, 96(12), 6745–6750.
- Andrews, G. K. (2000). Regulation of metallothionein gene expression by oxidative stress and metal ions. *Biochemical Pharmacology*, 59(1), 95–104.
- Androulakis, I. P. (2005). Selecting maximally informative genes. *Computers & Chemical Engineering*, 29(3), 535–546.
- Baer, C., Nees, M., et al. (2004). Profiling and functional annotation of mRNA gene expression in pediatric rhabdomyosarcoma and Ewing’s sarcoma. *International Journal of Cancer*, 110(5), 687–694.
- Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2), 101–113.
- Bassett, D. E., Jr., Eisen, M. B., et al. (1999). Gene expression informatics—it’s all in your mine. *Nature Genetics*, 21(Suppl. 1), 51–55.
- Biegler, L. T., Grossmann, I. E., et al. (1997). *Systematic methods of chemical process design*. Prentice Hall.
- Bittner, M., Meltzer, P., et al. (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, 406(6795), 536–540.
- Boscolo, R., & Sabatti, C., et al. (2004). Reconstructing hidden regulatory layers by network component analysis: Theory and application, <http://www.ee.ucla.edu/%7Ericcardo/NCA/Boscolo-TCBB-0516.pdf>.
- Bowtell, D. D. (1999). Options available – from start to finish – for obtaining expression data by microarray. *Nature Genetics*, 21(Suppl. 1), 25–32.
- Busygin, S., Prokopyev, O. A., et al. (2005). Feature selection for consistent biclustering via fractal 0–1 programming. *Journal of Combinatorial Optimization*, 10(1), 7–21.
- Cerisano, V., Aalto, Y., et al. (2004). Molecular mechanisms of CD99-induced caspase-independent cell death and cell–cell adhesion in Ewing’s sarcoma cells: Actin and zyxin as key intracellular mediators. *Oncogene*, 23(33), 5664–5674.
- Chen, Y. H., Beischlag, T. V., et al. (2006). Role of GAC63 in transcriptional activation mediated by the aryl hydrocarbon receptor. *Journal of Biological Chemistry*, 281(18), 12242–12247.
- Cheung, V. G., Morley, M., et al. (1999). Making and reading microarrays. *Nature Genetics*, 21(Suppl. 1), 15–19.
- Chilingaryan, A., Gevorgyan, N., et al. (2002). Multivariate approach for selecting sets of differentially expressed genes. *Mathematical Biosciences*, 176(1), 59–69.
- Dasika, M. S., Gupta, A., et al. (2004). A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. *Pacific Symposium in Biocomputing*, 474–485.
- Dettling, M., & Buhlmann, P. (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1), 106–131.
- Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2(1), 28–34.
- Dudoit, S., Yang, Y. H., et al. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1), 111–139.
- Duin, R. (2000). Classifiers in almost empty spaces. In *ICPR15 15th Int. Conference on Pattern Recognition*.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(11), 179–188.
- Floudas, C. A. (1995). *Nonlinear and mixed-integer optimization: Fundamentals and applications*. Oxford University Press.
- Floudas, C. A. (2005). Research challenges, opportunities and synergism in systems engineering and computational biology. *Aiche Journal*, 51(7), 1872–1884.
- Fraenzer, J. T., Pan, H., et al. (2003). Overexpression of the NF2 gene inhibits schwannoma cell proliferation through promoting PDGFR degradation. *International Journal of Oncology*, 23(6), 1493–1500.
- Freed, N., & Glover, F. (1981a). A linear programming approach to the discriminant problem. *Decision Sciences*, 12, 68–74.
- Freed, N., & Glover, F. (1981b). Simple but powerful goal programming for the discriminant problem. *European Journal of Operational Research*, 7, 44–60.
- Freed, N., & Glover, F. (1986). Evaluating alternative linear programming formulations for the discriminant problem. *Decision Sciences*, 17, 151–162.

- Gehrlein, W. V. (1986). General mathematical programming formulations for the statistical classification problem. *Operations Research Letters*, 5(6), 299–304.
- Glover, F. (1975). Improved Linear integer Programming Formulations of Non-linear Integer Problems. *Management Science*, 22(4), 455–460.
- Glover, F. (1990). Improved linear programming models for discriminant analysis. *Decision Sciences*, 21, 771–785.
- Glover, F., Keene, S., et al. (1988). A new class of models for the discriminant problem. *Decision Sciences*, 19, 269–280.
- Golub, T. R., Slonim, D. K., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439), 531–537.
- Guyon, I., Weston, J., et al. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3), 389–422.
- Herrgard, M. J., Covert, M. W., et al. (2004). Reconstruction of microbial transcriptional regulatory networks. *Current Opinion in Biotechnology*, 15(1), 70–77.
- Ho, T. (2002). A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis Applications*, 5, 102–112.
- Holter, N. S., Mitra, M., et al. (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(15), 8409–8414.
- Huang, H., Hara, A., et al. (2005). Altered expression of immune defense genes in pilocytic astrocytomas. *Journal of Neuropathology and Experimental Neurology*, 64(10), 891–901.
- Iannarilli, F. J., & Rubin, P. A. (2003). Feature selection for multiclass discrimination via mixed-integer linear programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6), 779–783.
- Jain, A., & Zongker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 153–158.
- Jegga, A. G., Sherwood, S. P., et al. (2002). Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Research*, 12(9), 1408–1417.
- Kafatos, F. C. (2002). A revolutionary landscape: The restructuring of biology and its convergence with medicine. *Journal of Molecular Biology*, 319(4), 861–867.
- Kao, K. C., Yang, Y. L., et al. (2004). Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2), 641–646.
- Khan, J., Wei, J. S., et al. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Natural Medicine*, 7(6), 673–679.
- Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11), 826–837.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.
- Lee, K., Berthiaume, F., et al. (2003). Profiling of dynamic changes in hypermetabolic livers. *Biotechnology and Bioengineering*, 83(4), 400–415.
- Lee, I. K., Lee, Y. M., et al. (2002). Hepatobiliary excretion of tributylmethylammonium in rats with lipopolysaccharide-induced acute inflammation. *Archives of Pharmacological Research*, 25(6), 969–972.
- Lee, H., Park, D. S., et al. (2002). Caveolin-1 mutations (P132L and null) and the pathogenesis of breast cancer: Caveolin-1 (P132L) behaves in a dominant-negative manner and caveolin-1(–/–) null mice show mammary epithelial cell hyperplasia. *American Journal of Pathology*, 161(4), 1357–1369.
- Liao, J. C., Boscolo, R., et al. (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26), 15522–15527.
- Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1), 51–60.
- Lin, H. J., Shaffer, K. M., et al. (2004). AF1q, a differentially expressed gene during neuronal differentiation, transforms HEK cells into neuron-like cells. *Brain Research. Molecular Brain Research*, 131(1–2), 126–130.
- Lipshutz, R. J., Fodor, S. P., et al. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(Suppl. 1), 20–24.
- Liu, H., & Motoda, H. (2000). *Feature selection for knowledge discovery and data mining*. Kluwer Academic Publishers.
- Lomas, J., Bello, M. J., et al. (2005). Genetic and epigenetic alteration of the NF2 gene in sporadic meningiomas. *Genes Chromosomes Cancer*, 42(3), 314–319.
- Luo, J., Duggan, D. J., et al. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Research*, 61(12), 4683–4688.
- Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13, 444–452.
- Mangasarian, O. L. (1968). Multi-surface method of pattern separation. *IEEE Transactions in Information Theory*, IT-14, 801–807.
- Matys, V., Fricke, E., et al. (2003). TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research*, 31(1), 374–378.
- Miao, W., Hu, L., et al. (2005). Transcriptional regulation of NF-E2 p45-related factor (NRF2) expression by the aryl hydrocarbon receptor-xenobiotic response element signaling pathway: Direct cross-talk between phase I and II drug-metabolizing enzymes. *Journal of Biological Chemistry*, 280(21), 20340–20348.
- Narendra, P., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions of Computers*, C-26(9), 917–926.
- NIST. (1998). *e-Handbook of Statistical Methods*. SEMATECH.
- Ogata, A., Chauhan, D., et al. (1997). IL-6 triggers cell growth via the Ras-dependent mitogen-activated protein kinase cascade. *Journal of Immunology*, 159(5), 2212–2221.
- Perou, C. M., Jeffrey, S. S., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16), 9212–9217.
- Pollack, J. R., Perou, C. M., et al. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genetics*, 23(1), 41–46.
- Qian, J., Lin, J., et al. (2003). Prediction of regulatory networks: Genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19(15), 1917–1926.
- Quandt, K., Frech, K., et al. (1995). MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research*, 23(23), 4878–4884.
- Rajjayabun, P. H., Garg, S., et al. (2001). Caveolin-1 expression is associated with high-grade bladder cancer. *Urology*, 58(5), 811–814.
- Raychaudhuri, S., Stuart, J. M., et al. (2000). Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Pacific Symposium on Biocomputing*, 455–466.
- Ross, D. T., Scherf, U., et al. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24(3), 227–235.
- Ryu, C. H., Kim, S. W., et al. (2005). The merlin tumor suppressor interacts with Ral guanine nucleotide dissociation stimulator and inhibits its activity. *Oncogene*, 24(34), 5355–5364.
- Sahinidis, N. V. (1996). BARON: A general purpose global optimization software package. *Journal of Global Optimization*, 8(2), 201–205.
- Schena, M., Shalon, D., et al. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–470.
- Schmitt, W. A., Jr., Raab, R. M., et al. (2004). Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. *Genome Research*, 14(8), 1654–1663.
- Schroer, K., Zhu, Y., et al. (2002). Obligatory role of cyclic adenosine monophosphate response element in cyclooxygenase-2 promoter induction and feedback regulation by inflammatory mediators. *Circulation*, 105(23), 2760–2765.
- Scott, M. J., Godshall, C. J., et al. (2002). Jaks, STATs, Cytokines, and Sepsis. *Clinical and Diagnostic Laboratory Immunology*, 9(6), 1153–1159.
- Shioda, R. (2003). *Integer optimization in data mining, operations research*. Boston: MIT.
- Sollars, V. E., McEntee, B. J., et al. (2002). A novel transgenic line of mice exhibiting autosomal recessive male-specific lethality and non-alcoholic fatty liver disease. *Human Molecular Genetics*, 11(22), 2777–2786.
- Sonveaux, P., Martinive, P., et al. (2004). Caveolin-1 expression is critical for vascular endothelial growth factor-induced ischemic hindlimb collateraliza-

- tion and nitric oxide-mediated angiogenesis. *Circulation Research*, 95(2), 154–161.
- Stam, A. (1997). Nontraditional approaches to statistical classification: Some perspectives on L_p-norm methods. *Annals of Operations Research*, 74(0), 1–36.
- Street, W. N. (2005). Oblique multicategory decision trees using nonlinear programming. *Inform Journal on Computing*, 17(1), 25–31.
- Sun, M., & Xiong, M. (2002). A mathematical programming approach for gene selection and tumor classification. *American Journal of Human Genetics*, 71(4), 229–229.
- Szabo, A., Boucher, K., et al. (2002). Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*, 176(1), 71–98.
- Taylor, M. S., Kai, C., et al. (2006). Heterotachy in mammalian promoter evolution. *PLoS Genet*, 2(4), e30.
- Teixeira, M. C., Monteiro, P., et al. (2006). The YEASTRACT database: A tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Research*, 34(Database issue), D446–D451.
- Tompa, M., Li, N., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1), 137–144.
- Tran, L. M., Brynildsen, M. P., et al. (2005). gNCA: A framework for determining transcription factor activity based on transcriptome: Identifiability and numerical implementation. *Metabolic Engineering*, 7(2), 128–141.
- Uney, F., Turkay, M. (in press). A mixed-integer programming approach to multi-class data classification problem. *European Journal of Operations Research*.
- Vemula, M., Berthiaume, F., et al. (2004). Expression profiling analysis of the metabolic and inflammatory changes following burn injury in rats. *Physiology Genomics*, 18(1), 87–98.
- Vrzal, R., Ulrichova, J., et al. (2004). Aromatic hydrocarbon receptor status in the metabolism of xenobiotics under normal and pathophysiological conditions. *Biomedical Papers of the Medical Faculty of University of Palacky Olomouc Czech Republic*, 148(1), 3–10.
- Wang, S., Ge, K., et al. (2004). Role of mediator in transcriptional activation by the aryl hydrocarbon receptor. *Journal of Biological Chemistry*, 279(14), 13593–13600.
- Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4), 276–287.
- Wei, G. H., Liu, D. P., et al. (2004). Charting gene regulatory networks: Strategies, challenges and perspectives. *Biochemistry Journal*, 381(Pt 1), 1–12.
- Weir, M. L., & Muschler, J. (2003). Dystroglycan: Emerging roles in mammary gland function. *Journal of Mammary Gland Biology Neoplasia*, 8(4), 409–419.
- Williams, T. M., Cheung, M. W., et al. (2003). Loss of caveolin-1 gene expression accelerates the development of dysplastic mammary lesions in tumor-prone transgenic mice. *Molecular Biology of Cell*, 14(3), 1027–1042.
- Williams, T. M., Hassan, G. S., et al. (2005). Caveolin-1 promotes tumor progression in an autochthonous mouse model of prostate cancer: Genetic ablation of Cav-1 delays advanced prostate tumor development in tramp mice. *Journal of Biological Chemistry*, 280(26), 25134–25145.
- Williams, T. M., Medina, F., et al. (2004). Caveolin-1 gene disruption promotes mammary tumorigenesis and dramatically enhances lung metastasis in vivo. Role of Cav-1 in cell invasiveness and matrix metalloproteinase (MMP-2/9) secretion. *Journal of Biological Chemistry*, 279(49), 51630–51646.
- Xiao, G. H., Chernoff, J., et al. (2003). NF2: The wizardry of merlin. *Genes Chromosomes Cancer*, 38(4), 389–399.
- Yang, E., Berthiaume, F., et al. (2006). An integrative systems biology approach for analyzing liver hypermetabolism. In *16th European Symposium on Computer Aided Process Engineering and 9th Int. Symp. Process Systems Engineering Garmisch-Partenkirchen*. Germany: Elsevier.
- Yang, E., Roth, C. M., et al. (2005). New approaches for enabling temporal expression profiling analysis. In *AICHE Annual Meeting*.
- Yang, Y. L., Suen, J., et al. (2005). Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics*, 6(1), 90.
- Zhu, X., Meng, G., et al. (2001). MHC class I-related neonatal Fc receptor for IgG is functionally expressed in monocytes, intestinal macrophages, and dendritic cells. *Journal of Immunology*, 166(5), 3266–3276.