

# Mathematical Programming Approaches for the Analysis of Microarray Data

Ioannis P. Androulakis

Biomedical Engineering Department and Chemical and Biochemical Engineering Department, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854  
yannis@rci.rutgers.edu

**Abstract.** One of the major challenges facing the analysis of high-throughput microarray measurements is how to extract in a systematic and rigorous way the biologically relevant components from the experiments in order to establish meaningful connections linking genetic information to cellular function. Because of the significant amount of experimental information that is generated (expression levels of thousands of genes), computer-assisted knowledge extraction is the only realistic alternative for managing such an information deluge. Mathematical programming offers an interesting alternative for the development of systematic methodologies aiming towards such an analysis. We summarize recent developments related to critical problems in the analysis of microarray data, namely: tissue clustering and classification, informative gene selection and reverse engineering of gene regulatory networks. We demonstrate how advances in non linear and mixed-integer optimization provide the foundations for the rational identification of critical features unraveling fundamental elements of the underlying biology thus enabling the interpretation of volumes of biological data. We conclude the discussion by identifying a number of related research challenges and opportunities for further research.

## 1.1 Microarrays and the New Biology

The genetic information is stored in the DNA, the double-stranded polymer composed of four basic molecular units (nucleotides) adenine (A), guanine (G), cytosine (C), and thymine (T). In order for the genome to direct, or affect, changes in the cell a transcriptional program must be activated eventually dictating all biological transformations. This program is regulated temporarily according to an intrinsic program or in response to changes in the environment. The expression of the genetic information, which is stored in DNA, takes place in two stages: transcription, during which DNA is transcribed into mRNA, a single stranded complimentary copy of the base sequence of the DNA, and translation, during which mRNA provides the blue-print for the production of specific proteins. Measuring the level of production of mRNA,

thus measuring the expression levels of the associated genes, provides a quantitative assessment of the levels of production of the corresponding proteins, the ultimate expression of the genetic information.

Innovative approaches such as cDNA and oligonucleotide microarrays were recently developed to extract genome-wide information related to gene expression (see Schena et al. [50], Bowtell [3], Brown and Botstein [7], Cheung et al. [9], and Lipshutz et al. [41]). During an expression experiment extracted mRNA is reverse-transcribed into more stable complementary DNA (cDNA), which is labeled using fluorescent dyes. Different colored dyes are used for different samples (probes). The probes are then tested by hybridizing to a DNA array holding thousands of spots, each containing a different DNA sequence. Once the probes have hybridized, they are washed off and the array is scanned to determine the relative amount of each cDNA probe bound to any given spot. Quantitative imaging coupled with clone database information allows measurement of the labeled cDNA that hybridized to each target sequence. Image processing and data normalization are among the first, and very critical, computational filters required before the actual quantification of the expression experiment is defined (Dudoit et al. [17]). Gene expression changes are usually measured relative to another sample. Comparative differences are used to assess the impact of gene expression to various regulatory pathways.

Gene expression microarray experiments have been celebrated as a revolution in biology, attracting significant interest, because they are slowly changing the working paradigm of biological research by allowing the analysis of the combined effects of numerous genetic and environmental components. The profound impact is that such a global approach will allow a fundamental shift from “. . . piece-by-piece to global analysis and from hypothesis driven research to discovery-based formulation and subsequent testing of hypotheses. . .” (see Kafatos [39]). One of the major challenges is to extract in a systematic and rigorous way the biologically relevant components from the array experiments in order to establish meaningful connections linking genetic information to cellular function. Because of the significant amount of experimental information that is generated (expression levels of thousands of genes) computer-assisted knowledge extraction is the only realistic alternative for managing such an information deluge.

## 1.2 Issues in Microarray Data Analysis

Among the numerous tasks that can be assisted by the data generated from microarray experiments we will focus mainly on three: tissue classification, gene selection, and construction of regulatory networks from temporal gene expression data. We do so because

- a) these tasks are critical and define the basis for a number or more complicated problems,

- b) they have clearly defined approaches based on mathematical programming techniques and can be used as excellent motivating examples.

In tissue classification, samples from multiple cell types (for example different cancer types, cancerous and normal cells etc.) are comparatively analyzed using microarray gene expression measurements. The question therefore becomes how to identify which genes provide consistent signatures that distinctly characterize the different classes. The problem can be viewed as either a supervised classification problem in which the classes are already known, or as an unsupervised clustering problem in which we attempt to identify the classes contained within the data. In gene selection the computational problem is equivalent to that of feature selection in multidimensional data sets. Identifying the minimum number of gene markers is however critical because this reduced set can provide information about the biology behind the experiment as well as define the basis for future therapeutic agents.

In time-ordered gene expression measurements, the temporal pattern of gene expression is investigated by measuring the gene expression levels at a small number of points in time. The continuous monitoring of the level of mRNA abundance has the ultimate goal of deriving the temporal evolution of the synergistic effects of multiple genes. By doing so, a regulatory network is constructed, that is a biologically plausible superstructure of gene interactions that interprets the data. Transcriptional regulatory networks are the key to understanding the sequence of events leading to an observed biological response. The tasks that we are about to discuss in this chapter have already been addressed by a number of approaches under the general umbrella defined as *data mining*. What we plan to present however is a definition of these tasks as mathematical programming problems exploring principles and advances of optimization. We will demonstrate the flexibility that mathematical programming and deterministic optimization provide, discuss some characteristic applications and finally conclude with a number of suggestions for future research.

## 1.3 Analysis of Gene Expression Data: Tissue Clustering and Classification

### 1.3.1 Clustering and Classification Preliminaries

Let us assume the data describing a particular process are expressed in the form of  $n$ -dimensional feature vectors  $x \in \mathbb{R}^n$ . An important goal of the analysis of such data is to determine an explicit or implicit function that maps the points of the feature vector from the input space to an output space (for example in regression). This mapping has to be derived based on a finite number of data, thus assuming that a proper sampling of the space has been performed. If the predicted quantity is categorical and if we know the

value that corresponds to each elements of the training set then the question becomes how to identify the mapping that connects the feature vector and the corresponding categorical value (class). This problem is known as the classification problem (supervised learning). If the class assignment is not known and we seek to: (a) identify whether small, yet unknown, number of classes exist; (b) define the mapping assigning the features to classes then we have clustering problem (unsupervised learning).

Although numerous methods exist for addressing these problems they will not be reviewed here. Nice reviews of classification that were recently presented include the papers by (Grossman et al. [33]; Hand et al. [35]). In this short introduction we will concentrate on solution methodologies based on reformulating the clustering and classification questions as optimization problems.

### **Tissue Classification**

Developing specific therapies to pathogenetically distinct tumor types is important for cancer treatment, because they maximize efficacy and minimize toxicity. Thus, precisely classifying tumors is of critical importance to cancer diagnosis and treatment. Diagnostic pathology has traditionally relied on macro- and microscopic histology and tumor morphology as the basis for tumor classification. Current classification frameworks, however, are unable to discriminate among tumors with similar histopathologic features, which vary in clinical course and in response to treatment. Recently, there is increasing interest in changing the basis of tumor classification from morphologic to molecular. In the past decade, microarray technologies have been developed that can simultaneously assess the level of expression of thousands of genes. Several studies have used microarrays to analyze gene expression in colon, breast, and other tumors and have demonstrated the potential power of expression profiling for classifying tumors. Gene expression profiles may offer more information than classic morphology and provide an alternative to morphology-based tumor classification systems (Zhang et al. [60]).

### **Mathematical Programming Formulations**

Classification and clustering, and for that matter most of the data mining tasks, are fundamentally optimization problems. Mathematical programming methodologies formalize the problem definition and make use of recent advances in optimization theory and applications for the efficient solution of the corresponding formulations. In fact, mathematical programming approaches, particularly linear programming, have long been used in data mining tasks. The pioneering work of Mangasarian [43, 44] demonstrated how to formulate the problem of constructing planes to separate linearly separable sets of points. In addition, early work by Freed and Glover [20, 21, 22], Gehrlein [26], Glover et al. [31], and Glover [30] skillfully discussed various aspects

of discriminant analysis from the point of view optimization. A more recent excellent review was presented in Stam [53], highlighting numerous developments that defined the field of applications of mathematical programming to statistical classification. It should be pointed out that one of the major advantages of a formulation based on mathematical programming is the ease in incorporating explicit problem specific constraints whose incorporation in classical statistical approaches is not evident in general.

Let us consider a two-class problem in which the sample points belong to either one of sets with their point coordinates be denoted by  $A$  and  $B$  respectively<sup>1</sup>. As discussed earlier a discriminant function can be derived based on a hyperplane of the form

$$P = \{x \in \mathbb{R}^n | x^\top \omega = \gamma\}.$$

The normal to this plane is

$$\frac{|\gamma|}{\|\omega\|_2}.$$

The classification problem thus becomes how to determine  $g$  and  $w$  such that the separating hyperplane  $P$  defines two open half spaces

$$\{x \in \mathbb{R}^n | x^\top \omega < \gamma\} \text{ and } \{x \in \mathbb{R}^n | x^\top \omega > \gamma\}$$

containing mostly points in  $A$  and  $B$  respectively. Unless the problem is linearly separable the hyperplane can only be derived within a certain error. Minimization of the average violations provides a possible approximation of the separating hyperplane

$$\min_{\omega, \gamma} \frac{1}{m} \|-A\omega + e\gamma + e\| + \frac{1}{k} \|-B\omega + e\gamma + e\|$$

where  $m$  and  $k$  denote the number of samples belonging to classes  $A$  and  $B$  respectively. Bradley et al. [5] discusses various implementations including a particularly effective robust linear programming reformulation suitable for large-scale problems:

$$\min_{\omega, \gamma, y, z} \frac{1}{m} e^\top y + \frac{1}{k} e^\top z$$

subject to

$$\begin{aligned} -A\omega + e\gamma + e &\leq y \\ B\omega - e\gamma + e &\leq z \\ y, z &\geq 0. \end{aligned}$$

Fung et al. [23] demonstrated how to extend the aforementioned formalism to account for non-linear kernel functions that generate non-linear optimal separating surfaces.

---

<sup>1</sup>For simplicity we use the symbols  $A$  and  $B$  to denote both the classes and the matrices containing the coordinates.

While the approaches just described aim at minimizing an error in separating the given data, Support Vector Machines (SVM, Vapnik [57]) incorporate also the structured risk minimization, which minimizes an upper bound of the generalization error. In fact a very interesting analysis on the learning stability characteristics of SVMs, in dealing with uncertainty, is demonstrated by Bousquet and Elisseeff [1]. The general idea behind SVM is illustrated by considering the case where a linear separating surface is to be generated. In that case, SVMs determine, among the infinite number of possible planes separating the two classes, the one that also maximizes the margin separating the two classes.

SVMs are based on an analysis of the general problem of learning the classification boundary between positive and negative samples. This is a particular case of the problem of approximating a multivariate function from sparse data. Regularization theory is a classical approach to solving it by formulating the approximation problem as a variational optimization problem of finding the function  $f$  that minimizes the functional

$$\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(x_i)) + \lambda \|f\|_2$$

where  $\ell$  is the number of training samples,  $V(\cdot)$  is the loss function, and  $\|\cdot\|_2$  a suitable norm. In order to derive a linear separating surface between the two classes the above-mentioned problem is equivalent to the solution of the following optimization problem (Cortes and Vapnik [11]):

$$\min_{w,b} \frac{1}{2} w^\top w + C \sum_{i=1}^{\ell} \xi_i$$

subject to

$$\begin{aligned} y_i(wx_i + b) &\geq 1 - \xi_i & i = 1, \dots, \ell \\ \xi_i &\geq 0 & i = 1, \dots, \ell. \end{aligned}$$

In this formulation  $y_i$  denotes the class of sample  $i$  and it is either  $+1$  or  $-1$ . The solution to this problem not only minimizes the misclassifications (second part of the objective) but also identifies the hyperplane, with normal vector  $w$ , that provides the maximum margin between the two classes.

In general however, the separating surface will be nonlinear. In this case, we have to think of a non-linear projection of the original data for which we seek a linear separating surface. In that case, the linear separating surface in the projected feature space will correspond to a non-linear separating surface in the original space. In that case, we can write the following optimization:

$$\min_{w,b} \frac{1}{2} w^\top w + C \sum_{i=1}^{\ell} \xi_i$$

subject to

$$\begin{aligned} y_i (w\phi(x_i) + b) &\geq 1 - \xi_i & i = 1, \dots, \ell \\ \xi_i &\geq 0 & i = 1, \dots, \ell. \end{aligned}$$

The functional  $\phi(\cdot)$  defines the nature of the nonlinear kernel.

Support Vector Machines (SVM) have been applied with great success in clustering and classification problems in microarray experiments (see Brown et al. [6], Furey et al. [24], Guyon et al. [34], Rifkin et al. [48]). It will be shown later that analysis of the coefficients of the separating hyperplanes, of non-linear kernels, can provide some indications as to which features are more significant. Therefore, a byproduct of clustering and classification analysis within such an optimization framework will also be feature (gene) selection.

### Multi-class SVMs

The solution to binary classification problems using SVM has been well developed, tested and documented. However, extending the method to multi-class problems remains an open research issue. The standard approach, within an SVM framework, is to treat the multi-class problem as a collection of two-class (binary) classification problems. Recently, however, multi-class methods considering a much larger problem encompassing all classes at once have been proposed. The drawback of course is the requirement for the solution of a much larger problem. Recently (Hsu and Lin [37] and Nguyen and Rajapakse [47]) discuss a number of alternatives for the development of SVM-based multi-class classifiers.

#### *One-against-all (OAA) classifier*

This method constructs  $k$  SVM models where  $k$  is the number of classes. The  $j^{\text{th}}$  SVM is trained to classify the members of the  $j^{\text{th}}$  class, assuming to have positive labels, against the samples of all the other classes, which are assumed to have negative labels. Therefore, given  $\ell$  training data in the form  $(x_1, y_1), \dots, (x_\ell, y_\ell)$  where  $x_i \in \mathbb{R}^n$  and  $y_i \in \{1, \dots, k\}$  ( $i = 1, \dots, \ell$ ), the  $j^{\text{th}}$  SVM solves the following problem:

$$\min_{w^j, b^j, \xi^j} \frac{1}{2} (w^j)^\top w^j + C \sum_{i=1}^{\ell} \xi_i^j$$

subject to

$$\begin{aligned} (w^j)^\top \phi(x_i) + b^j &\geq 1 - \xi_i^j & i = 1, \dots, \ell : y_i = j \\ (w^j)^\top \phi(x_i) + b^j &\geq -1 + \xi_i^j & i = 1, \dots, \ell : y_i \neq j \\ \xi_i^j &\geq 0 & i = 1, \dots, \ell. \end{aligned}$$

Minimizing the first term in the objective function,  $\frac{1}{2}(w^j)^\top w^j$ , means that large values of the margin between the two groups of data,  $2/\|w^j\|$ , are favored. The second term in the objective function,  $\sum_{i=1}^{\ell} \xi_i^j$ , favors a reduction in the number of training errors for the case where the problem is not linearly separable. Solving this problem for  $j = 1, \dots, k$  generates  $k$  decision functions:

$$(w^j)^\top \phi(x) + b^j \quad (j = 1, \dots, k).$$

Sample  $x$  belongs to the class which has the largest value of the decision function:

$$\text{class of } x = \arg \max_{j=1, \dots, k} [(w^j)^\top \phi(x) + b^j].$$

*One-against-one (OAO) classifier*

This method constructs  $k(k-1)/2$  classifiers each of which is trained on data from two classes,  $j$  and  $j'$  ( $j, j' = 1, \dots, k, j' > j$ ):

$$\min_{w^{jj'}, b^{jj'}, \xi^{jj'}} \frac{1}{2} (w^{jj'})^\top w^{jj'} + C \sum_{i=1}^{\ell} \xi_i^{jj'}$$

subject to

$$\begin{aligned} (w^{jj'})^\top \phi(x_i) + b^j &\geq 1 - \xi_i^{jj'} & i = 1, \dots, \ell : y_i = j \\ (w^{jj'})^\top \phi(x_i) + b^j &\geq -1 + \xi_i^{jj'} & i = 1, \dots, \ell : y_i \neq j \\ \xi_i^{jj'} &\geq 0 & i = 1, \dots, \ell. \end{aligned}$$

Feature testing based on binary classifiers is not trivial. However, a standard technique is based on majority voting: weighted sum of the outputs of all pair wise classifiers defines the predicted class. A particular implementation of the OAO classifier prediction uses the concept of *Directed Acyclic Graphs* (DAG). Each node is a classifier between two classes. Given a test sample  $x$  and starting at the root node, the binary decision function is evaluated. Then it moves to either the left or the right of the tree depending on the output value.

Weston and Watkins [58] proposed the construction of a likewise linear separation of the  $k$  classes in a single optimization formulation. The original formulation is generalized as follows:

$$\min_{w, b, \xi} \frac{1}{2} \sum_{j=1}^k (w^j)^\top w^j + C \sum_{i=1}^{\ell} \sum_{j=1, j \neq y_i}^k \xi_i^j$$

subject to

$$\begin{aligned} w^{y_i} x_i + b^{y_i} &\geq w^j x_i + b^j + 2 - \xi_i^j & i = 1, \dots, \ell; j = 1, \dots, k; y_i \neq j \\ \xi_i^j &\geq 0 & i = 1, \dots, \ell; j = 1, \dots, k; y_i \neq k. \end{aligned}$$

Once again we assume the existence of  $k$  classes and  $\ell$  objects, and  $y_i$  is an integer indicating the class of object  $i$ . Effectively, the method is a generalization of the OAA approach where the classifiers are estimated simultaneously through the solution of a larger optimization problem. In this case, the discriminating function becomes  $\arg \max_{j=1,\dots,k} (w^j x + b^j)$ . Similar in spirit is the formulation proposed by Crammer and Singer [12]. The formulation is similar to the one proposed by Weston and Watkins [58] with the only difference being that the constraints are defined such that a smaller number of slack variables is required.

### Classification of Microarray Data Using SVM

SVM are becoming one of the favorite classification methods for the classification of microarray data primarily due to their sound mathematical foundation. In this section we will outline just a few illustrative examples. The first application aims at classifying cancerous cells based on the measurement of expression values, whereas the second application aims at functionally classifying genes.

#### *Molecular cancer classification*

Modern cancer treatments rely upon macroscopic examination to classify tumors according to anatomical site of origin. DNA microarrays generate information potentially able to formulate molecular-based predictors circumventing the subjectivity associated with the examination of macroscopic characteristics. Rifkin et al. [48] present a computational method, based on SVM, aiming at classifying tumor data in an attempt to derive a general, multi-class molecular-based cancer classification based solely on gene expression data. The case study concerned the analysis of 198 samples from 14 different cancer types, using microarray data recording the activity (expression) levels of 16,063 probes. Both the OAA and OAO approaches were computationally evaluated in terms of their ability to correctly predict unknown samples. This work demonstrated the ability of SVM to effectively and efficiently classify large microarray data sets in computationally reasonable times. In a somewhat similar study, Williams et al. [59] evaluate the ability of SVM to develop prognostic classification tools for relapsing tumor.

#### *Gene functionality classification*

Brown et al. [6] introduced a method of functionally classifying genes by using gene expression data from DNA microarray experiments based on SVM. The approach is motivated by the realization that genes of similar functionality yield similar expression patterns in microarray experiments. As data from such experiments begin to accumulate in increasing rates, it will become essential to have means for extracting biological significance and using data to assign functions to genes. The authors experimented with a number

of nonlinear kernels, including a dot product based measuring the similarity between two gene expression vectors  $K(X, Y) = X \cdot Y$  as well as various  $d$ -fold generalizations of the form  $K(X, Y) = (X \cdot Y + 1)^d$ , and a Gaussian kernel  $K(X, Y) = \exp(-\|X - Y\|^2 / (2\alpha^2))$ . The study considered 2,467 yeast genes for which functional annotation was available. SVM were trained to recognize six functional families: tricarboxylic acid (TCA) cycle, respiration, cytoplasmic ribosomes, proteasome, histones and helix-turn-helix proteins. The computational evaluation of the SVM was based on a three-way cross validation, repeated a number of times. SVMs were compared with other standard supervised learning techniques, including Parzen windows, Fisher's linear discriminant analysis and decision trees (MOC1 and C4.5), and were found to outperform all of them providing superior performance.

### 1.3.2 Feature Selection Preliminaries

Machine learning algorithms are known to be prone to deteriorating performance when faced with many irrelevant or correlated features (see Kohavi and John [40]). A universal, therefore, problem is to decide on which aspects, i.e., features, of a problem are relevant. Narendra and Fukunaga [46] 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 [40] examines a number of issues associated with the problem of feature selection. More recently, Liu and Motoda [42] also present ideas related to the coupling of information theory and feature selection.

Feature selection is a very healthy and vibrant area of research in the machine learning community and has gained increased significance with the recent advances in functional genomics that resulted in the creation of very high-dimensional feature sets. A number of recent publications (Golub et al. [32], Chilingaryan et al. [10], Szabo et al. [55], Dettling and Buhlmann [14]) have devised various approaches for extracting critical, differentially expressed genes in a systematic manner. The advantages of multivariate methods are that (a) they attempt to take into account collaborative effects of gene expression activities (b) they do not simply characterize genes based on arbitrary  $n$ -fold increased/decreased activities.

### Feature Selection in Almost Empty Spaces

A fundamental problem in machine learning is the development of accurate classifiers in sparsely populated datasets, i.e., *almost empty spaces* (see Duin [18]). As noted earlier the 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 [38]. The inability of sparse data to properly capture the complexity of a classification problem was also analyzed by Ho [36]. A nice discussion of the impact of the small

sample size problem in array expression data is presented in Dougherty [15]. 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. 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.

### Gene Selection using Support Vector Machines

Reducing the number of noisy measured variables reduces potential noise, hence avoids pointless over-fitting. Selecting the optimal number of features is a complicated task: too few genes will not discriminate or predict, too many genes might be introducing noise to the model rather than information. Therefore, the identification of informative genes is a significant component of an integrated computer assisted analysis of array experiments. However, in current practice the identification of such a critical sub-set of genes whose expression is informative is accomplished as a by-product of some other activity. For instance by analyzing patterns in “heat maps” in hierarchical clustering, the loadings of singular vectors, or by assessing the ability of certain genes to maximize the separability between classes. In most cases the question of identifying differentially expressed genes is restated as a hypothesis-testing problem in which the null hypothesis of no association between expression levels and responses of interest is tested (see Dudoit et al. [17]).

Support vector machines (SVM) are powerful classifiers based on regularization techniques for regression (see Vapnik [57]). Guyon et al. [34] discuss a recursive forward selection procedure for ranking features in gene expression experiments. Since the method, in general, attempts to identify a surface separating different classes, the assumption is that the weights of the feature in the decision function should also serve to quantify the importance of each feature. Specifically, Guyon et al. [34] follow the formalism of Cortes and Vapnik [11] in which the following problem is considered. Given a set of training examples  $\{x_k\}$ ,  $x_k \in \mathbb{R}^n$  and class labels for each example  $\{y_k\}$ , defined as either +1 or -1, a separating surface is defined as the solution of an optimization problem as defined earlier. The hyperplane  $D = w \cdot x + b = 0$  is the one that separates the training examples belonging to the two classes with a maximal margin. A metric for the ranking of the features is based on the quantity  $w_i^2$ . Guyon et al. [34] developed a recursive feature elimination procedure, which successively ranked and eliminated features and demonstrated the ability of the SVM-based procedure to extract reduced sets of biologically relevant genes. The general observation was that the quality of the SVM classifier improves once irrelevant features are removed. Alternatively, Bradley and Mangasarian [4] presented a variant of the basic SVM which augments the objective by

the addition of the term  $\lambda w^\top \cdot w/2$  which appropriately weights the scarcity of the vector defining the separating hyperplane. They also discuss possible reformulations of this formulation that render the problem one of minimizing a concave objective subject to linear constraints. Despite the fact that the problem is nonconvex, it can be efficiently solved. The issues of non-convexity and global optimality will be revisited later.

### 1.3.3 Simultaneous Gene Selection and Tissue Classification

A mixed-integer linear formulation was recently proposed by Sun and Xiong [54] and will be used for the purposes of our discussion. Feature selection is always considered within the framework of a given analysis. This could be model development/fitting, classification, clustering etc. In other words we want to extract the minimum number of required independent variables necessary to perform a particular task. Therefore, an objective measuring the “goodness of fit” will be required. The parameters associated with the model naturally define a continuous optimization problem. The notion of selection a sub-set of variables, out of superset of possible alternatives, naturally lends itself to a combinatorial (integer) optimization problem. Therefore, depending on the model used to describe the data the problem of feature selection will end up being a mixed integer (non) linear optimization problem. Furthermore, this problem is a multi-criteria optimization since one wishes to simultaneously minimize the model error and the number of features used. Sun and Xiong [54] propose the use of a linear discriminator, similar to a Support Vector Machine to be discussed later. Let  $m$  denote the number of observations for a two-class problem such that  $k$  and  $\ell$  denote the number of samples in each class (for example number of benign and cancerous cells respectively). We also denote as  $I_1$  and  $I_2$  the indices of the corresponding samples and  $I = I_1 \cup I_2$  denotes the entire set of samples. Finally, the set  $J$  denotes the set of all genes recorded in the observations and  $J' \subset J$  denotes the set of genes (features) that are required to develop an accurate model. The expression data are presented in the form  $x_{ij}$ ,  $i = 1, \dots, I$ ,  $j = 1, \dots, J$ . A linear classifier is constructed as:

$$\begin{aligned} \beta_0 + \sum_{j \in J} \beta_j x_{ij} &< 0 & i \in I_1 \\ \beta_0 + \sum_{j \in J} \beta_j x_{ij} &> 0 & i \in I_2. \end{aligned}$$

However, because the observations are not, in general, perfectly separable by a linear model a goal programming formulation can be proposed whose goal is to estimate the coefficients that minimize the deviations from the classifier model. That is:

$$\min \sum_{i \in I_1} d_i^1 + \sum_{i \in I_2} d_i^2$$

subject to

$$\begin{aligned}
\beta_0 + \sum_{j \in J'} \beta_j x_{ij} - d_i^1 + d_i^2 &= -\delta & i \in I_1 \\
\beta_0 + \sum_{j \in J'} \beta_j x_{ij} - d_i^1 + d_i^2 &= \delta & i \in I_2 \\
\beta_j &\in \mathbb{R} & j \in J \cup \{0\} \\
d_i^1, d_i^2 &\in \mathbb{R}^+ & i \in I_1 \cup I_2
\end{aligned}$$

where  $\delta$  is a small constant. It can either be fixed based on user preferences or be added to the objective to be minimized. In order to minimize the number of variables used in the classifier, hence extract the most relevant features for the specific linear model, binary variables need to be introduced to define whether a particular variable is used in the model or not. Therefore:

$$y_j = \begin{cases} 1 & j \in J' \\ 0 & j \notin J' \end{cases}$$

The number of “active” genes can therefore be constrained (that is introduced parametrically in the formulation in order to avoid the solution of a multi-criteria optimization problem. According to the e-constraint method one additional constraint of the form

$$\sum_{j \in J'} y_j \leq \epsilon$$

is introduced. The complete MIP formulation thus becomes:

$$\min \sum_{i \in I_1} d_i^1 + \sum_{i \in I_2} d_i^2$$

subject to

$$\begin{aligned}
\beta_0 + \sum_{j \in J'} \beta_j x_{ij} - d_i^1 + d_i^2 &= -\delta & i \in I_1 \\
\beta_0 + \sum_{j \in J'} \beta_j x_{ij} - d_i^1 + d_i^2 &= \delta & i \in I_2 \\
\sum_{j \in J'} y_j &\leq \epsilon \\
\beta_j &\leq M y_j \\
-\beta_j &\leq M y_j \\
\beta_j &\in \mathbb{R} & j \in J \cup \{0\} \\
d_i^1, d_i^2 &\in \mathbb{R}^+ & i \in I_1 \cup I_2 \\
y_j &\in \{0, 1\}.
\end{aligned}$$

## 1.4 Inferring Regulatory Networks

### 1.4.1 Mixed Integer Formulations

It would have been misleading to assume that gene expression experiments define static and time-independent observations. Temporal, i.e., dynamic, measurements of gene expression activities exhibit the wealth of complexity characterizing the genomic response to external stimuli. A complete understanding of the organization and dynamics of gene regulatory networks is an essential first step towards realizing the goal of deciphering the complex regulation underlying gene expression (see Bower and Bolouri [2], Dasika et al. [13]).

Unlike the preceding discussion the expression level of a gene is now considered to be a function of time,  $Z_i(t)$ . The expression of any given gene  $i$  is however regulated by the expression of some other gene  $j$  with an effective delay  $\tau$ . From a biological point of view, the time delay in gene regulation characterizes the various underlying processes such as transcription and translation introduced earlier in this chapter. The strength of the time regulation is denoted by  $w_{ij\tau}$ . The sign denotes either activation or inhibition of expression. In order to derive biologically relevant activation/inhibition relations logical constraints are imposed to denote the existence of these restrictions. Specifically:

$$Y_{ij\tau} = \begin{cases} 1 & \text{if gene } j \text{ regulates gene } i \text{ with time delay } \tau \\ 0 & \text{otherwise.} \end{cases}$$

Dasika et al. [13] derived the following optimization problem to estimate the potential connectivity and interaction matrix for a given set of temporal gene expression experiments (expression on  $N$  genes measured at  $T$  time points):

$$\min \frac{1}{NT} \sum_{i=1}^N \sum_{j=1}^N [e_i^+(t) - e_i^-(t)]$$

subject to

$$\begin{aligned} \dot{Z}_i(t) - \sum_{\tau=0}^{\tau_{\max}} \sum_{j=1}^N \omega_{ij\tau} Z_j(t-\tau) &= [e_i^+(t) - e_i^-(t)] & i = 1, \dots, N; t = 1, \dots, T \\ \omega_{ij\tau} &\geq \Omega_{ji}^{\min} Y_{ji\tau} & i, j = 1, \dots, N; t = 1, \dots, \tau_{\max} \\ \omega_{ij\tau} &\leq \Omega_{ji}^{\max} & i, j = 1, \dots, N; \tau = 1, \dots, \tau_{\max} \\ \sum_{\tau=0}^{\tau_{\max}} Y_{ji\tau} &\leq 1 & i, j = 1, \dots, N \\ \sum_{\tau=0}^{\tau_{\max}} \sum_{j=1}^N Y_{ji\tau} &\leq N_i & i = 1, \dots, N \\ Y_{ji\tau} &\in \{0, 1\} & i, j = 1, \dots, N; \tau = 1, \dots, \tau_{\max} \\ e_i^+(t), e_i^-(t) &\in \mathbb{R}^+ & i = 1, \dots, N; t = 1, \dots, T. \end{aligned}$$

$N_i$  denotes the maximum number of regulatory inputs for gene  $i$ ,  $e_i^\pm$  denotes positive and negative error variables respectively expressing the deviation from the experimentally measured gene expression values,  $\tau_{\max}$  denotes the maximum allowed time delay in the model and  $\Omega_{ji}^{\max}$  denote maximum values for the regulatory coefficients. Dasika et al. [13] demonstrate an effective solution of the proposed formulation based on a sequential bound relaxation scheme. This work demonstrates nicely how a mathematical programming formalism can assist in the analysis of temporal data which present a significant increase in problem complexity compared to the time-independent data discussed earlier.

#### 1.4.2 Multicriteria Optimization for Generic Network Modeling

Approaches like the one described in the previous section, attempt to reverse engineer genetic networks from microarray data. A major problem, however, is how to reliably find interactions when faced with a relatively small number of arrays compared with data (small sample size problem discussed earlier). To address this dimensionality problem, prior biological knowledge needs to be incorporated, in the form of constraints, about the genetic networks. This can be modeled in terms of limited connectivity, redundancy, stability and robustness. Recently, van Someren et al. [56] presented a multi-objective formulation to address these issues. The problem addressed concerns the definition of appropriate genetic interactions from a set of temporal gene expression data. Specifically, we are given a set  $g_i(t)$  representing the expression level of gene  $i$  at time point  $t$  and let  $N$  genes be measured at each time point. The expression state of the organism is thus defined as  $g(t) = [g_1(t), \dots, g_N(t)]^\top$ . The concatenated expression levels at each time  $t$  are defined as  $x^q = g(t)$ . Van Someren et al. [56] assumed the simplest dynamic relation for their model, i.e., linear. That is the state of the system at  $t + 1$  is a linear function of the state of the system at time  $t$ :  $x^{q+1} = W \cdot x^q$ . The matrix of interactions  $W$  is termed the gene regulation matrix (GRM). As previously stated a nonzero entry  $w_{ij}$  denotes the existence of a regulatory connection between genes  $i$  and  $j$ , the sign defines a activating action ( $> 0$ ) or and inhibiting action ( $< 0$ ). In order to learn the gene regulation matrix, we simply require that the predicted states of gene  $i$  are close as possible to the target (measured) states. The corresponding error is represented by the mean square error criterion defined as:

$$f^{\text{MSE}}(w_i) = \frac{1}{Q-1} \sum_{q=1}^{Q-1} (w_i \cdot x^q - x^{q+1})^2.$$

The authors model two biologically relevant constrains. The first one deals with the knowledge that a particular node is influenced only by a limited number of other genes. The connectivity is defined as the number of non-zero weights in the  $W$  matrix

$$f^c(w_i) = \sum_{j=1}^N c_{ij}$$

where

$$c_{ij} = \begin{cases} 1 & \text{if } w_{ij} \neq 0 \\ 0 & \text{if } w_{ij} = 0. \end{cases}$$

The second constraint deals with the realization that gene networks are robust in the respect to noise. The robustness is here defined as the inherent ability not to propagate forward in time small perturbations in the current expression state. A metric for the robustness is the first derivative of models' output and it is minimized by minimizing the sum of the squared (or absolute) first derivatives  $f^S(w_i) = \sum_{j=1}^N w_{ij}^2$ . Van Someren et al. [56] demonstrate how the Pareto-front can be generated efficiently in order to balance the requirement for accuracy in the model and robustness and stability in the predictions.

## 1.5 A Final Comment

It is clear from the preceding discussion that *feature selection*, *clustering*, and *classification* are tasks intimately connected. Numerous techniques have been developed that addressed each problem independently. One of the major advantages of mathematical programming (MP) formulations is that they can bring these tasks explicitly together within a similar framework. The goal of this short exposition was not only to show, by example, how some key questions in biology can be advanced by formulating them as MP problems, but also to demonstrate that one of the major advantages of MP-based approaches is the integrated and highly flexible formulations that capitalize on our advanced understanding of large scale mixed integer (non) linear optimization theory. It should be pointed out that a number of other optimization (continuous and mixed-integer) reformulation of data mining have been proposed recently by Glen [27, 28, 29]. We have chosen, however, to focus on methods that have found direct application to microarray expression data and hence left their presentation out of this short review. We do however encourage the interested reader to follow up with such methods because we believe that they will become critical enablers for addressing some of the important open issues such as the ones discussed in the following section.

## 1.6 Research Challenges

Numerous issues can be raised for future research. In fact the advantage of a MP-based formalism is the tremendous flexibility it provides.

*Multi-objective optimization*

Interpretation of biological information needs to tackle multiple simultaneous objectives. In this short review we discussed simultaneous optimization of accuracy and size of classifier (number of features). In clustering applications the number of clusters is yet another level of complexity, hence an additional decision variable. Therefore, multicriteria trade-off curves (Pareto solutions) have to be developed for these high-dimensional mixed integer (non) linear optimization problems.

*Incorporation of biological constraints*

One of the advantages of using mathematical programming techniques is that constraints can be readily accounted for. Thus far microarray analyses approaches treat the array data as raw unconstrained measurements. One of the targets of microarray analysis is to identify potential correlations among the data. However, prior biological knowledge is not taken into account mainly because most data mining methods cannot handle implicit or explicit constraints. Recently Sese et al. [51] demonstrated the need to account for biological driven constraints when clustering expression profiles.

*Large-scale combinatorial optimization*

The development of scalable algorithms is a daunting task in optimization theory. With the recent developments in genomics we should be expecting routinely that the analysis of gene arrays composed of tens of thousands of probes (hence tens of thousands of binary variables in the MIP gene selection formulation). Duarte Silva and Stam [16], Gallagher et al. [25], and Rubin [49] discuss various mixed-integer reformulations to the classification problem. Undoubtedly, the biological sciences will greatly benefit by the anticipated advances in optimization theory and practice when used to target problems such as the ones just described. The recent work of Shioda [52] identified opportunities for successful reformulations of various data mining tasks in the context of linear integer optimization. Busygin et al. [8] present some more recent ideas for addressing the bi-clustering problem as a fractional 0-1 optimization problems. Undoubtedly, integer optimization will play a prominent role in feature algorithmic developments as recent results demonstrate the complementarity of the different methodologies, suggesting that a unified approach may help to uncover complex genetic risk factors not currently discovered with a single method (see Moscato et al. [45]).

*Global optimization*

The development of general non-linear, non-convex separating boundaries naturally leads to requirements of solving large-scale combinatorial non-linear problems to global optimality. Recent advances in the theory and practice of deterministic global optimization are also expected to be critical enablers (see Floudas [19]).

*Multi-class problems*

Most of the recent developments on mathematical programming-driven approaches are based on two-class problems. The simplest multi-class extension is the one-against-all by constructing  $k$  SVM models, where  $k$  is the number of classes. The  $i^{\text{th}}$  SVM classifies the examples of class  $i$  against all the other samples in all other classes. Another alternative builds one-against-one classifiers by building  $k(k-1)/2$  models where each is trained on data from two classes. Hsu and Lin [37] discuss a computational comparison of the models. The emphasis of current research is on novel methods for generating all the decision functions through the solution of a single, but much larger, optimization problem.

*Analyzing almost empty spaces*

The sparseness of the data set is a critical roadblock. Accurate models can be developed using convoluted optimization approaches. However, we would constantly lack appropriately populated datasets in order to achieve a reasonable balance between the thousands of independent variables (genes measured) and necessary measurements (tissue samples) for a robust identification. Information theoretic approaches accounting for complexity (Akaike and Bayesian Information Criteria) should be developed to strike a balance between the complexity and the accuracy of the model so as to avoid pointless over fitting of the sparsely populated datasets.

*Uncertainty considerations*

Noise and uncertainty in the data is a given. Therefore, data mining algorithms in general and mathematical programming formulations in particular have to account for the presence of noise. Issues from robustness and uncertainty propagation have to be incorporated. However, an interesting issue emerges: how do we distinguish between noise and an infrequent, albeit interesting observation? This in fact maybe a question with no answer especially if we consider the implications of sparsely populated data sets

*Mixed integer dynamic optimization*

We demonstrated how researchers begin to explore the dynamic component of the gene expression data. This type of analysis however is expected to be enabled tremendously by upcoming advances in efficient algorithms for addressing large-scale mixed integer dynamic optimization problems. Once the models become non-linear and non-convex the issue of global optimality will once again become pertinent.

*Reformulations*

Undoubtedly some of the most critical advances in the practice of mathematical programming-based methods for the analysis of microarray data in general and data mining in particular, have been the result of fundamental advances in terms of reformulating large scale optimization problems and devising ingenious solutions methodologies. To that effect the pioneering work of Mangasarian [43, 44] deserves particular mention. Stating the data mining tasks as optimization problems is but the beginning. The most appealing characteristic of gene expression analysis is the enormous dimensionality of the resulting optimization problem. High performance computing will without a doubt have a profound effect, however, true advances will be the result of ingenious algorithmic developments. This is a critical step so that rigorous optimization methods become true competitors for the simpler, yet very efficient, statistics-based analysis methods.

*Interpretation and visualization*

The ultimate goal of data mining is the understanding of the data and the development of actionable strategies based on the conclusions. We need to improve not only the interpretation of the derived models but also the knowledge delivery methods based on the derived models. Optimization and mathematical programming need to provide not just the optimal solution but also some way of interpreting the implications of a particular solution including the quantification of potential crucial sensitivities.

**1.7 Acknowledgments**

The author wishes to thank the National Science Foundation (NSF-0519563) and the Environmental Protection Agency (EPA-GAD R 832721-010) for financial support.

**References**

- [1] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(3):499–526, 2002.
- [2] J.M. Bower and H. Bolouri, editors. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, 2004.
- [3] D.D. Bowtell. Options available – from start to finish – for obtaining expression data by microarray. *Nature Genetics*, 21(1 Suppl):25–32, 1999.
- [4] D. Bradley and O.L. Mangasarian. 15th international conference on machine learning (icml’98). In Morgan Kaufman, editor, *Pacific Symposium on Biocomputing*, San Francisco, CA, 1998.

- [5] P.S. Bradley, U.M. Fayyad, and O.L. Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.
- [6] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Walsh Sugnet, T.S. Furey, M. Ares, Jr., and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, 97(1):262–267, 2000.
- [7] P.O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21(1 Suppl):33–37, 1999.
- [8] S. Busygin, O.A. Prokopyev, and P.M. Pardalos. Feature selection for consistent biclustering via fractional 0-1 programming. *Journal of Combinatorial Optimization*, 10(1):7–21, 2005.
- [9] V.G. Cheung, M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs. Making and reading microarrays. *Nature Genetics*, 21(1 Suppl):15–19, 1999.
- [10] A. Chilingaryan, N. Gevorgyan, A. Vardanyan, D. Jones, and A. Szabo. Multivariate approach for selecting sets of differentially expressed genes. *Mathematical Biosciences*, 176(1):59–69, 2002.
- [11] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [12] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47(2-3):201–233, 2002.
- [13] M.S. Dasika, A. Gupta, and C.D. Maranas. A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. In *Pacific Symposium on Biocomputing*, pages 474–485, 2004.
- [14] M. Dettling and P. Buhlmann. Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, 90(1):106–131, 2004.
- [15] E.R. Dougherty. Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, 2(1):28–34, 2001.
- [16] A.P. Duarte Silva and A. Stam. A mixed integer programming algorithm for minimizing the training sample misclassification cost in two-group classification. *Annals of Operations Research*, 74(0):129–157, 1997.
- [17] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, 12(1):111–139, 2002.
- [18] R.P.W. Duin. Classifiers in almost empty spaces. In *15th International Conference on Pattern Recognition (ICPR'00), Volume 2*, 2000.
- [19] C.A. Floudas. *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications*. Oxford University Press, Oxford, U.K., 2000.
- [20] N. Freed and F. Glover. A linear programming approach to the discriminant problem. *Decision Sciences*, 12:68–74, 1981.
- [21] N. Freed and F. Glover. Simple but powerful goal programming for the discriminant problem. *European Journal of Operational Research*, 7:44–60, 1981.
- [22] N. Freed and F. Glover. Evaluating alternative linear programming formulations for the discriminant problem. *Decision Sciences*, 17:151–162, 1986.
- [23] G.M. Fung, O.L. Mangasarian, and A.J. Smola. Minimal kernel classifiers. *Journal of Machine Learning Research*, 3(2):303–321, 2003.

- [24] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
- [25] R.J. Gallagher, E.K. Lee, and D.A. Patterson. Constrained discriminant analysis via 0/1 mixed integer programming. *Annals of Operations Research*, 74(0):65–88, 1997.
- [26] W.V. Gehrlein. General mathematical programming formulations for the statistical classification problem. *Operations Research Letters*, 5(6):299–304, 1986.
- [27] J.J. Glen. Classification accuracy in discriminant analysis: a mixed integer programming approach. *Journal of the Operational Research Society*, 52(3):328–339, 2001.
- [28] J.J. Glen. An iterative mixed integer programming method for classification accuracy maximizing discriminant analysis. *Computers & Operations Research*, 30(2):181–198, 2003.
- [29] J.J. Glen. Mathematical programming models for piecewise-linear discriminant analysis. *Journal of the Operational Research Society*, 56(3):331–341, 2005.
- [30] F. Glover. Improved linear programming models for discriminant analysis. *Decision Sciences*, 21:771–785, 1990.
- [31] F. Glover, S. Keene, and B. Duea. A new class of models for the discriminant problem. *Decision Sciences*, 19:269–280, 1988.
- [32] 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*, 286(5439):531–537, 1999.
- [33] R.L. Grossman, C. Kamath, and V. Kumar. *Data Mining for Scientific and Engineering Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001.
- [34] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.
- [35] D.J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, Cambridge, MA, 2001.
- [36] T.K. Ho. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis & Applications*, 5:102–112, 2002.
- [37] C.W. Hsu and C.J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.
- [38] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.
- [39] F.C. Kafatos. A revolutionary landscape: the restructuring of biology and its convergence with medicine. *Journal of Molecular Biology*, 319(4):861–867, 2002.
- [40] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.
- [41] R.J. Lipshutz, S.P. Fodor, T.R. Gingeras, and D.J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(1 Suppl):20–24, 1999.
- [42] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Oxford University Press, Oxford, U.K., 2000.
- [43] O.L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

- [44] O.L. Mangasarian. Multi-surface method of pattern separation. *IEEE Transactions on Information Theory*, IT-14:801–807, 1968.
- [45] P. Moscato, R. Berretta, M. Hourani, A. Mendes, and C. Cotta. Genes related with Alzheimer’s disease: A comparison of evolutionary search, statistical and integer programming approaches. In F. Rothlauf et al., editor, *Applications of Evolutionary Computing*, pages 84–94, Berlin, Germany, 2005. Springer-Verlag.
- [46] P. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, C-26(9):917–926, 1977.
- [47] M.N. Nguyen and J.C. Rajapakse. Multi-class support vector machines for protein secondary structure prediction. *Genome Informatics*, 14:218–227, 2003.
- [48] R. Rifkin, S. Mukherjee, P. Tamayo, S. Ramaswamy, C.-H. Yeang, M. Angelo, M. Reich, T. Poggio, E.S. Lander, T.R. Golub, and J.P. Mesirov. An analytical method for multiclass molecular cancer classification. *SIAM Review*, 45(4):706–723, 2003.
- [49] P.A. Rubin. Solving mixed integer classification problems by decomposition. *Annals of Operations Research*, 0:51–64, 74.
- [50] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.
- [51] J. Sese, Y. Kurokawa, M. Monden, K. Kato, and S. Morishita. Constrained clusters of gene expression profiles with pathological features. *Bioinformatics*, 20(17):3137–3145, 2004.
- [52] R. Shioda. *Integer Optimization in Data Mining*. Ph.d. thesis, Massachusetts Institute of Technology, Operations Research, 2003.
- [53] A. Stam. Nontraditional approaches to statistical classification: Some perspectives on  $L_p$ -norm methods. *Annals of Operations Research*, 74(0):1–36, 1997.
- [54] M. Sun and M. Xiong. A mathematical programming approach for gene selection and tissue classification. *Bioinformatics*, 19(10):1243–1251, 2003.
- [55] A. Szabo, K. Boucher, W.L. Carroll, L.B. Klebanov, A.D. Tsodikov, and A.Y. Yakovlev. Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*, 176(1):71–98, 2002.
- [56] E.P. van Someren, L.F.A. Wessels, E. Backer, and M.J.T. Reinders. Multi-criterion optimization for genetic network modeling. *Signal Processing*, 83(4):763–775, 2003.
- [57] V.N. Vapnik. *The nature of Statistical Learning*. Springer-Verlag, Berlin, Germany, 1995.
- [58] J. Weston and C. Watkins. Multi-class support vector machines. In *Proceedings of ESANN99*, Brussels, Belgium, 1999. D. Facto Publishers.
- [59] R.D. Williams, S.N. Hing, B.T. Greer, C.C. Whiteford, J.S. Wei, R. Natrajan, A. Kelsey, S. Rogers, C. Campbell, K. Pritchard-Jones, and J. Khan. Prognostic classification of relapsing favorable histology Wilms tumor using cDNA microarray expression profiling and support vector machines. *Genes, Chromosomes & Cancer*, 41(1):65–79, 2004.
- [60] H. Zhang, C.Y. Yu, B. Singer, and M. Xiong. Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 98(12):6730–6735, 2001.