Analysis of Gene Expression Microarrays for Phenotype Classification

Andrea Califano[1], Gustavo Stolovitzky, Yuhai Tu

IBM Computational Biology Center, T.J. Watson Research Center, PO Box 704, Yorktown Heights, NY 10598

{acal, gustavo, yuhai}@us.ibm.com

Abstract

Several microarray technologies that monitor the level of expression of a large number of genes have recently emerged. Given DNA-microarray data for a set of cells characterized by a given phenotype and for a set of control cells, an important problem is to identify "patterns" of gene expression that can be used to predict cell phenotype. The potential number of such patterns is exponential in the number of genes.

In this paper, we propose a solution to this problem based on a supervised learning algorithm, which differs substantially from previous schemes. It couples a complex, non-linear similarity metric, which maximizes the probability of discovering discriminative gene expression patterns, and a pattern discovery algorithm called SPLASH. The latter discovers efficiently and deterministically *all* statistically significant gene expression patterns in the phenotype set. Statistical significance is evaluated based on the probability of a pattern to occur by chance in the control set. Finally, a greedy set covering algorithm is used to select an optimal subset of statistically significant patterns, which form the basis for a standard likelihood ratio classification scheme.

We analyze data from 60 human cancer cell lines using this method, and compare our results with those of other supervised learning schemes. Different phenotypes are studied. These include cancer morphologies (such as melanoma), molecular targets (such as mutations in the p53 gene), and therapeutic targets related to the sensitivity to an anticancer compounds. We also analyze a synthetic data set that shows that this technique is especially well suited for the analysis of sub-phenotype mixtures.

For complex phenotypes, such as p53, our method produces an encouragingly low rate of false positives and false negatives and seems to outperform the others. Similar low rates are reported when predicting the efficacy of experimental anticancer compounds. This counts among the first reported studies where drug efficacy has been successfully predicted from large-scale expression data analysis.

Keywords: Gene Expression Microarrays. Gene Expression analysis. Gene Expression Patterns. Tissue Classification. Phenotype Classification. Clustering.

1. Phone: (914)784-7827. Fax: (914)784-6223.

1 Introduction

Recent advances in DNA microarray technology [1, 2] are for the first time offering us exhaustive snapshots of some of the cell's most intimate genetic mechanisms. These are creating a unique opportunity to improve our knowledge of the cellular machinery. Previous work on DNA microarrays has concentrated primarily on the identification of coregulated genes [3, 4, 5, 6, 7, 8] to decipher the underlying structure of genetic networks and/or the molecular classification of diseased tissues [9, 10, 11].

A few data analysis schemes [10, 11, 12, 13, 14, 15] have been proposed which aim at extracting useful information from microarray data. Two general strategies have been followed, based on either supervised [11, 15] or unsupervised [10, 12, 13, 14] learning algorithms. In these approaches, the quantitative expression of n genes in k samples are considered as either n vectors in k-dimensional space or as k vectors in n-dimensional space. Various metrics [16] measuring distance between these vectors have been proposed.

In the unsupervised learning category, hierarchical clustering [12] has been used to produce a hierarchical dendrogram in which genes with similar expression patterns (according to standard correlation coefficient metric) are adjacent, and adjacency is interpreted as functional similarity. Other cluster analysis approaches [10, 13, 14] attempt to make a partition of the n genes and/or the k samples into homogeneous and well separated groups whose elements are interpreted as functionally related.

Supervised learning methods [11, 15] have been designed to assess whether or not cells belong to a class characterized by a known phenotype, based on the cells' gene expression profiles. This is also known as the cell phenotype prediction problem. Typically, these algorithms are first trained on two example sets: a *positive example set* or *phenotype set*, which contains data for cells characterized by a predefined phenotype, and a *negative example set* or *control set*, which includes data for cells that do not exhibit that phenotype. In [11], a vector of "marker genes," or signature, is used for classification. Marker genes are selected based on the discriminative power of their individual statistics. In [15] it is shown that Support Vector Machine (SVM) based algorithms outperform other standard learning algorithms such as decision trees, Parzen windows and Fisher's linear discriminant.

In this paper we introduce a novel supervised learning algorithm for cell phenotype prediction. While our objec-

tives are similar in spirit to those reported in [9], [10], [11], and [15] our method differs on several counts. First and foremost, rather than relying on a unique best-fit model, which optimally discriminates between the phenotype and control set, we use multiple, optimally discriminative models. As supported by our results, this improves the analysis of complex phenotypes over single model techniques, especially when those phenotypes are mixtures of multiple, simpler sub-phenotypes at the molecular level. Also, as opposed to [11], our analysis selects genes based on their collective discriminative power, rather than on their individual one. In this latter sense our method can be though of as complementary to that of [11].

Our algorithm finds gene groups whose expression is tightly clustered in a subset of the phenotype set and not tightly clustered in any subsets of the control set. (We shall rigorously define what we mean by *tightly* in Section 2.1.) Among these groups, an optimal subset is then used for classification. Given a microarray with N_g genes, there are 2^{N_g} potential such groups.

Our approach has four basic steps.

First, we transform the gene expression axis using a gene-dependent non-linear metric. This improves the chances of finding tight clusters in the phenotype set that are unlikely to occur in the control set (Section 2.1).

Second, rather than exploring all potential gene groups by brute-force, we use an efficient pattern discovery algorithm called SPLASH [17]. This is discussed in Section 2.3. Each discovered pattern is defined by a subset of the genes and by the subset of the phenotype set over which these genes are tightly clustered.

Third, we discard all patterns that are not be statistically significant under a null hypothesis (Section 2.4), which takes into account some statistical properties of the control set, except for the correlations in the expression of different genes. In practice, genes are correlated and some patterns may be deemed statistically significant even though they are not. We shall call these *promiscuous* patterns. Promiscuous patterns, however, are easily identified (see discussion in Section 2.5) and do not create serious consequences in the remaining of the analysis.

Fourth, an optimal set of patterns is chosen among the statistically significant ones using a greedy set covering algorithm [18]. For typical cases, this set is small, consisting of one to three patterns. The combined pattern set is used to build a multivariate probability density model. A "control" probability density model is also built for each gene from the samples in the control set. A standard classification scheme based on the ratio of the two probability densities [19] is then used. This scheme allows us to decide whether or not a previously unseen cell belongs to the phenotype set.

We have applied our supervised learning scheme to the classification of 60 human cancer cell lines [21], from data obtained with Affymetrix HU6800 GeneChips [1]. Cell lines have been analyzed according to a variety of phenotypes. The term "cell phenotype," is generally used in the literature to indicate a common property of a set of cells. For instance, a cancer morphology, such as melanoma, is a typical phenotype. More subtle phenotypes are also possible and useful. In Section 3.1, for instance, we refer to a "p53 phenotype" which identifies cancer cells with mutations in the p53 oncogene. Also, by measuring the drug concentration required to inhibit by 50% the cell line growth, the so-called GI_{50}, it is possible to define a drug-sensitivity phenotype. This can be used to divide cells in two groups: one with cells that are inhibited by low concentrations of the drug (i.e., that are highly sensitive to it) and the other with cells that require high concentrations (i.e., that are resistant to it). A classification method can then be used to predict whether an unknown cell line is likely to be sensitive or resistant to a given drug.

Some complex phenotypes, such as the p53-related one, are likely to be mixtures of simpler unknown sub-phenotypes at the molecular level, each one characterized by a possibly independent pattern. Methods that rely on a single model are likely to perform poorly with these complex cases, as truly there is no single model that describes the entire set. This is verified in Section 3.1, where a systematic comparative analysis of the classification performance of several algorithms, including [11] and [15], has been performed using a standard leave-one-out cross-validation scheme. Results are analyzed based on the sum of false positive and false negative probabilities.

For simple phenotypes, such as a specific cancer morphology, a unique model is sufficient. In this case, performance is similar for all methods. In more complex cases such as with the p53-related phenotype, where multiple models clearly emerge, our technique outperforms the other two. We have also performed tests on synthetic datasets that mimic the statistics of the human cancer cell lines. This is useful in determining the practical limits of the technique, whether promiscuous patterns have a negative impact on classification performance, and whether the technique overfits the data, a typical problem of supervised learning algorithms with many degrees of freedom. Our method is shown to perform well on all counts.

Finally, the analysis of the sensitivity to the drug Chlorambucil [22] shows consistently good results for both our method and SVM. These are among the first results where drug efficacy is predicted based only on large-scale gene expression data from microarrays. In general, results clearly indicated that accurate and sensitive phenotype prediction, from absolute gene expression levels is possible.

The implication of our study is twofold. It could validate the possibility of creating diagnostic tools for the classification and identification of several diseases. It could also help devise new tools to predict which one, among a set of alternative therapies, may have the highest chances of success with a pathology linked to a specific cell phenotype. How-

ever, we must be extremely cautious in extrapolating from the current analysis. The samples we have used are from cell lines rather than patient tissue. As a result, they could be much more homogeneous than real tissue samples. Also, since these are all cancer cells, the set statistics could be quite skewed. Therefore, these results would not be immediately useful as a diagnostic tool. However, we believe this paper constitutes proof of concept that successful phenotype prediction can be accomplished from microarray data using pattern discovery. Given more statistically sound phenotype and control sets, this approach could be used to discover multiple sets of marker genes both for diagnostic and therapeutic purposes.

2 Methods

In the following sections we will (1) describe rigorously the metric used for distance/similarity measure; (2) describe the pattern discovery algorithm; (3) determine the probability of patterns to arise spontaneously in the control set; (4) show how patterns can form the basis of a supervised learning algorithm.

2.1 Metric Definition and Data Normalization

In order to find useful gene expression clusters or patterns, we must first define a metric in the space of gene expression values to best distinguish the "signal" (phenotype set) from the "noise" (control set). In other words, in designing an appropriate metric for our problem, we wish to minimize the probability of discovering patterns that would be likely to occur in the negative example set.

Suppose that the expression level u of the g-th gene, in the control set, is distributed according to a given probability density $P^-_g(u)$. This density is estimated empirically, by using a sum of Gaussian densities centered around each expression measurement in the control set. The standard deviation is a function of *u* and it is also computed empirically from repeatability studies. A sufficient number of samples is required to measure this density with a sufficient degree of accuracy. In Fig.1a a possible shape for P^-_g is plotted along with four expression values from hypothetical phenotype cells *a*, *b*, *c*, and *d*: u_a, u_b, u_c and u_d. Although the Euclidean distance between u_c and u_d is smaller than that between u_a and u_b, the likelihood of getting the former values by chance is higher because they are very close to the maximum of the expression probability density. In other words, if we want to minimize the probability of finding random clusters in the control set, we must choose a metric such that u_c and u_d would be considered further away than u_a and u_b. A natural choice is to renormalize the expression axis so that the distance between two points on the new axis is equal to the integral of the $P^-_g(u)$ in the previous coordinate system. This is accomplished by defining a new variable v obtained by transforming the original variable u with the following (gene specific) non-linear transformation f_g:

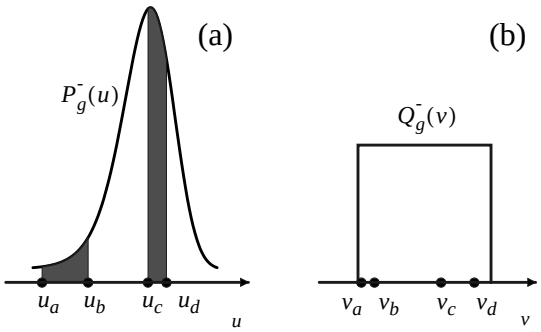

Fig. 1: Control-based data normalization. The shaded area between two points (a) measures the distance between them.

$$v \equiv f_g(u) = \int_{-\infty}^{u} P^-_g(x)dx \qquad (1)$$

In this new variable, the corresponding probability density $Q^-_g(v)$ for the control set is uniformly distributed and normalized in the interval [0,1]. In Fig. 1b, the probability density is plotted together with the transformed values for u_a, u_b, u_c, and u_d. As expected, the Euclidean distance between u_a and u_b is now smaller than that between u_c and u_d, which makes them much more likely candidates for a cluster. Indeed, if $v = f_g(u)$ and $v' = f_g(u')$ are two transformed expression values, we shall take the Euclidean metric in v as our measure of similarity, or distance, between u and u':

$$D(u, u') \equiv |v - v'| = \left| \int_{u}^{u'} P^-_g(x)dx \right| . \qquad (2)$$

The above equation is intuitively equivalent to the definition given before. In other words, the distance between two expression values is chosen to be equal to the integral of the gene expression probability density in the control set between these two values. Since, the number of measurements in the control set falling between u and u' is proportional to the integral in Equation (2), it follows that the more measurements in the control set fall between two values, the further apart they are in the new coordinate system and vice versa.

In the following sections, we will use the transformed gene expression space. One significant advantage is that, since the probability density for all genes in the control set is uniformly distributed in the transformed space over the interval $[0, 1]$, it is now possible to analytically compute the statistics of the patterns discovered in the control set. Based on that, we can assign a statistical significance to patterns discovered in the phenotype.

2.2 Definition of Gene Expression Patterns

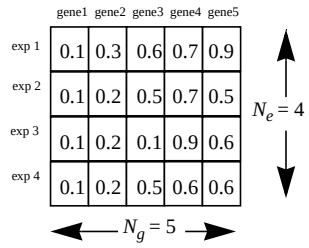

Fig. 2: Example of a gene expression matrix. Entries have values between 0 and 1.

Having defined our measure of similarity, we can now rigorously define what we mean by a gene expression pattern. We shall also define a few terms used consistently throughout this paper.

Gene Expression Matrix: the result of a DNA microarray experiment is a collection of gene expression levels, the level for each gene being roughly proportional to the concentration of the mRNA transcribed from that particular gene in the cell. N_g is the number of gene probes in the microarray. N_e is the number of microarray samples (i.e., experiments or cells). Thus, a set of DNA microarray experiments is conveniently represented by an $N_e \times N_g$ gene expression matrix $\boldsymbol{V} = \{v_{eg}\}$, where *e* is the experiment index and *g* is the gene index. From the last section, we take v_{eg} to be the *transformed* expression level according to Equation (1) of the *g*-th gene in the *e*-th sample. If the set is the control set, then the transformed gene expression values v_{eg} will be approximately uniformly distributed.

Gene vector and Experiment vector: A list of gene ids $G = \{g_1, \ldots, g_k\}$, with $1 \le g_1 < g_2 < \ldots < g_k \le N_g$ is called a gene vector. A list of experiments $E = \{e_1, \ldots, e_j\}$, with $1 \le e_1 < e_2 < \ldots < e_j \le N_e$ is an experiment vector.

δ-valid *jk*-patterns: Let $\boldsymbol{V}$ be a gene expression matrix, then a gene vector *G* and an experiment vector *E* uniquely define a $j \times k$ submatrix $\boldsymbol{V}_{E,G} = \{v_{e_i g_m}\}$ of $\boldsymbol{V}$. Given $\delta > 0$, $\boldsymbol{V}_{G,E}$ is a δ-valid *jk*-pattern if each column is *tightly clustered* in an interval of size up to δ. By this we mean that the maximum and the minimum value of each column must differ by less than δ. The length of the experiment vector *j* is called the *support* of the *jk*-pattern. Intuitively, if δ is small, each gene in a *jk*-pattern is expressed at approximately the same level across all the experiments in the experiment vector. However, because these are transformed values, the actual gene expression interval may be large.

Maximal patterns: A δ-valid *jk*-pattern is maximal if the following two conditions hold: (1) it cannot be extended into a δ-valid *jk'*-pattern, with $k' > k$, by adding genes to its gene vector, and (2) it cannot be extended into a δ-valid *j'k*-pattern, with $j' > j$, by adding experiments to its experiment vector.

Example. Consider the gene expression matrix $\boldsymbol{V}$ of Fig. 2, with $N_g = 5$ genes and $N_e = 4$ experiments. Given a gene vector $G = \{1, 3, 4\}$ and an experiment vector $E = \{1, 2, 4\}$, $\boldsymbol{V}_{G,E}$ is

$$\boldsymbol{V}_{G,E} = \begin{bmatrix} 0.1 & 0.6 & 0.7 \\ 0.1 & 0.5 & 0.7 \\ 0.1 & 0.5 & 0.6 \end{bmatrix}. \quad (3)$$

$\boldsymbol{V}_{G,E}$ is not a (δ=0.05)-valid ($j = 3, k = 3$)-pattern because the values in its 2nd and 3rd column are spread over an interval greater than 0.05. The same pattern, π_1 in Fig. 3, is ($\delta = 0.1$)-valid but not maximal for the matrix of Fig. 2, because adding gene 2 to *G*, produces π_2 which is still δ-valid. Pattern π_2 is maximal because adding any other gene or experiment yields submatrices that are no longer ($\delta = 0.1$)-valid. π_3 shows a ($\delta = 0.1$)-valid ($j = 2, k = 5$)-pattern.

2.3 The Pattern Discovery Algorithm

Full details of the SPLASH algorithm are given in [17]. In that paper, SPLASH was introduced as an algorithm to discover patterns in strings, where all possible relative strings alignment are allowed. Also, a density constraint is introduced to limit the impact of random matches occurring over large distances on the string. For the equivalent association discovery problem, relevant in this context, the approach is analogous as we can imagine each row in the matrix to be equivalent to a string. However, the strings are prealigned in the present case. In addition, the density constraint criteria introduced in [17] is no longer meaningful here, as the first and last genes are as likely to form patterns as two corresponding to contiguous matrix columns.

Using the notation of [17], the canonical *seed set* P_s has a single pattern with no genes, all the rows, and an offset of 0 for each row. The histogram operator T_h is implemented by simply sorting the values in each column and then selecting all subsets of continuous values that are δ-valid. Non maximal subsets that are completely contained within another subset are removed. Each subset is a potential *superpattern* of a maximal pattern. The enumerate operator T_e is then applied iteratively to create all possible maximal combinations of these superpatterns. As a results, all patterns that exist in the data are generated hierarchically by combining together smaller superpatterns, with fewer genes. Non maximal branches are eliminated at each iteration, as soon as their corresponding superpattern arises. This contributes to the efficiency of the algorithm.

2.4 Statistical significance of Patterns in Gene Expression Matrices

When gene expression values are organized in a gene expression matrix, *jk*-patterns may occur for any given value of δ. Can any of these patterns occur merely by chance? In this subsection we address this question by studying the statistics of patterns in any $N_e \times N_g$ matrix, whose elements are statistically independent of each other and have the same proba-

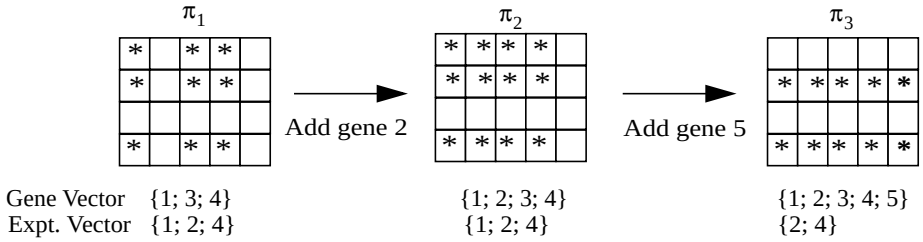

Fig. 3: (see matrix in Fig. 2) Pattern π_1 is extended to pattern π_2 by adding gene 2, with no change in the Exp. vector. π_1 is non-maximal. Pattern π_2 is maximal. Adding any gene to the gene vector of a maximal pattern, e.g., gene 5, produces a new pattern π_3 with a smaller experiment list.

bility distribution as that of the control set in the transformed variable space.

An important observation is in order at this junction: we do not mean that the expression values of different genes in "real-life" gene-expression matrices are independent random variables. Rather, we intend to use such a model as the null hypothesis of our statistical framework precisely to identify any skew or co-regulation in the phenotype set. This null hypothesis definition is based on two assumptions: (a) that the probability densities for the expression levels of each gene are the same as in the control set, and (b) that the gene expression levels in different experiments and/or those of different genes are independently distributed. When discovering patterns in the phenotype set, the statistical relevant patterns will be those for which the null hypothesis is rejected. These are patterns whose constituent genes are either distributed differently in the phenotype set than in the control set, and/or are expressed in a correlated fashion. Both of these features are actually the kind of behavior that we are seeking to differentiate the two sets.

Of course, many genes are not independently distributed in the control set. Therefore patterns may arise that reject the null hypothesis and yet are likely to occur in the control set. We shall call these *promiscuous* patterns. Promiscuous patterns, are easily eliminated in a post-processing phase and do not contribute significantly to remaining analysis. This is verified by experimental results in Section 3.2, where any correlation in the control set is artificially removed. Results for this set are not different from those on the real data, showing that gene correlation in the control set is not an issue in the present context.

Our main result on the statistics of patterns is the following: given $\delta > 0$, an $N_e \times N_g$ gene expression matrix $\boldsymbol{V}$, a *k*-dimensional gene vector *G* and a *j*-dimensional experiment vector *E*, the probability that the submatrix $\boldsymbol{V}_{G,E}$ is a maximal δ-valid *jk*-pattern is[1]:

$$P_\delta(j,k,N_e,N_g) \approx \zeta^k[1-\zeta]^{N_g-k}[1-(1+j^{-1})^k\delta^k]^{N_e-j} \qquad (4)$$
$$\zeta = j\delta^{(j-1)} - (j-1)\delta^j$$

Therefore, the average number of maximal δ-valid *jk*-patterns in $\boldsymbol{V}$ is

$$N_{jk} = N_t P_\delta(j,k,N_e,N_g), \text{ where}$$
$$N_t = \binom{N_g}{k} \times \binom{N_e}{j}. \qquad (5)$$

N_t is the total number of ways in which one can choose a gene and experiment vector. In Equation (4), we make an approximation that is valid when δ is small. This is consistent with the values used in the experimental section, typically $0.05 \le \delta \le 0.15$.

To verify the validity of Equation (5), and also to show the behavior of N_{jk} as a function of *j* and k, we have performed simulations by running pattern discovery on synthetic gene expression matrices, using the statistics of the null hypothesis. Fig. 4 shows an excellent agreement between various theoretical and experimental values of N_{jk} for $N_g = 400$, $N_e = 10$, and $\delta = 0.2$. Similar agreement is observed for other value of the parameters. In Fig. 4, N_{jk} attains its maximum for $k_{peak} \approx N_g[j\delta^{j-1} - (j-1)\delta^j]$.

From the analysis of the variance of the number of *jk*-patterns (not reported here) we have also verified that this is typically very close to the mean number of *jk*-patterns, especially when the latter is small. This result suggests that the distribution of the number of patterns could be well approximated by a Poisson distribution. Indeed the histogram of the number ν of *jk*-patterns is very well fitted by the distribution $N_{jk}^\nu e^{-N_{jk}}/\nu!$ (data not shown).

We can use the previous observations to assess the statistical significance of a pattern in the phenotype set with respect to the randomized control set. Using classical statistics reasoning, we reject maximal δ-valid *jk*-patterns that would be likely to occur in the randomized control set. Under the null hypothesis, the probability p_{jk} that one or more *jk*-patterns occur in the phenotype set is

$$p_{jk} = 1 - e^{-N_{jk}}. \qquad (6)$$

This will be the *p*-value or significance level of our statistical test. Thus, setting a reasonable threshold P_0, we can say that

1. The derivation of this formula is lengthy and will be published elsewhere. The interested reader can request a draft of the derivation to the authors.

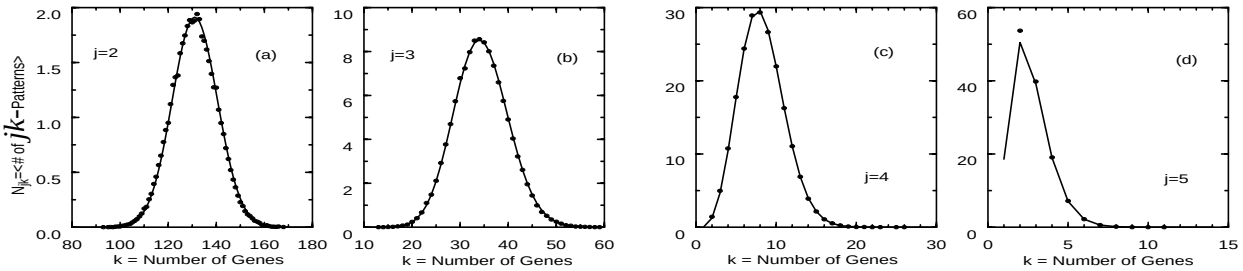

Fig. 4: Average number of **O**-valid *jk*-patterns in randomly generated synthetic data. Continuous curves are the theoretical values. Dots are experimental values obtained with Splash. $N_g = 400$, $N_e = 10$, $\delta = 0.2$.

if we observe one or more *jk*-patterns in the phenotype set with $p_{jk} < P_0$, we reject the null hypothesis and conjecture that such *jk*-patterns could be specific of the cell phenotype under study.

2.5 Classification

Once the statistically significant patterns are found in the phenotype set, we can use them as *classifiers* to build a discriminant function. This function should determine whether or not a previously unseen sample $\mathbf{v} = (v_1, \ldots, v_{N_g})$ belongs to the phenotype or the control set. To this end we build a model for the probability density function of the expression level for each statistically significant pattern π_l of the phenotype set. Each *i*-th gene of π_l contributes with a factor $P_i^+(v)$. The probability density P_i^+ is chosen to be normally distributed with mean equal to the average of the cluster for the *i*-th gene and standard deviation $\sigma_i = \delta/2$. This method is preferred over an empirical derivation from actual measurements because patterns are typically supported by too few samples.

For samples in the control set, the same gene would be expressed according to a different probability density $P_i^-(v)$. The latter can be built empirically because all the samples in the control set can be used. As discussed earlier, we use a sum of Gaussian densities with a mean equal to the value of the sample measurement and a standard deviation derived from repeatability experiments.

On a first order approximation, we shall assume independence between genes and take the multivariate distribution to be equal to the product of the probability densities of the individual genes, both in the phenotype and in the control set. This assumption is necessary because we do not have enough data to construct a realistic multidimensional probability density for either set.

Promiscuous patterns, which arise from correlations in the control set, are likely to play a minor role in the classification as described in the following discussion. To determine if a new microarray sample fits the phenotype model of a *jk*-pattern π_l for the expression values $(v_1, v_2, \ldots, v_k)$ over the *k* genes that constitute π_l, we score it by the logarithm of the ratio of the two probability densities [19]:

$$S_l = \log\left[\frac{P^+(v_1, \ldots, v_k)}{P^-(v_1, \ldots, v_k)}\right] \approx \sum_{i=1}^{k} \log[P_i^+(v_i)] - \sum_{i=1}^{k} \log[P_i^-(v_i)] \quad (7)$$

Using this score, we can easily determine whether promiscuous patterns are contained in the set of statistically significant patterns. Patterns with positive values of S_l for samples taken from the control set are considered promiscuous. Next we assign the statistically significant patterns a *promiscuity* index:

$$\wp_l = \sum_{S_l(\mathbf{v}) > 0} S_l(\mathbf{v} \in \text{Control Set}), \quad (8)$$

where the sum runs over all the samples in the control set for which $S_l > 0$. Patterns whose $S_l < 0$ for all samples in the control set have a promiscuity index of zero. Patterns can now be sorted according to the promiscuity index, with the least promiscuous pattern first.

The next step is to associate each pattern π_l with a coverage set, which includes all the samples in the phenotype set $\mathbf{v}^{ph}$ with a positive score $S_l(\mathbf{v}^{ph}) > 0$.

Finally, an optimal set of patterns is selected using a greedy set covering algorithm [18] to optimally cover the phenotype set. The set covering algorithm tries to use the patterns in sort order according to the promiscuity index: the least promiscuous and most covering pattern is chosen first. The smallest subset of patterns whose coverage sets optimally cover the phenotype set is then used for classification purposes. Therefore, if a non-promiscuous set that optimally covers the phenotype set exists, it will be selected over a promiscuous one.

Typically N_c patterns are selected, where this value ranges between one and three. The score of a previously unclassified sample $\mathbf{v}$ is defined as

$$S(\mathbf{v}) = max(S_l(\mathbf{v}), l = 1, \ldots, N_c). \quad (9)$$

Given a threshold S_c, the sample *fits* the phenotype model only if $S(v) \geq S_c$. The theoretical false positive (FP) and false negative (FN) probabilities can be easily estimated by integrating P^+ and P^- over the region where their ratio is greater or smaller than the threshold. If a single classifier is used, $S_c = 0$ minimizes the sum of false positive and false negative probabilities [19]. In the multivariate model, S_c

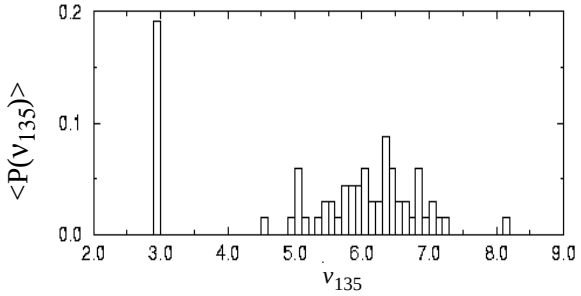

Fig. 5: Histogram of the log expression of gene 135.

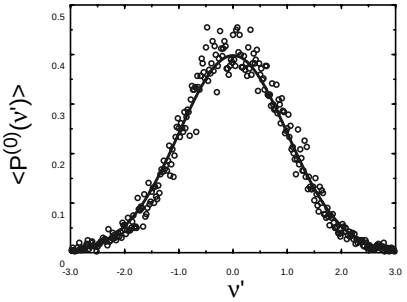

Fig. 6: Normalized probability density: log-expression of the nonbasal activity for all genes.

must be tuned. S_c is an useful tunable parameter practically since different problem requires different balance between FP and FN.

3 Results

An Affymetrix HU6800 GeneChip [1] has been used to monitor the gene expression levels of 6,817 full length human genes in 60 human cancer cell lines [21]. These are organized into a set of panels for leukemia, melanoma, and cancer of the lung, colon, kidney, ovary, and central nervous system. The identity of the genes is not known to the authors. They are therefore identified by a numeric id.

Genes with expression values of 20 or less are considered switched off. From the 6,817 original genes, a subset of 418 was selected by means of a variational filter to eliminate genes that did not change significantly across samples (variational filters were also used in [12, 13].)

The fluorescence intensity ϕ_g of each gene, roughly proportional to the mRNA concentration, appears to be lognormally distributed. The value of variable u is then chosen as $u = \log(\phi_g)$. In Fig. 5, the histogram of a typical gene's expression over the 60 samples is shown. This distribution is clearly bimodal. There is a peak at the basal level, corresponding to the gene being switched off in some experiments. The non-basal expression values, on the other hand, are distributed with a well behaved mean and standard deviation. Thus, we write the corresponding probability density as:

$$P_i^-(u_i) = \alpha_i \delta(u_i - u_0) + (1 - \alpha_i) P_i^{(0)}(u_i) \ , \qquad (10)$$

where α_i is the percentage of expression data being at the basal level u_0, and $P_i^{(0)}(u_i)$ is the density function for the non-basal expression values. For each gene, we determine its basal level α_i, the mean $\bar{u}_i$, and standard deviation σ_i of its non-basal density $P_i^{(0)}(u_i)$. Non-basal values of u_i are normally distributed in accord with the observed lognormal distribution of the ϕ_i. This is shown in Fig. 6, where we plot the combined distribution obtained by shifting and rescaling the nonbasal activity of each gene as $u' = (u_i - \bar{u}_i)/\sigma_i$. The change of variables corresponding to Equation (1) is:

$$v_i = \int_{u_0}^{u_i} P_i^-(x)dx = \alpha_i H(u_i - u_0) + (1 - \alpha_i) \int_{u_0}^{u_i} P_i^{(0)}(x)dx \qquad (11)$$

In Equation (11), $H(x)$ is the Heaviside function, ($H(x \geq 0) = 1$ and $H(x < 0) = 0$), which results from the integration of the delta function in the u_i distribution. For $\alpha_i \neq 0$, different values of u_i can correspond to the same value of v_i. This is not a problem unless α_i is of the same order of δ, in which case we discard the values at u_0 which would not be discriminative.

3.1 Phenotype Analysis

Given a gene expression matrix **V**, with renormalized values v_{lm}, the pattern discovery algorithm SPLASH is used to find all maximal δ-valid *jk*-patterns for $k \geq 4$ and $j \geq 4$. These parameters are chosen because patterns with too few genes are not specific enough, while patterns with too small a support do not characterize a significant consensus in the dataset. δ is chosen between 0.05 and 0.15, depending on the dataset, such that a sufficient number of patterns is discovered, typically on the order of 10 to 50 statistically significant patterns. Larger values of δ are possible but they increase the probability of finding promiscuous pattern and reduce performance. In general, the smaller value of δ one can choose and still discover patterns, the better the results. The threshold of significance P_0 is chosen to be 10^{-4}.

We discuss experimental results on the classification of 7 samples in the melanoma panel, of 17 samples with mutations in the p53 gene, and of 10 samples whose growth is highly inhibited by the drug Chlorambucil. For each experiment, we plot the sum $p_{Tot} = p_{FP} + p_{FN}$ of false positives and false negative probabilities as a function of the matching threshold S_c.

Three methods are studied: our pattern discovery method (PD), the support vector machine (SVM) method of [15]; and the gene by gene method (GBG) of [11]. For each given phenotype, we use its complement in the NCI-60, excluding

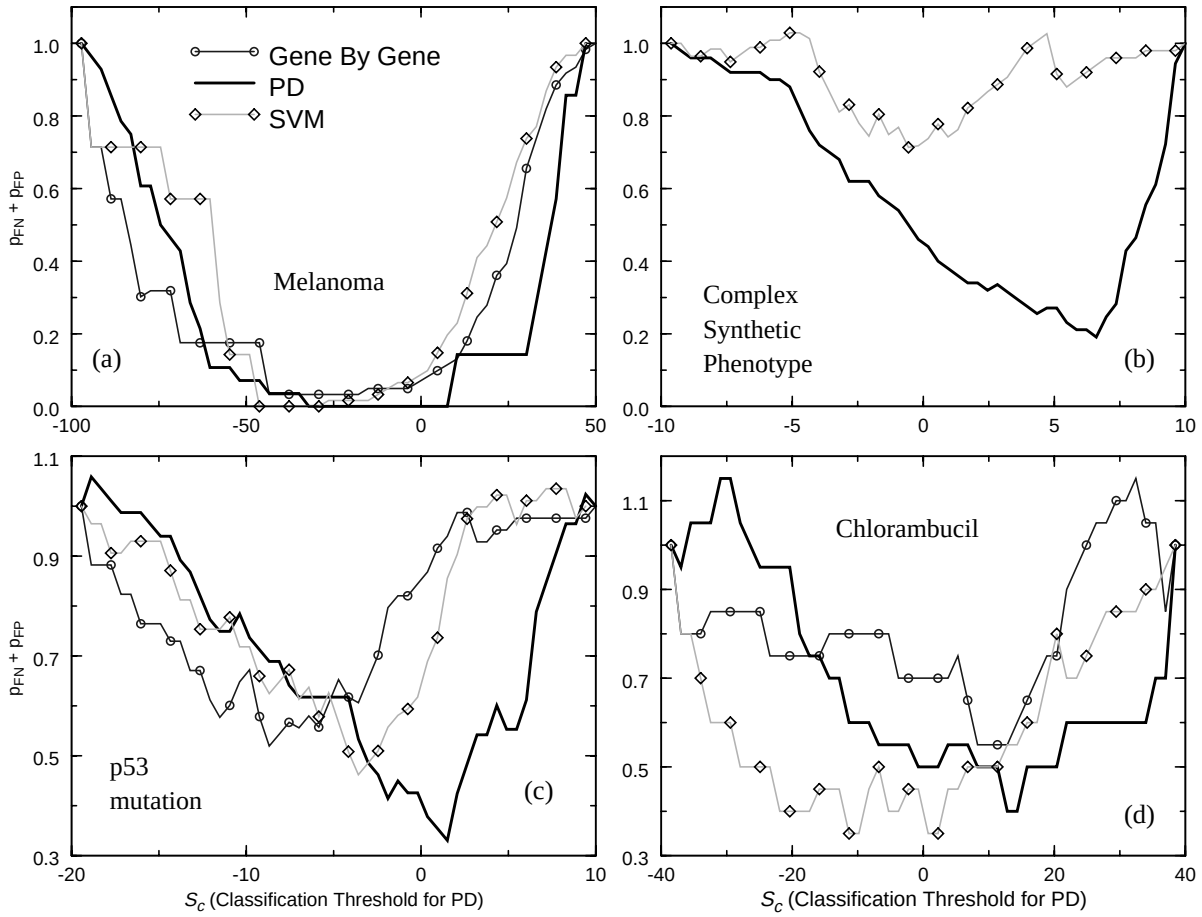

Fig. 7: Classification performance for (a) Melanoma, (b) Complex Synthetic phenotype, (c) p53 and (d) Chlorambucil. The sum of the false positive and false negative probabilities is plotted as a function of the classification score threshold. PD is shown by a thick Solid line, SVM by a dashed line with diamonds, and GBG by a dotted line with circles.

the samples whose phenotype cannot be accurately determined (neutral samples), as the control set.

Given the limited number of samples in the NCI-60 set, false positive and false negative ratios are computed by cross validation. Each sample both in the phenotype and in the control set is removed in turn. The algorithm is trained using the remaining samples, this includes gene axis transformation, pattern discovery, and set covering. Finally, the previously removed sample is classified as described in Section 2.5. When a phenotype set sample is misclassified it is considered a false negative. When a control sample is misclassified it is considered a false positive. All computation times reported are relative to a 450MHZ Pentium II.

Melanoma: The melanoma panel includes 7 samples. There are also 14 neutral samples. These have been selected by biologists prior to this analysis. When the complete set of melanoma samples is used for the training, there is only one statistically significant gene expression pattern that is selected after the set covering phase. Fig. 7a shows the performance of the complete analysis, with $\delta = 0.12$, as described in the previous section. Both SVM and PD show a significant range of the match threshold S_c where both false positive and false negative probabilities are zero. This is considered perfect recognition. The GBG produces results which are very similar, although a fraction less accurate. The time required to classify a sample with the PD method is approximately 10 seconds.

p53 Mutation: A more challenging phenotype is that of 17 samples for cells with mutations in the p53 gene. The corresponding set of cancer morphologies is considerably more complex. It includes 5 melanoma, 3 renal cancer samples, 2 samples for cancer of the central nervous system, leukemia, ovarian cancer, and breast cancer, and 1 sample for colon cancer. As mentioned earlier, this is likely to have several sub-phenotypes at the molecular level. This is confirmed by our analysis, which also highlights a much wider range of variability for the various methods. As shown in Fig. 7c, the GBG method performs quite poorly with a $min(p_{Tot}) = 0.51$ The SVM method improves on that result, bringing that value to about 0.46. Our pattern discovery based approach, with $\delta = 0.12$, has the best result at

$min(p_{tot}) = 0.33$. Three distinct, rather orthogonal patterns are used on average for each sample classification. If only one pattern is allowed, results become close to that of the SVM method. The time required to classify a sample with the PD method is approximately 20 seconds.

Chlorambucil GI_{50}: Some truly interesting phenotypes are associated to the ability of a given drug to inhibit cell growth. These are relevant because many experimental anti-cancer compounds exhibit relatively poor growth inhibition rates across large variety of cancer cells of similar morphology. If one could however correlate the effectiveness of a compound to the much richer space of the gene expression profile of a cell, it could be possible to determine a-priori which cells are most likely to be inhibited by a drug. To test this scenario, we have selected Chlorambucil (NSC 3088) from the NCI anti-cancer database. Since the growth inhibition rate is distributed rather continuously over the entire NCI-60 spectrum, we have split the samples in three groups. The phenotype group, contains the 10 cells that are most inhibited by Chlorambucil. The control group contains the 20 samples whose growth is least inhibited by the compound. The third set of 30 cells is considered neutral. As shown in Fig. 7d, the SVM and PD methods perform similarly, with a slight advantage towards the former. Best values for p_{Tot} are 0.35 and 0.4 respectively. For the PD method, a value of $\delta = 0.12$ is used. The other method cannot to better than 0.55. The time required by the PD method to classify a sample is 10 seconds.

3.2 Synthetic data analysis

Several cross-validation checks using synthetic or randomized data have been performed to validate our approach. Three synthetic data sets are analyzed.

The first test has been designed to evaluate the theoretical performance of the algorithms in the case of phenotype mixtures. A synthetic data model has been generated with the same gene by gene statistics as the control set for the p53 study. A set of 48 control samples have been generated from this model at random. Their gene by gene probability density is virtually identical to that of the real control set in the p53 study. A set of 18 phenotype samples has been synthetically generated. This set consist of three independent sub-phenotypes of 6 samples each. Each sub-phenotype is characterized by 10 marker genes clustered around a tight interval. Remaining genes are modeled as in the control set. Marker genes are different for different sub-phenotype with some overlap. In particular, sub-phenotype 1 and 2 have 6 marker genes in common; sub-phenotype 2 and 3 have 2 marker genes in common. Marker genes are expressed differently than in the control according to the following criteria: 1) they have a different mean located about 0.5σ away from the control mean, 2) they have a smaller standard deviation 0.33σ, where σ is the standard deviation of the same gene in the control set. As shown in Fig. 7b there is a dramatic difference between the performance of the SVM method and that of the PD method. The minimum of p_{Tot} is about 0.7 for the SVM method and 0.19 for the PD method. About 80% of the genes composing the sub-phenotypes are correctly identified by the 3 resulting patterns. Of course, we must be careful not to draw conclusions from this simple minded exercise. Yet, this seems a good indication that the PD method is a suitable choice for the classification of some complex phenotypes that may be mixtures of simpler sub-phenotypes.

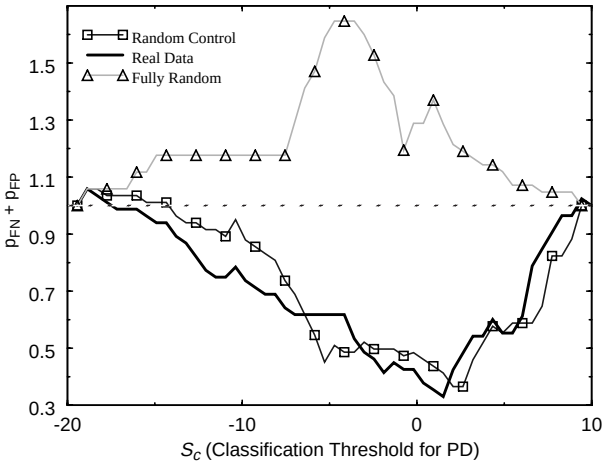

Fig. 8: Performance of the PD method with the actual p53 training set (thick solid line), a randomized control set (squares) and randomized control and phenotype sets (triangles).

The second synthetic data set has been designed to determine whether correlation of genes in the control set is a significant factor and could reduce the performance of the technique. To accomplish this goal, the values of the genes have been randomly permuted only across the control set on a gene by gene basis. This has the effect of leaving the expression probability density for each gene for the control virtually unchanged, while removing any possible correlation between the values of genes in the same sample. Results for the classification are shown in Fig. 8 (squares). There are no major differences with respect to the same curve for the real data of Fig. 7c, and reproduced in Fig. 8 as the thick solid line. This proves that correlation in the genes even though present in the data does not result in classifiers that are highly correlated over the control set.

Finally, we have designed a test to determine whether this approach may suffer from overfitting the data. To that end, classification has been performed using the same data and criteria as for the p53 phenotype study but after the expression values of the individual genes have been randomly permuted across all the samples on a gene by gene basis. The results of the classification are shown as triangles in Fig. 8. As clearly shown, performance is very poor, with a value of p_{Tot} consistently larger than 1 over the entire classification threshold interval. The PD method, as well as the SVM and GBG methods, exhibits no predictability for these

data sets, i.e., the sum of the false positive and the false negative rates is close to 1, as it should be.

4 Summary and Conclusions

Based on a combinatorial multivariate approach, we have developed a systematic framework for cell phenotype classification from gene expression microarray data. We have used SPLASH, a deterministic pattern discovery algorithm, to discover all gene expression patterns on a given data set.

We have then evaluated analytically the statistical significance the patterns. The set of statistically significant patterns form the basis our classification scheme. A scoring system, based on the ratio of the probabilities between the phenotype and the control sets is constructed in accordance with the multivariate nature of the patterns.

We have applied our classification method to gene expression data from 60 human cancer cell lines. Results for the classification of melanoma, p53 mutations and the GI_{50} activity of Chlorambucil are excellent. They range from 0% to about 40% sum of false positive and false negative probability. The high sensitivity and specificity of the method for complex phenotypes such as p53 show that the method can successfully deal with multiple independent sub-phenotypes. We also report results from one of the first attempts to predict drug effectiveness from gene expression data.

We compare our method with other supervised learning schemes: the gene by gene method of [11] and the support vector machine method of [15]. The classification based on pattern discovery performed almost as good or better than the other methods. However these comparisons depend strongly on the data set under classification. One relevant conclusion in this context is that our approach is especially well suited to treat complex phenotypes, composed of several sub-phenotypes. We showed that such is the case for the p53-related phenotype and for a synthetic data set.

Besides the better predictive power, another advantage of our method over SVM is that our method highlights the relevant marker genes, their expression range in the phenotype, and the independent patterns that relate them. Such information is highly desirable for discovering the mechanism for various diseases at the molecular level. The SVM method, on the other hand, is more like a black box for classification.

The method described in this paper is a significant contribution to the set of tools for the analysis of gene expression microarray data. It constitutes proof of concept for a range of important practical applications for diagnostics and for the design of highly specific therapies. It may also help understanding the structure of the underlying gene regulatory networks and ultimately the mechanism responsible for various diseases.

5 Acknowledgments

Special thanks go to the Whitehead Institute team, Jill Mesirov, Donna Slonim, and Pablo Tamayo for the Gene Expression datasets. We also thank Ajay Royyuru, Reece Hart, Laxmi Parida, and Ruud Bolle for many useful discussions on pattern discovery and statistics, Donna Slonim for a careful reading of the manuscript and Hilary Coller for very helpful insights on the NCI-60 data.

6 References

[1] Lockhart, D. J.; Dong, H.; et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnol. 14:1675-1680.

[2] Brown, P.O.; and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays, Nature Genetics Suppl. 21:33-37.

[3] DeRisi, J.; Penland, L; et al. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nature Genetics 14:457-460.

[4] Wodicka, L.; Dong, H.; et al. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nature Biotechnol. 15:1359-1367.

[5] Cho, R. J.; Campbell, M. J.; et al. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2:65-73.

[6] Chu, S.; DeRisi,J.L.; et al. (1998). The transcriptional program of sporulation in budding yeast. Science 282:699-705.

[7] Iyer, V. R.; Eisen, M.B.; et al. (1999). The transcriptional program in the response of human fibroblasts to serum. Science 283:83-87.

[8] DeRisi, J.L.; Iyer, V.R.; et al. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278:680-686.

[9] Perou, Ch.; Jeffrey, S.S.; et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc Natl Acad Sci U S A 96:9212-9217.

[10] Alon, U.; Barkai, N.; Notterman, D.A.; Grish, K.; Ybarra, S.; Mack, D.; and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 96:6745-6750.

[11] Golub, T.R.; Slonim, D.K. ; et al. (1999), Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286**:**531-537.

[12] Eisen, M. B., Spellman, P. T.; et al. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95:14863-14868.

[13] Tamayo, P., Slonim, D.; et al. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A 96: 2907-2912.

[14] Ben-Dor A and Yakhini Z. (1999). Clustering Gene Expression Patterns. In *Proc. of the 3rd International Conference on Computational Molecular Biology,* 33-42. Lyon, France: ACM Press.

[15] Brown, M.P.S.; Grundy, W.N.; Lin, D.; Cristianini, N.; Sugnet, C.; Ares, M.; and Haussler, D. (1999). Support Vector Machine Classification of Microarray Gene Expression Data, University of California Technical Report USCC-CRL-99-09. Available at: http://www.cse.ucsc.edu/research/compbio/genex.

[16] D'haeseller, P.; Wen X.; Fuhrman, S.; and Somogyi, R. (1998). Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data. In *Information Processes in Cells and Tissues*, 203-212, Paton, R.C. and Holcombe, M. Eds., Plenum Publishing.

[17] Califano, A. (1999), SPLASH: Structural Pattern Localization Algorithm by Sequential Histograming. Bioinformatics, in press. Preprints available at http://www.research.ibm.com/topics/popups/deep/math/html/splashexternal.PDF.

[18] Chvatal, V. (1979) . A greedy heuristics for the set covering problem. Math. Oper. Res. 4:233-235.

[19] Welch, B.L. (1939). Note on Discriminant Functions. Biometrika 31:218-220.

[20] Lehmann, E.L. (1986). Testing Statistical Hypotheses. Second Edition. Springer-Verlag.

[21] Weinstein, J. N.; Myers, T. G.; et al. (1997). An information-intensive approach to the molecular pharmacology of cancer. Science 275:343-349.

[22] Various activity data for Chlorambucil on the NCI-60 cell lines can be found by using NSC access number 3088 on site http://dtp.nci.nih.gov.