BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 555–565

Binary analysis and optimization-based normalization of gene expression data

Ilya Shmulevich and Wei Zhang

Cancer Genomics Laboratory, Department of Pathology, University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd, Box 85, Houston, TX 77030, USA

Received on August 3, 2001; revised on October 11, 2001; accepted on November 23, 2001

ABSTRACT

Motivation: Most approaches to gene expression analysis use real-valued expression data, produced by high-throughput screening technologies, such as microarrays. Often, some measure of similarity must be computed in order to extract meaningful information from the observed data. The choice of this similarity measure frequently has a profound effect on the results of the analysis, yet no standards exist to guide the researcher.

Results: To address this issue, we propose to analyse gene expression data entirely in the binary domain. The natural measure of similarity becomes the Hamming distance and reflects the notion of similarity used by biologists. We also develop a novel data-dependent optimization-based method, based on Genetic Algorithms (GAs), for normalizing gene expression data. This is a necessary step before quantizing gene expression data into the binary domain and generally, for comparing data between different arrays. We then present an algorithm for binarizing gene expression data and illustrate the use of the above methods on two different sets of data. Using Multidimensional Scaling, we show that a reasonable degree of separation between different tumor types in each data set can be achieved by working solely in the binary domain. The binary approach offers several advantages, such as noise resilience and computational efficiency, making it a viable approach to extracting meaningful biological information from gene expression data.

Contact: is@ieee.org

1 INTRODUCTION

One of the most popular approaches to genome-wide expression analysis has been clustering, which is a type of unsupervized analysis. The goal of clustering is to group together objects that are similar. The objects can either be genes or samples, such as different tissues. For example, if two tissues have similar gene expressions, then they would be grouped together. Most approaches to gene expression analysis employ real-valued expression data. For instance, various clustering algorithms that have been used for gene expression data, such as hierarchical clustering (Eisen *et al.*, 1998), *k*-means clustering (Tavazoie *et al.*, 1999), self-organizing maps (Tamayo *et al.*, 1999), and others (Ben-Dor and Yakhini, 1999; Lukashin and Fuchs, 2001; Herrero *et al.*, 2001) typically assume that similarity measurements have been computed in the continuous, rather than the discrete, domain. For example, linear correlation, rank-based or ordinal correlation such as Spearman's ρ or Kendall's τ (Kendall and Gibbons, 1990), and norm-based distances (e.g. Euclidean distance), typically operate on continuous-level expression values.

One major difficulty is that the manner in which we quantify the notion of similarity has a most profound effect on the resulting clustering. Furthermore, no theory exists on how to choose the best similarity measure (Brazma and Vilo, 2000). This is hardly surprising as different measures emphasize different aspects and structure of the underlying data and the choice should expose those characteristics that would best reflect the goals of the analysis. To make matters worse, different clustering algorithms often produce drastically different resulting configurations, thus necessitating the reliance on prior biological knowledge in order to make meaningful interpretations (D'haeseleer *et al.*, 2000) and spurring the development of quantitative clustering validation methods (Yeung *et al.*, 2001; Zhang and Zhao, 2000). On a more fundamental level, the lack of 'golden standards' or consensus on the choice of similarity measures or clustering algorithms may simply reflect the pitfalls of attempting to use quantitative techniques to make qualitative interpretations or trying to understand and capture global coarse-grained phenomena with fine-grained models and methods.

Perhaps one of the earliest paradigm shifts towards coarse-scale qualitative modeling in genomics appeared approximately thirty years ago in the area of genetic network modeling, with the introduction of *Boolean Networks* (Kauffman, 1969; Glass and Kauffman, 1973). In this model, gene expression is quantized to only two levels: ON and OFF. The expression level (state) of each gene is functionally related to the expression states of

some other genes, using logical rules. Computational models that reveal these logical interrelations have since then been successfully constructed (Yuh *et al.*, 1998). Recent research seems to indicate that many realistic biological questions may be answered within the seemingly simplistic Boolean formalism, which emphasizes fundamental, generic principles rather than quantitative biochemical details (Huang, 1999). This model system has yielded insights into the overall behavior of large genetic networks (Somogyi and Sniegoski, 1996; Szallasi and Liang, 1998; Wuensche, 1998; Thomas *et al.*, 1995; Shmulevich *et al.*, 2001) and allows the study of large data sets in a global fashion. While Boolean models have certainly earned their place in computational and theoretical modeling of genetic regulatory networks, few attempts have been made to apply them to real gene expression data.

Generally speaking, there are several advantages to working in the binary domain. First, there are many algorithms already available for supervized learning in the binary domain, such as logical analysis of data (Boros *et al.*, 1997). Boolean-based classification algorithms have already been considered for gene expression based cancer classification (Akutsu and Miyano, 2001). Second, the quantization to binary values provides a certain level of robustness to noise in the observed data and may even improve accuracy of classification and simplify the obtained models (Pfahringer, 1995; Dougherty *et al.*, 1995). Finally, there is usually a significant computational advantage to working in the binary domain, such as reduced training times. Of course, there may be reasons why one may not wish to use this approach in certain problem settings. For example, since binarizing gene expression data results in significant loss of information as far as the actual expression levels are concerned, some applications, such as studying the dynamics of gene expression in cell-cycle regulation, would inherently preclude the use of this approach. Therefore, as always, the set of tools and data representation with which we work, along with their advantages and limitations, should be chosen in view of and guided by our goals of analysis and modeling.

In this paper, we explore the possibility of analyzing gene expression data in the binary setting. The main question with which we will concern ourselves is whether or not a sufficient level of detail required by our goals of analysis can be preserved when gene expression is quantized to only two levels. We will go about answering this question by considering two sets of gene expression data, produced by two different cDNA microarray platforms: membrane-based microarray with radioisotope-labeled cDNAs and glass-based microarray with fluorophore-labeled cDNAs. The first set of data was gathered from glioma tissue specimens from 26 patients, constituting four types of tumors. The tumor types, diagnosed according to the recently revised World Health Organization Classification of Tumours of the Nervous System (WHO 2000), were: Glioblastoma Multiforme (GM), Anaplastic Astrocytoma (AA), Anaplastic Oligodendroglioma (AO), and Low-grade Oligodendroglioma (OL). The second set of data, using the glass-based microarrays, contained gene expression data from leiomyosarcoma specimens from three excised sarcomas from an equal number of different patients. For two of the tumors, replicate experiments were performed three times using the same RNA samples. From the third tumor, carefully mapped specimens from four different spatial locations were used.

As one of the contributions of this paper, it will be shown, using both data sets, that a reasonable degree of separation between different tumor types can be achieved by working entirely in the binary domain. The tool that we will use for this analysis is Multidimensional Scaling (MDS; Borg and Groenen, 1997). MDS exposes the underlying similarity by trying to directly preserve all interpoint distances. We prefer standard nonmetric MDS to other unsupervized learning methods because it is more 'objective,' as it treats all distances equally. Some other methods, such as Sammon's mapping or the self-organizing map, emphasize the preservation of local distances. Yet another popular approach, principal component analysis, cannot take into account nonlinear structures, since it describes the data in terms of a linear subspace (Kaski, 1997).

Since we ultimately need some measure of similarity between tissue types, perhaps the most prudent approach would be to select a similarity measure that distinctly reflects the notion of similarity used by biologists when comparing gene expressions from different tissue samples. The latter essentially amounts to counting the number of genes that show significant differential expression between the two tissues. The lower the number, the more similar are the two tissues. The well-known Hamming distance is a natural choice for this purpose, once both tissues are represented by suitably discretized expression values.

Mathematically, this can be represented as follows. For each gene, a suitable threshold is chosen such that all genes whose expression levels exceed this threshold get assigned a label of '1' while the rest of the genes get the label '0.' In other words, all genes get quantized to two levels. Then, the number of mismatches or equivalently, the Hamming distance, between two tissues is used as a measure of (dis)similarity. Thus, the similarity simply reflects the number of genes that significantly 'disagree' in their expression levels. We will also discuss a procedure for selecting the above-mentioned thresholds.

The above task is confounded, however, by the fact that

some arrays may have different overall (average) intensities, due to various conditions such as photomultiplier gain or other parameter settings, amounts of exposure, etc. It is commonly assumed that the sources of error are multiplicative and thus, the true expression levels are modified by a multiplicative factor (Hartemink *et al.*, 2001). A recently introduced model for the error structure in microarrays by Rocke and Durbin (2001) also supports this claim, especially for genes that are at least moderately expressed.

Because of the above phenomenon, any gene belonging to an array with an overall higher intensity has a greater chance of getting set to 1. This problem requires normalization of the gene expression data—a crucial preprocessing procedure that is employed by nearly all gene expression studies in which data from one array must be compared to data on another array. A number of approaches may be taken. The data can be normalized with respect to some statistic, such as the mean, median, maximum, and standard deviation (e.g. Golub *et al.*, 1999). Alternatively, they can be normalized with respect to some set of 'housekeeping' genes (e.g. GAPDH; Kim *et al.*, 2000). Once again, there is no currently accepted golden standard (Celis *et al.*, 2000) and thus, the chosen method should be motivated by the application at hand and the goals of the data analysis. Another contribution of this paper is a data-dependent optimization-based normalization procedure, especially well-suited for the binarization task.

The paper is organized as follows. Section 2 contains some mathematical preliminaries and definitions as well as a motivation for normalization, as a preprocessing step for binarization. Section 3 presents the proposed optimization-based normalization method as well as a simulation example. Finally, Section 4 describes the binary approach to analyzing gene expression data and provides several examples.

2 DEFINITIONS AND MOTIVATION FOR NORMALIZATION AS A NECESSARY STEP FOR BINARIZATION

Let n be the number of genes and k be the number of tissues/arrays. Let $G_{i,j}$ be the observed expression of gene i in array j. Then, the *gene profile* of gene i ($1 \leqslant i \leqslant n$) is the vector of values $G_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,k})$. Thus, G is a matrix whose rows are gene profiles and columns are *arrays*. As an estimate of the average intensity of array j, we can consider either the sample mean $\mu_j = n^{-1} \cdot \sum_{i=1}^{n} G_{i,j}$ or the sample median $\nu_j = \text{med}\,(G_{1,j}, \ldots, G_{n,j})$, $1 \leqslant j \leqslant k$. The *mean* and *median profiles* are defined as $\mu_G = (\mu_1, \ldots, \mu_k)$ and $\nu_G = (\nu_1, \ldots, \nu_k)$, respectively. The median is often preferred as it is known to be very robust in the presence of outliers (Huber, 1981). That is, it's value is much less sensitive to extreme values (i.e. very highly expressed genes) than the mean and gives a more realistic estimate of the 'average' array intensity.

In order to illustrate the necessity of normalization, let us first informally consider an example from the Glioma data set, consisting of $n = 597$ genes and $k = 26$ tissue samples. Figure 1a shows the gene profile for the APO-2 Ligand (APO2L). By simply looking at this profile, one may be tempted to conclude that this gene is highly expressed in tissue #21, an oligodendroglioma, or at least, higher than in other tumor tissues. However, it is informative to consider the median profile as well, shown in Figure 1b. It can be readily seen that for tissue #21, the 'average' expression of the genes in the entire array is higher than in other arrays. In fact, the entire shape of the profile of APO2L is similar to the median profile. Indeed, the correlation between the APO2L profile and the median profile ν_G is equal to 0.95. This fact leads us to believe that the observed expression value of any particular gene may be highly 'influenced' by the characteristics of the array in which it is observed. Thus, normalization, or more generally, standardization is a crucial preprocessing step that must be taken prior to the comparison of data from different microarray experiments. One obvious possibility is to correct for the differences in average (median) intensities of the arrays by making all medians equal—this can be accomplished by simply dividing each gene's expression value by the median of that array. Thus, the median normalized gene profile of gene i is equal to

$$\tilde{G}_i = (G_{i,1}/\nu_1, G_{i,2}/\nu_2, \ldots, G_{i,k}/\nu_k). \qquad (1)$$

Clearly, the medians of the median normalized arrays are all equal to 1; that is, $\nu_{\tilde{G}} = (1, 1, \ldots, 1)$. Figure 1c shows the median normalized gene profile of APO2L. It can be seen, for example, that the expression of this gene in tissue #21 is no longer the highest and the correlation between this profile and the median profile is equal to 0.26.

Let's continue by considering the above phenomenon for all gene profiles. As above, we can compute the correlation $\rho(G_i, \nu_G)$ between each (unnormalized) gene profile G_i and the median profile ν_G and then look at the distribution of the resulting correlations[†]. If there are many high correlations, then it is likely that the characteristic of each array has a strong effect on the expression values of individual genes.

Let us consider the Glioma data set again. Figure 2 shows the distribution (density) of these 597 correlations as a solid line (using a kernel density estimate). It can be seen that this distribution is heavily skewed towards high correlation values (mean = 0.56, SD = 0.27),

[†] We use the symbol ρ to denote arbitrary correlation measures. In this example, we consider the standard linear correlation coefficient, but in the sequel we will also use ordinal correlations.

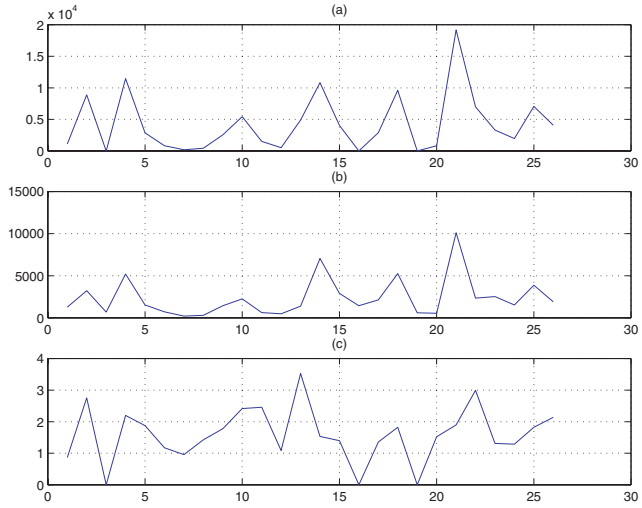

Fig. 1. (a) Gene profile of unnormalized APO2L; (b) median profile ν_G; (c) normalized gene profile of APO2L.

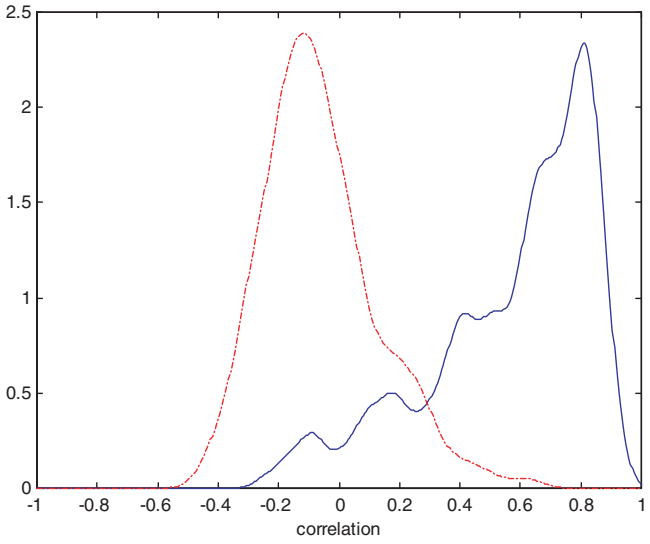

Fig. 2. Distributions of correlations between each gene profile and the median profile (solid) and each median normalized gene profile and the median profile (dashed).

indicating that many gene profiles are strongly coupled with the median profile. Then, we performed the median normalization as described above and computed 597 correlations $\rho(\tilde{G}_i, \nu_G)$ between each normalized gene profile $\tilde{G}_i$ and the median profile ν_G. The distribution of these correlations is shown as a dashed line in Figure 2. As can be seen, most correlations are small (mean $= -0.06$, SD $= 0.19$), implying that the average[‡] array intensity has been mostly 'decoupled' from the gene profiles.

We should note here that we could have also considered the correlations between every pair of gene profiles, producing a 597×597 matrix of correlations. Indeed, if we plot the distributions of these correlations for both normalized and unnormalized arrays, we will see the same effect as in Figure 2: normalizing by the medians tends to partially decorrelate the gene profiles. This is shown in Figure 3.

However, we must be cautious about how we interpret these results. An important first question is whether or not we're actually interested in decorrelating the gene profiles. After all, some genes are functionally related and thus exhibit similar expression profiles, even under ideal conditions. In fact, that is the central goal of clustering analysis—to discover groups of genes with similar expression profiles. Additionally, one of the goals in this paper is to quantize the gene expression profiles into discrete values, e.g. $0/1$, by choosing thresholds for each profile. If we were to decorrelate the gene profiles, we would be mapping them, by means of a linear whitening

[‡] We use the term 'average' loosely and here it signifies the median as a measure of array-wide expression.

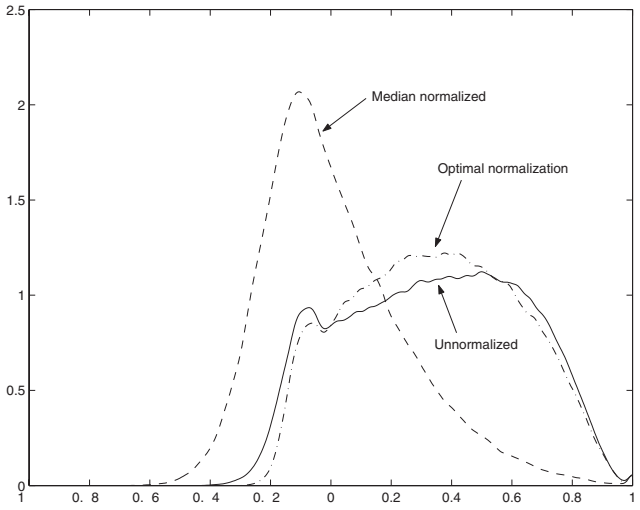

Fig. 3. Distributions of correlations between every pair of unnormalized gene profiles (solid), between every pair of median normalized gene profiles (dashed), and between every pair of gene profiles normalized by optimal scaling parameters (dash–dot; see Section **Simulation**).

transformation, into another space in which the values of the mapped profiles would no longer represent gene expression levels, but rather their linear combinations. This method has recently been applied to time series data in Alter *et al.* (2000) and Holter *et al.* (2000) and while being useful for analyzing large data sets by reducing the

dimensionality and filtering out noise, it would preclude the application of any subsequent method that expects (possibly normalized) gene expressions as inputs, such as binarization or clustering.

Thus, our goals are driven by the following requirements: (a) to remove as much as possible the effect of global array characteristics on the gene profiles so as to make them comparable across experiments; (b) to preserve the interpretability of the transformed profiles as gene expression profiles. Because of the above considerations, the first requirement can be met by decorrelating the gene profiles from the median profile as much as possible, as shown in Figure 2, while the second requirement implies that our transformation must consist of a simple scalar multiplication of each array, such as in (1). To reformulate the latter in terms of a linear transformation, our linear transformation matrix A must be diagonal. If the scalars are just the medians, as discussed above, then

$$A = \begin{bmatrix} \nu_1^{-1} & 0 & \cdots & 0 \\ 0 & \nu_2^{-1} & \ddots & 0 \\ 0 & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \nu_k^{-1} \end{bmatrix} \tag{2}$$

and $\tilde{G} = G \times A$. A natural question to ask is: can we do better at decorrelating the gene profiles from the median profile than by using the median profile itself?

3 PROPOSED NORMALIZATION METHOD

As a first step, we must decide on a criterion that we can use to judge the amount of decorrelation achieved. Intuitively speaking, this should be reflected in the distribution of the correlations after normalization, as shown in Figure 2. A good decorrelation should produce a density that is centered around zero and whose standard deviation is small. More formally, let the sample mean and the sample standard deviation of the correlations be given by

$$\mu_\rho = \frac{1}{n}\sum_{i=1}^{n} \rho(\tilde{G}_i, \nu_G) \tag{3}$$

and

$$\sigma_\rho = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\rho(\tilde{G}_i, \nu_G) - \mu_\rho)^2}, \tag{4}$$

respectively, where the $\tilde{G}_i = (G_{i,1}/a_1, G_{i,2}/a_2, \ldots, G_{i,k}/a_k)$ are the normalized profiles by k scaling parameters $a_1, a_2, \ldots, a_k$. We could have also used other (robust) estimates of location and scale of the distribution, such as the median and Median of Absolute Deviations from the median (MAD) in place of (3) and (4), respectively.

The method we propose consists of choosing the k scaling parameters $a_1, a_2, \ldots, a_k$, or equivalently, the diagonal transformation matrix A, such that some function of μ_ρ and σ_ρ is minimized. One choice for such a function is $f(\mu_\rho, \sigma_\rho) = |\mu_\rho| + \sigma_\rho$ and corresponds to the weighted sum strategy in multiobjective optimization, where the weights are equal. Of course, we could also assign weights to the mean and standard deviation to emphasize their relative importance. Additionally, we may impose the constraints $1 \leqslant a_j \leqslant \max_i(G_{i,j})$, for all $j = 1, \ldots, k$ (it is always the case that $\max_i(G_{i,j}) > 1$). Since all expression values are positive, this guarantees that the maximum values of the normalized arrays cannot exceed the maximum values of the unnormalized arrays, nor can the maxima be less than 1. Thus, the optimization problem may be posed as

$$\begin{gathered} \text{minimize } |\mu_\rho| + \sigma_\rho \\ \text{subject to } 1 \leqslant a_j \leqslant \max_i(G_{i,j}),\ \ j = 1, \ldots, k. \end{gathered}$$

One other issue merits our attention. It is well known that the standard linear (Pearson) correlation coefficient can be unduly influenced by extreme values. Because of this, the optimization algorithm would choose very high or very low scaling parameters such that the correlations would be as small as possible, in line with minimizing the objective function $f(\mu_\rho, \sigma_\rho)$. In other words, many of the parameters $a_1, a_2, \ldots, a_k$ would be very likely to attain the imposed bounds $1 \leqslant a_j \leqslant \max_i(G_{i,j})$. This is, of course, undesirable, since the normalized values in $\tilde{G}_i$ would also be unduly influenced. The solution to this problem is to use ordinal correlation coefficients, such as Spearman's ρ or Kendall's τ, which operate on ranks of the data rather than on the data values themselves, and are thus much less sensitive to outliers and unplanned defects in the data. In general, these nonparametric correlation coefficients are much more indicative of true underlying correlation than the linear correlation coefficient. So, the values of μ_ρ and σ_ρ would be computed from the distribution of ordinal correlation coefficients.

Unfortunately, ordinal correlation coefficients are also highly nonlinear and discontinuous. That is, slightly changing the value of one of the data points can potentially change its rank and consequently, all the other ranks. Similarly, changing the values of some data points may result in no change in the ranks. This precludes the use of traditional gradient-based optimization algorithms (Luenberger, 1989), which make some continuity assumptions about the gradient. Moreover, the use of ordinal correlation coefficients is likely to result in many local minima—a major weakness of gradient-based methods. Because of these considerations, we have decided to employ a Genetic Algorithm (GA) for the optimization (Goldberg, 1989). GAs are a class of stochastic sampling methods that make no assumptions about the characteristics of the problem, such as continuity or

differentiability, and proceed through the search space by making stochastic decision rules. They are well-suited for multi-modal function optimization as they are much less likely to get stuck in local optima. GAs proceed through the search space of a function by using a simulated evolution strategy, that is, the survival of the fittest. Thus, good solutions are more likely to survive and reproduce than bad solutions, resulting in an improvement in the next generations. GAs explore all regions of the search space and exploit promising areas via the operations of mutation, crossover, and selection applied to individuals of the population.

3.1 Simulation example

We applied the proposed optimization-based normalization method to the Glioma data set. As explained above, we have used a GA-based optimization and the Spearman's ρ ordinal correlation coefficient for computing the objective function $f(\mu_\rho, \sigma_\rho)$. When we normalize all gene profiles by the median profile, as in (1), the value of the objective function $f(\mu_\rho, \sigma_\rho)$ is 0.2525 whereas when we normalize by the optimum scaling parameters $a_1, a_2, \ldots, a_k$ produced by the algorithm, the value of $f(\mu_\rho, \sigma_\rho)$ is 0.1809. Figure 4 shows the resulting distribution of ordinal correlations. The solid-line plot in Figure 4 shows the distribution of correlations between the median profile and each unnormalized gene profile. As can be seen, this is similar to the solid-line plot in Figure 2, which shows the distribution of linear (Pearson) correlations. The dashed-line plot of Figure 4 shows the distribution of correlations when the gene profiles are normalized by the median profile. Finally, the dash–dot plot on the same figure shows the distribution of correlations when the gene profiles are normalized by the optimum scaling parameters produced by the proposed algorithm. It can be seen that the proposed algorithm results in many more small correlations and fewer large correlations than the median-normalization.

We again would like to stress that, in line with our goals and expectations, the proposed algorithm does not achieve a decorrelation between the gene profiles themselves. In fact, quite to the contrary, we expect that the correlations that do exist between the optimally normalized gene profiles are more likely to be due to biological reasons than the corresponding correlations between the unnormalized profiles. A distribution of correlations between every pair of optimally normalized gene profiles is shown in Figure 3. As can be seen, this distribution is quite similar to the unnormalized case, even though the distributions of correlations between the optimally-normalized gene profiles and the median profile are entirely different (Figure 4).

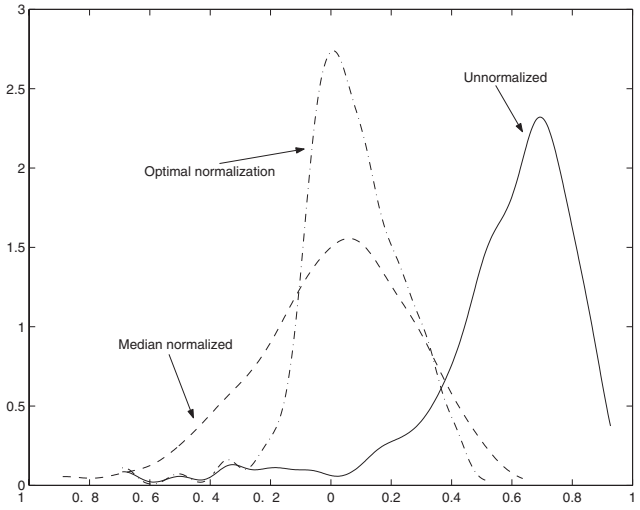

Fig. 4. The solid-line plot shows the distribution of ordinal correlations between the median profile and each unnormalized gene profile; the dashed-line plot shows the distribution of correlations between the median profile and median-normalized gene profiles; the dash–dot plot shows the distribution of correlations between the median profile and each gene profile normalized by the optimal scaling parameters produced by the proposed algorithm.

4 GENE EXPRESSION DATA ANALYSIS IN THE BINARY DOMAIN

Having discussed the proposed optimization-based procedure for normalizing gene expression data, we are now ready to consider the data analysis in the binary setting. Naturally, in order to quantize real-valued gene expressions into binary values, we need a procedure for selecting a threshold. This is discussed next.

Different genes manifest different ranges of gene expression levels. Consequently, selecting a global threshold for each array is not appropriate and there should be an individually selected threshold for each (normalized) gene profile. While there are many ways to select a threshold, the basic idea we use is that the location of the threshold should be where the separation between low and high expression values is greatest. In other words, if we were to sort all expression values in a given gene profile, the threshold would correspond to the location where the first 'big jump' occurs, that is, where the finite differences between successive sorted values first exceed some predefined value.

Figure 5 illustrates this idea graphically. The figure contains a normalized gene profile and the corresponding sorted gene profile. It can be seen by considering the latter that there are 15 tissues in which this gene has rather low expression. The first 'jump' occurs between the 15th and 16th position, which corresponds to the jump between

tissues #21 and #10. Thus, tissue #10 and all tissues with expression values higher than it in this gene profile should be set to 1, while the others below it should be set to 0. The 11 tissues for which this gene is equal to 1 are shown with circles. The only other remaining issue is how to define 'big jump' in the sorted profile. The most difficult case to deal with would be when all gene expression values are equally spaced between the minimum and maximum in which case the sorted profile would be a straight line and every difference between successive values would be equal. We will use this difference as our threshold for deciding when the first big jump occurs. Thus, we have the following algorithm for converting a (normalized) gene profile $G_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,k})$ into a binary-valued profile $B_i = (B_{i,1}, B_{i,2}, \ldots, B_{i,k})$.

Algorithm 1 Binarize: INPUT G_i, OUTPUT B_i

$S_i \leftarrow \text{sort}(G_{i,1}, G_{i,2}, \ldots, G_{i,k})$
for $j = 1$ to $k - 1$ **do**
  $D_{i,j} \leftarrow (S_{i,j+1} - S_{i,j})$
end for
$t \leftarrow (S_{i,k} - S_{i,1})/(k-1)$
$m = \min\{j : D_{i,j} > t\}$
for $j = 1$ to k **do**
  if $G_{i,j} \geqslant S_{i,m+1}$ **then**
    $B_{i,j} \leftarrow 1$
  else
    $B_{i,j} \leftarrow 0$
  end if
end for

Of course, other 'edge detector'-type algorithms could also be employed to detect the jump in the sorted gene profile. For example, the sample variance calculated in a running-window fashion will be one possibility.

One additional difficulty needs to be addressed. Due to the inherent noise present in microarrays, genes that are not expressed still exhibit some variation. Consequently, the direct application of Algorithm 1 to a gene profile in which some or all of the values are within the typical level of noise is likely to result in spurious binarization. The reason for this is that Algorithm 1 is scale invariant; in other words, multiplying the entire gene profile by a constant does not change the output of the algorithm. This calls for the need to first screen out those gene expression values that are below a certain noise floor threshold. The latter can be estimated by using negative controls and/or background. Having determined this global threshold, all gene values below it are set to 0 and the remaining genes are subjected to Algorithm 1. As a result, some gene profiles may be completely set to 0 after which the binarization algorithm would have no effect.

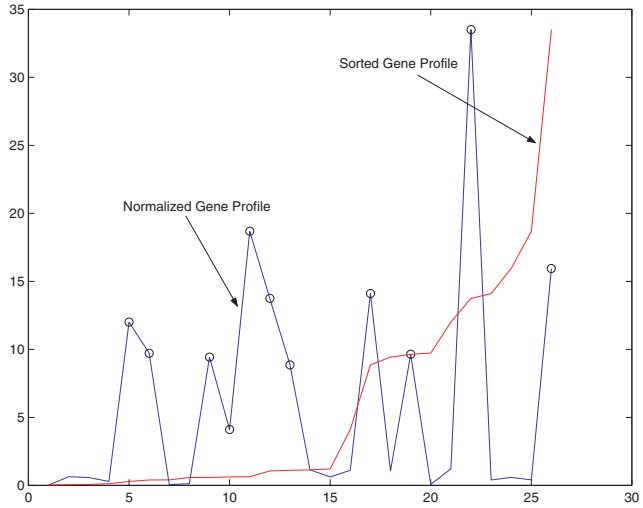

Fig. 5. An example of a normalized gene profile (blue) and the sorted profile (red), illustrating the binarization procedure. The circles indicate those tissues/arrays for which this gene is set to 1.

4.1 Data analysis examples in the binary domain

We present examples of two different data sets being analysed in the binary domain. The first example uses the Glioma data set, containing 597 genes. The data collection procedure was described in Fuller *et al.* (1999). We first normalized the data by the optimal parameters, as described in Section 3, and then performed the above binarization procedure. Afterwards, we obtained a 26×26 matrix containing Hamming distances between all pairs of tumors and subjected it to standard nonmetric MDS, visualized in 3-dimensions. It can be seen from Figure 6 that the glioblastomas (circles) and oligodendrogliomas (squares) are well separated, consistent with the pathological characterizations, being the most and least aggressive types of tumors, respectively. Also not surprising is the fact that the AAs and AOs, exhibiting intermediate characteristics, are positioned between the GMs and OLs. Especially encouraging is the fact that the results are consistent with prior biological knowledge, considering that only binary values were used for the analysis.

The second example uses the Sarcoma data set, containing a total of 10 arrays (three replicates of two different tumors and four spatially different specimens from a third tumor) and 4800 gene measurements. Unlike in the Glioma data set, these microarrays were cDNA glass slides to which matched tumor and normal samples, labeled with Cy3 and 5 respectively, were cohybridized. It is common practice to present data from microarray experiments as ratios of test samples against a common reference sample that functions as an internal standard to avoid gross errors caused by uneven hybridization. Since

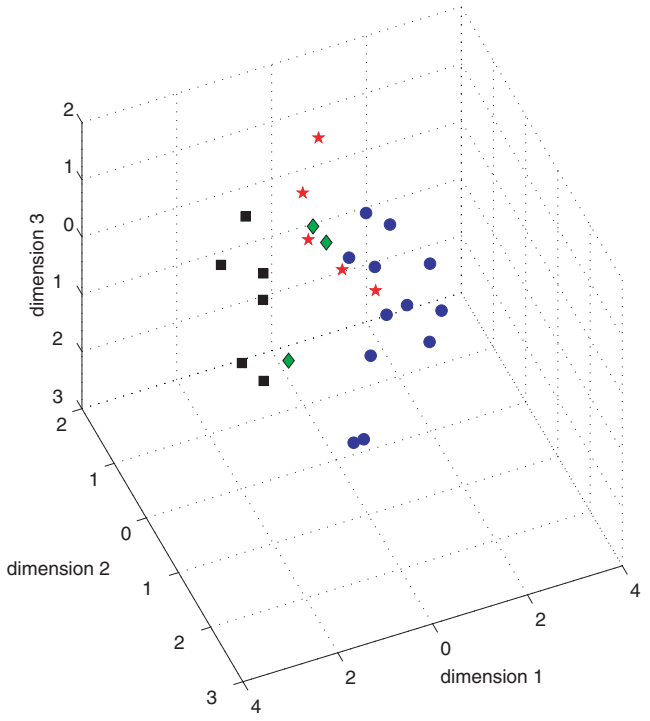

Fig. 6. A 3D MDS solution using Hamming distances between pairs of tumors in the Glioma data set. The four tumor types are: circles—GM, stars—AA, diamonds—AO, squares—OL.

many genes typically have low or undetectable expression levels, using ratios to study similarity between different samples is risky. Being ultimately interested in examining the similarity between tissues on the basis of their genetic 'signatures,' it would be preferable to simply use the single channels corresponding to the samples of interest without their corresponding reference channels. This is, of course, contingent on the premise that each channel is stable in the sense that it measures the true expression levels and is minimally affected by other variations, for example, due to uneven hybridization. This can be tested by assessing the reproducibility of the same tissue on different slides.

To test the reproducibility of our single-channel measurements, we performed several experiments in which a tumor sample from a patient was cohybridized to the array with a normal control tissue from the same patient. This was repeated three times for two different patients. If the reproducibility were high, we would expect that the three correlations (between first and second, first and third, and second and third measurements) would be high as well. We computed the standard (Pearson's) correlation coefficients as well as Spearman's ρ correlation coefficients. Table 1 shows these correlations between the three

Table 1. Pearson's and Spearman's correlation coefficients between replicates of single channels

	r_{12}, ρ_{12}	r_{13}, ρ_{13}	r_{23}, ρ_{23}
Tumor 1	0.89, 0.86	0.93, 0.89	0.89, 0.86
Tumor 2	0.94, 0.91	0.91, 0.88	0.95, 0.93

replicates for two different tumors. All correlations are quite high and significant with $p \ll 0.0001$. For purposes of comparison, we computed Spearman's ρ correlations between tumors 1 and 2, for each pair of replicates. These were equal to: 0.46, 0.49, and 0.53 for the three replicates. These correlations are considerably lower than those in Table 1, and it is not entirely surprising that they are not close to 0. After all, one would not expect most genes to exhibit drastically different expression levels between two tumors. The important thing is that correlations between tumors are lower than correlations between replicates of the same tumor.

Having convinced ourselves of the high stability of the single channels, we can proceed with the binary analysis only in terms of these single channels, without making use of the reference channel. Once again, we normalized the data by the optimal parameters chosen by the optimization method and then performed the binarization procedure. Having obtained a 10×10 matrix of Hamming distances, we performed MDS. The solution is shown in Figure 7. It can be seen that a reasonable degree of separation between the three tumors is achieved entirely in the binary domain. The data suggest, in addition, that a binary-valued tumor-specific gene 'signature' may exist.

5 DISCUSSION

Coarse-scale qualitative modeling approaches, in genomics in particular, can be justified from two major standpoints. From a practical standpoint, limited amounts of data and the noisy nature of the measurements can make useful quantitative inferences problematic. For example, in the case of genetic regulatory networks, it is difficult at the present time to predict the true expression levels of genes from those of other genes, in a global large-scale fashion. From a notional standpoint, the resolution of a chosen model should reflect the intended goals of analysis and the characteristics of the underlying physical phenomenon that we intend to capture. Once again, in the case of genetic networks, we may be interested, at least as a first step, in discovering generic qualitative principles rather than quantitative biochemical details. Boolean modeling is a natural framework for this purpose.

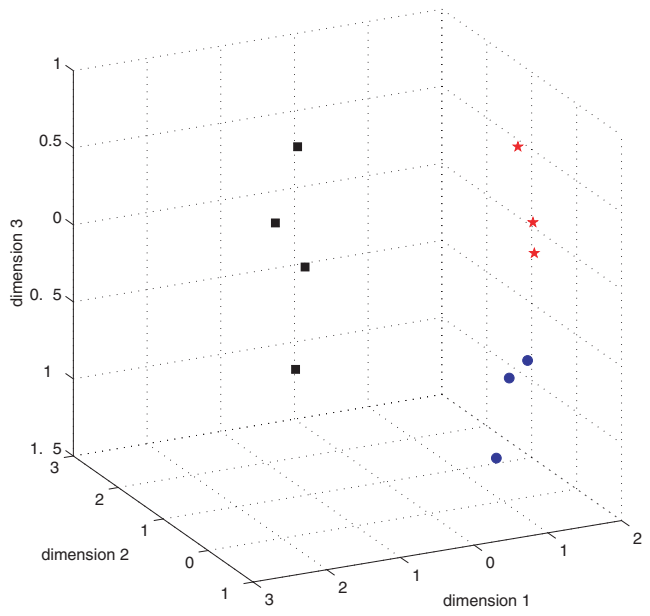

Fig. 7. A 3D MDS solution using Hamming distances between pairs of samples in the Sarcoma data set. Stars and circles correspond to the three replicates of tumors 1 and 2, respectively. The squares correspond to the four spatially different samples from tumor 3.

A question that is often posed is whether or not binary representations, by their very nature, lose essential information. A simple and concise answer is, of course, yes. However, to properly address such a question, it is necessary to make a clear distinction between what is understood by useful information and noise. Perhaps it would be helpful to draw an analogy with the field of image processing. An important problem in image analysis is to detect, identify, and possibly count various objects in an image. Depending on the process used to form the image (e.g. synthetic aperture radar, satellite remote sensing, astronomical imagery, microscope imaging, computed tomography, digital imaging, etc.), the image may be corrupted by various kinds of noise (e.g. due to atmospheric interference, transmission errors, etc.), blurring effects (e.g. due to motion, optical systems), and other distortions. In addition, and perhaps more importantly, the content of the image itself may naturally exhibit great variation from one image to another. For example, in a microscope image, it is natural to expect that all cells will have different shapes, sizes, and locations within the image. Yet, robust detection of objects and their structures should be possible in the face of all this uncertainty. A great number of powerful approaches based on so-called multiresolution or multiscale methods have proven to be highly effective for such image analysis problems (e.g. Starck *et al.*, 1998). In these approaches, several coarser-scale versions (at different resolutions) of the image are used in a concerted manner. Perhaps somewhat paradoxically, often a lower resolution image containing less *information* (in terms of pixel content) is in fact more *informative* (in terms of extracting objects and characterizing their properties).

In some sense, the binary representation of gene expression is the coarsest possible. We have explored the question of whether or not a sufficient level of detail required by our goals of analysis can be preserved when gene expression is quantized to only two levels. The ability to separate different tumor types based only on binary data seems to suggest the feasibility of performing classification analysis or inferring genetic network structure in the binary domain. Moreover, there may be other advantages such as resilience to noise, computational savings, simplification of models, and possible improvement of accuracy of classification. As an added benefit, it essentially avoids ambiguity in the choice of similarity measures and seems to reflect the notion of similarity that biologists use in their every-day language and scientific literature.

While the proposed normalization method is more effective under the chosen objective function than standard median normalization, there are some added difficulties. First, because of the discontinuous and nonlinear nature of rank correlations and the existence of many local optima, stochastic optimization methods become necessary. A common property shared by all stochastic optimization methods is that they can potentially produce different solutions every time they are executed. This is the price one has to pay for obtaining optimal or near-optimal solutions to very difficult optimization problems. On a more positive note, in practice, repeated executions of these methods typically produce very similar and often identical results, depending on the problem. In our particular application, the indeterminacy of the normalization parameters does not present an obstacle to effective data exchange—all that is necessary is to send the optimal normalization parameters together with the unnormalized data. Another potential drawback of the proposed method is the speed and convergence of the optimization algorithm. Despite some recent theoretical efforts (e.g. Rudolph, 1994), it is difficult to make general statements about convergence of GAs. In our case, the optimization algorithm as well as the rank correlation functions were implemented in MATLAB on a Pentium III 933 MHz, and required less than 5 min of computer time to produce the optimal scaling parameters.

Finally, as already mentioned in the Section **Introduction**, we have made an implicit assumption that the sources of error are multiplicative and thus, the true expression levels are modified by a multiplicative factor. That is why our normalization is also multiplicative. However, as new models of measurement error become

available (Rocke and Durbin, 2001; Baggerly *et al.*, 2001) and validated, more sophisticated estimation methods will need to be developed.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Latha Ramdas for valuable discussions and the referees for many helpful suggestions. This work was partially supported by the Tobacco Settlement Funds as appropriated by the Texas State Legislature, by a generous donation from the Michael and Betty Kadoorie Foundation, and by a grant from the Texas Higher Education Coordination Board under grant 003657-0039-1999.

REFERENCES

Akutsu,T. and Miyano,S. (2001) Selecting informative genes for cancer classification using gene expression data. In *Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing* (*NSIP*). Baltimore, MD, pp. 3–6.

Alter,O., Brown,P.O. and Botstein,D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10 101–10 106.

Baggerly,K.A., Coombes,K.R., Hess,K.R., Stivers,D.N., Abruzzo,L.V. and Zhang,W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J. Comput. Biol*.

Ben-Dor,A. and Yakhini,Z. (1999) Clustering gene expression patterns. In *Proceedings of the 3rd International Conference on Computational Molecular Biology* 3342. ACM Press, Lyon, France.

Borg,I. and Groenen,P. (1997) *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York.

Boros,E., Hammer,P.L., Ibaraki,T. and Kogan,A. (1997) Logical analysis of numerical data. *Math. Program.*, **79**, 163–190.

Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett.*, **480**, 17–24.

Celis,J.E., Kruhøffer,M., Gromova,I., Frederiksen,C., Østergaard,M., Thykjaer,T., Gromov,P., Yu,J., Pálsdóttir,H., Magnusson,N. and Ørntoft,T.F. (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett.*, **480**, 2–16.

D'haeseleer,P., Liang,S. and Somogyi,R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, **16**, 707–726.

Dougherty,J., Kohavi,R. and Sahami,M. (1995) Supervised and unsupervised discretization of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning*. Morgan Kaufmann, Tahoe City, CA, pp. 194–202.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Fuller,G.N., Rhee,C.H., Hess,K.R., Caskey,L.S., Wang,R., Bruner,J.M., Yung,W.K.A. and Zhang,W. (1999) Reactivation of insulin-like growth factor binding protein 2 expression in glioblastoma multiforme: a revelation by parallel gene expression profiling. *Cancer Res.*, **59**, 4228–4232.

Glass,K. and Kauffman,S.A. (1973) The logical analysis of continuous, non-linear biochemical control netoworks. *J. Theor. Biol.*, **39**, 103–129.

Goldberg,D. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hartemink,D.G., Jaakkola,I. and Young,R. (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. *Microarrays: Optical Technologies and Informatics* (*Proceedings of SPIE*). pp. 4266.

Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Holter,N.S., Mitra,M., Maritan,A., Cieplak,M., Banavar,J.R. and Fedoroff,N.V. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl Acad. Sci. USA*, **97**, 8409–8414.

Huang,S. (1999) Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J. Mol. Med.*, **77**, 469–480.

Huber,P.S. (1981) *Robust Statistics*. Wiley, New York.

Kaski,S. (1997) Data exploration using self-organizing maps. In *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series* 82. Finnish Academy of Technology, Espoo, Finland.

Kauffman,S.A. (1969) Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.*, **22**, 437–467.

Kendall,M. and Gibbons,J. (1990) *Rank Correlation Methods*, 5th edn, Oxford University Press, New York.

Kim,S., Dougherty,E.R., Chen,Y., Sivakumar,K., Meltzer,P., Trent,J.M. and Bittner,M. (2000) Multivariate measurement of gene expression relationships. *Genomics*, **67**, 201–209.

Kleihues,P. and Cavenee,W.K. (eds.) (2000) In *Pathology and Genetics of Tumours of the Nervous System*, 2nd edn (*World Health Organzation Classification of Tumours of the Nervous System*), Oxford University Press, New York.

Luenberger,D.G. (1989) *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA.

Lukashin,A.V. and Fuchs,R. (2001) Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics*, **17**, 405–414.

Pfahringer,B. (1995) Compression-based discretization of continuous attributes. In Prieditis,A. and Russell,S. (eds), *Machine Learning: Proceedings of the Twelfth International Conference*. Morgan Kaufmann, San Francisco.

Rocke,D.M. and Durbin,B. (2001) A model for measurement error for gene expression arrays. *J. Comput. Biol.*, in press.

Rudolph,G. (1994) Convergence analysis of canonical genetic algorithms. *IEEE Trans. Neural Netw.*, **5**.

Shmulevich,I., Dougherty,E.R., Kim,S. and Zhang,W. (2001) Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, in press.

Somogyi,R. and Sniegoski,C. (1996) Modeling the complexity of gene networks: understanding multigenic and pleiotropic regulation. *Complexity*, **1**, 45–63.

Starck,J.-L., Murtagh,F. and Bijaoui,A. (1998) *Image Processing and Data Analysis: the Multiscale Approach*. Cambridge University Press, Cambridge.

Szallasi,Z. and Liang,S. (1998) Modeling the normal and neoplastic cell cycle with realistic Boolean genetic networks: their application for understanding carcinogenesis and assessing therapeutic strategies. *Pac. Symp. Biocomput.*, **3**, 66–76.

Tavazoie,S., Hughes,D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.

Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.

Thomas,R., Thieffry,D. and Kaufman,M. (1995) Dynamical behaviour of biological regulatory networks—I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.*, **57**, 247–276.

Wuensche,A. (1998) Genomic regulation modeled as a network with basins of attraction. *Pac. Symp. Biocomput.*, **3**, 89–102.

Yeung,K.Y., Haynor,D.R. and Ruzzo,W.L. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309–318.

Yuh,C.-H., Bolouri,H. and Davidson,E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.

Zhang,K. and Zhao,H. (2000) Assessing reliability of gene clusters from gene expression data. *Funct. Integ. Genom.*, **1**, 156–173.