Generative Model Based Analysis of Cancer Associated Gene Expression Matrices

Mattias Wahde
Div. of Mechatronics
Chalmers University of Technology
412 96 Göteborg, Sweden
mwahde@me.chalmers.se

Zoltan Szallasi
Department of Pharmacology,
Uniformed Services University of the Health Sciences
Bethesda, MD, 20814
zszallas@mxc.usuhs.mil

Abstract

One of the main aims of analyzing cancer associated gene expression matrices is to identify a subset of genes that is consistently mis–regulated in a given type of tumor samples. Such a subset of genes forms, together with an appropriate function, a separator that can distinguish between normal and tumor samples. Separators can appear accidentally due to the high level of gene expression diversity detected in cancer. Various statistical methods can be used to estimate whether the appearance of a given separator is due to chance. However, the accuracy of all these tests will depend on the null hypothesis provided by the data structure. In this paper we are introducing generative models in order to simulate random, discrete gene expression matrices that retain the key features of massively parallel measurements in cancer. These include the number of changeable genes and the level of gene co–regulation as reflected in their pair–wise mutual information content. We show that the probability of the chance appearance of separators can be underestimated by many orders of magnitude if random and independent selection of mis–regulated genes is assumed instead of using the generative model outlined in this paper.

Introduction

The recent publication of several cancer associated large–scale gene expression matrices has clearly indicated that tumor biology has entered a new phase of analytical approaches. These matrices contain quantitative information about a large number of directly measured parameters, usually gene expression levels, that are typically listed as the rows of the matrix. The columns in these experiments correspond to different phenotypes such as different types of tumors or different treatments of either normal or neoplastic cells. Current computational biology is expected to define the different levels of analysis on these massively parallel data sets e.g. to what extent should knowledge based systems be involved.

In this paper we are focusing on analytical approaches that will use only the information contained in the gene expression matrices. There are two obvious ways of exploiting cancer associated gene expression matrices. Identification of separators or gene expression functions (Szallasi, 1998) determines a subset of genes the status of which, when coupled by an appropriate rule, will define the phenotypic state of cells. The classification of phenotypic samples on the other hand is supposed to identify subsets of samples with above average molecular similarity. These subsets can be later used to search for common genetic markers. The aim of this procedure, which was recently termed as tumor class discovery in cancer research (Golub *et al.*, 1999), is supposed to yield a group of tumor samples sharing a common set of genetic markers.

Cancer associated gene expression patterns show a high level of diversity. The average number of mis–regulated genes is on the order of 10% of all genes expressed in the given cell type (Perou *et al.* 1999). This variability will inevitably lead to the accidental appearance of separators and clusters in these data sets. The main aim of this paper is to introduce generative models in order to estimate the probability of accidental features of cancer associated gene expression data sets.

In this paper we will be focusing on discretized data. Continuous cDNA microarray measurements can be converted into ternary data as described by Chen *et al.* (1997). Their algorithm first calibrates the data internally to each microarray and statistically determines whether the data justifies the conclusion that a given gene is up– or down–regulated at a certain confidence level.

Separators

The purpose of **separators** is to identify patterns of gene expression indicative of neoplasticity. Thus, a separator $S = S(g_1, g_2, \ldots, g_K)$ is a discrete function of several inputs which takes the value 1 if the corresponding sample is in a neoplastic state and 0 otherwise. Using ternary data sets, the expression level of each gene can take one of three values, namely -1 (down–regulated), 0 (unchanged), or 1 (up–regulated). We will consider here the case when all samples are in the neoplastic state (i.e. S=1), and the down– or up–regulation is measured relative to an appropriate normal control. The analysis for the more general and complex case of both neoplastic (S=1) and normal tissue samples (S=0) will be treated elsewhere (Wahde and Szallasi, 2000). Let N denote the number of genes in each sample, M_- and M_+ the number of down– and up–regulated genes, respectively, and M their sum, i.e. $M = M_- + M_+$. The number of samples is denoted E. According to the assumptions above, the da-

ta contains examples of gene expression patterns for which $S = 1$. Clearly, any set of genes $(g_1, ..., g_K)$ for which there exists at least one sample such that $g_1 = g_2 = ... = g_K = 0$ *cannot* describe a separator, since some change in the expression levels is needed to arrive at the neoplastic state. Thus, the first step in identifying a separator of K inputs, is to find all combinations of K genes such that, in each sample, at least one of the K genes is down– or up–regulated. Any such combination of genes defines a separator. However, the high level of gene expression diversity in cancer samples makes it probable that separators can occur by chance even in the extreme case when gene expression patterns are generated by the random and independent selection of the mis–regulated genes.

Generative models

The probability of chance appearance of separators can be estimated by analytical tools only in relatively simple cases. For example, the accidental appearance of a single gene separator in a gene expression matrix produced by random and independent selection can be estimated by combinatorics (Wahde et al, 2001). However, in more complex cases, analytical calculations become intractable but computer simulations can still be used to obtain estimates of probabilities. The aim of a generative model is to produce an artificial data matrix which shares the essential characteristics of the original data matrix. The artificial data obtained by means of the generative model can then be used to form null hypotheses for the estimation of the probability of separators discovered in the real data set, thus making it possible to distinguish chance separators from actual separators.

Generative models can be derived from either theoretical considerations or empirical observations. In cancer research, theory–based generative models can use either genetic network modeling or aneuploidy driven gene mis–regulation as their starting point. Malignant transformation can be considered as an attractor transition of a self–organizing gene network (Kauffman 1993, Szallasi and Liang 1998) providing numerical estimates about the overall quantitative features of attractor transition like the expected number of up– or down–regulated (with a common term, mis–regulated) genes. There is an increasing evidence of the ploidy regulation of gene expression levels as well (Galitski *et al*, 1999). Thus, the aneuploidic distribution of chromosomes can also be used to model the expected gene expression patterns in cancer (Rasnick and Duesberg, 1999). At the current stage of theory and available data sets, however, we can best rely on generative models based on empirical observations. This approach starts with extracting overall quantitative features of cancer associated gene expression matrices. These include the number of genes that can be mis–regulated, the ratio of up– versus down–regulated genes and the level of co–regulation of mis–regulated gene groups.

In this paper, two methods for generating artificial data will be introduced and described. The first method simply forms a randomized gene expression matrix while preserving certain overall features of the real data matrix, such as the number of mis–regulated genes in each sample. Mutual information based generative models, which is the second method introduced here, preserve additional features of the real data, namely the co–regulation of genes. Note that we will use the terms generative model and generative algorithm interchangeably in this paper.

Randomization based models

As noted above, the gene expression diversity in cancer samples is so high as to make it probable that chance separators can occur, even in the case when gene expression patterns are generated by the random and independent selection of the mis–regulated genes. Such chance separators must be removed in order for the true separators to be discovered.

The simplest method of generating artificial data consists simply of inserting, for each sample, M_+ 1's and M_- -1's randomly in a null $N \times E$ matrix. In general, the values of M_- and M_+ will of course vary from sample to sample, so either an average value or the actual values of M_-^i and M_+^i $(i = 1, \ldots, E)$ from the real data can be used. It turns out that the formula for the expected number of separators is very sensitive to the values of M_-^i and M_+^i, and therefore the use of average values is not to be recommended. The randomization method that uses the actual values of M_-^i and M_+^i, will be referred to as **simple randomization**.

Consider first the case of $K = 2$ inputs. Assume that two genes, denoted g_1 and g_2, are being studied. In a given sample i, the approximate probability $p_s^i(2)$ of at least one of these two genes being changed (up– or down–regulated) is

$$p_s^i(2) = 1 - (p_0^i)^2, \tag{1}$$

where $p_0^i = (N - M^i)/N$ denotes the probability of a given gene being unchanged ($M^i = M_+^i + M_-^i$, where M_+^i and M_-^i denote, as before, the number of up– and down–regulated genes in sample i, respectively). Note that the approximation is valid as long as $1 << M^i << N$. In a typical neoplastic sample it is safe to make this assumption, since $\sim$10% of the genes are changed (i.e. $M^i \sim 0.1N$). The probability of at least one of the genes being changed in each of the E samples equals

$$P_s(2) = \prod_{i=1}^{E} p_s^i(2) \equiv \prod_{i=1}^{E} (1 - (p_0^i)^2). \tag{2}$$

Thus, the expected number of such separators is

$$N_s(2) = \binom{N}{2} P_s(2). \tag{3}$$

Generalizing these formulae, it is easy to see that the expected number of separators of K inputs is

$$N_s(K) = \binom{N}{K} P_s(K) \equiv \binom{N}{K} \prod_{i=1}^{E} p_s^i(K) \tag{4}$$

where

$$p_s^i(K) = 1 - (p_0^i)^K. \tag{5}$$

This analysis gives an estimate of the *total* number of separators of K inputs expected in a randomized artificial data set. Using similar methods, the approximate probability of

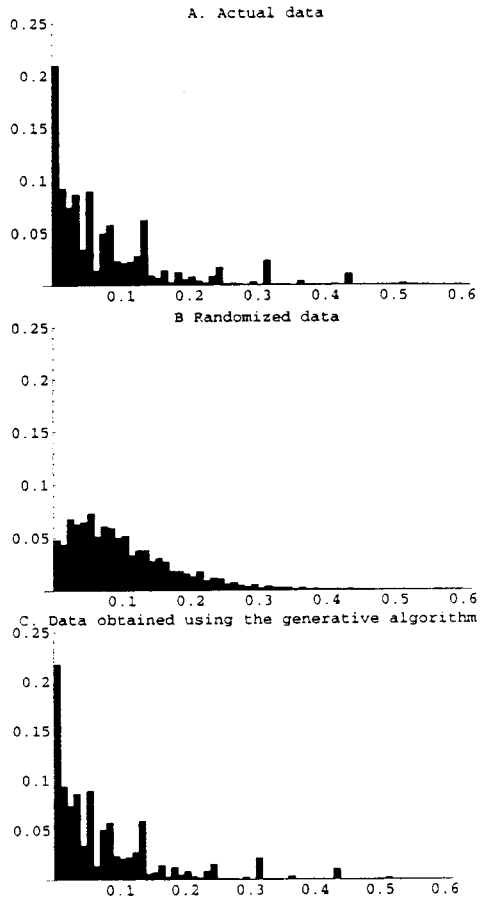

Figure 1: The distribution of pair–wise mutual information content. Panel A: Actual data set from Perou *et al.* (see text); B. Data randomized using simple randomization; C. Data obtained by running the generative algorithm.

discovering any specific separator in artificial data can also be obtained. In the case of K inputs, the total number of combinations of the input variables equals 3^K. The estimate of the probability of a specific separator begins by the computation of the probability, for one sample i, of obtaining one of those combinations for which $S = 1$. This probability is denoted p_R^i. The expected number of separators in the data set is then given by

$$N_R = \binom{N}{K} P_R \equiv \binom{N}{K} \prod_{i=1}^{E} p_R^i. \qquad (6)$$

As an example, consider a separator defined by the entries of Table 1. In any given sample i, the probability of having $S = 1$ equals

$$p_R^i = p_{-1,-1}^i + p_{0,-1}^i + p_{1,1}^i = (p_-^i)^2 + p_0^i p_-^i + p_+^i p_+^i, \quad (7)$$

g_1	g_2	S
-1	-1	1
-1	0	0
-1	1	0
0	-1	1
0	0	0
0	1	0
1	-1	0
1	0	0
1	1	1

Table 1: A $K = 2$ separator. The final column shows the value of the function S (the separator) for the given input configuration.

where $p_0^i = (N - M_-^i - M_+^i)/N$, $p_-^i = M_-^i/N$, and $p_+^i = M_+^i/N$. The approximate number of expected separators of this type is then

$$N_R = \binom{N}{2} P_R = \binom{N}{2} \prod_{i=1}^{E} p_R^i. \qquad (8)$$

Artificial data matrices obtained through simple randomization are, as we have seen, easy to handle analytically but not altogether realistic. For example, a real data matrix has a distribution of pair–wise mutual information which differs significantly from that of a matrix generated by simple randomization (Fig. 1). In particular, the randomized data generally lacks the spikes seen in the real data at high mutual information values. We now proceed to describe a generative model which does preserve the mutual information structure.

Mutual information based generative models

In the previous section we have discussed gene expression matrices in which a given number of gene mis–regulation appears by random and independent selection. In these matrices chance separators appear with a certain frequency that can be calculated as described above. This frequency, however, may significantly increase by restrictions on the selection of mis–regulated genes. Biological systems display the following two restrictions. First, not every gene can be mis–regulated. The number of changeable genes can be calculated as described elsewhere (Wahde et al, 2001) by conditional probabilities. Second, mis–regulated genes are not independently selected. Gene expression levels in cancer are determined by several factors, such as the regulatory input of other genes and the actual DNA–copy number of the given gene present in a cell (Galitski *et al*, 1999). This will obviously lead to a high level of interdependence between gene expression levels which is readily quantified by mutual information content. As we will be showing below, retaining the high level of mutual information content in a gene expression matrix will significantly influence the number of separators appearing by chance. Consequently, the aim of our generative model is to simulate gene expression patterns by randomly selecting the mis–regulated genes while retaining the actual size of the pool of changeable genes and also

their level of co–regulation as detected in actual cancer samples, and measured by the mutual information distribution of the gene pairs.

Algorithm The generative algorithm begins by generating a random data matrix R, by rearranging the matrix elements of the real data set D. A simple algorithm for arriving at a data set of this type is defined as follows: Loop through all genes. For each gene, loop through each sample, select randomly another sample, and swap the corresponding matrix elements. Note that, with this procedure, the values of M_{-}^{i} and M_{+}^{i} will change, since they are measured column-wise. However, since the computation of mutual information (see below) is based on comparison of genes (rows in the expression matrix), rather than samples (columns) this is the correct way to randomize the matrix in this case. This randomization method will be referred to as **permutative randomization**.

Once the permutative randomization has been performed a histogram of pairwise mutual information values is generated. A similar histogram is also generated for the real data set, and the distance between the two histograms is computed as

$$\Delta(H_G, H_D) = \frac{1}{N_{\text{bins}}} \sum_{m=1}^{N_{\text{bins}}} \frac{|H_G(m) - H_D(m)|}{\max(H_D(m), 1)}, \quad (9)$$

where N_{bins} is the number of bins in the histograms, for which the bin width thus equals $1/N_{\text{bins}}$. The algorithm then proceeds with the calculation as follows: A gene j is selected at random among the N genes, and its contribution to the histogram is computed by checking the pairwise mutual information between gene j and all other genes. The contribution of gene j to the histogram is subtracted, and the matrix elements in the corresponding row of the data matrix are rearranged, with probability p_{swap}, by the same swapping procedure as was used in the permutative randomization algorithm.

Then, the new contribution of gene j to the histogram is computed and the histogram thus obtained is compared with the histogram present before the rearrangement of gene j. If the distance is smaller than before the rearrangement, the new histogram (and, of course, the corresponding matrix) is kept. If not, the old matrix, and the old histogram, are retained. Thus, only improvements are kept, and the algorithm can be considered to be a simple implementation of an evolution strategy (Bäck *et al*, 1991). This procedure – selection of a random gene, subtraction from the histogram, partial rearrangement, formation of the new histogram, and finally selection of either the old or the new configuration – is repeated many times, until the distance between the histogram for the artificial data and that of the actual data is smaller than a user–defined critical value Δ_c. Usually, Δ_c was taken to be of order 10% of the initial distance between D and R.

A pseudo–code representation of the algorithm is given in Fig. 2. Normally, p_{swap} is given a large value in the beginning of a run, when the difference between the two histograms is large. The value of p_{swap} is then gradually lowered as the two histograms approach each other. There are

Perform permutative randomization and set $G = R$;

Compute mutual information distribution for G and D by going through all $N(N-1)/2$ gene pairs.

Set the number of bins to N_{bins} (and thus the bin width to $1/N_{\text{bins}}$) for the histograms (see below).

Compute the histogram H_D of pairwise mutual information content for the original data set D, using the mutual information data computed above.

Compute the histogram H_G of pairwise mutual information content for G using the mutual information data computed above.

Compute the difference Δ in mutual information content between G and D as follows:

$\Delta(H_G, H_D) = \frac{1}{N_{\text{bins}}} \sum_{m=1}^{N_{\text{bins}}} \frac{|H_G(m) - H_D(m)|}{\max(H_D(m), 1)}$.

Repeat

Pick a random gene j and compute the contribution $h_{G;j}$ of gene j to the histogram G.

Subtract $h_{G;j}$ from H_G to form H'_G:
$H'_G(m) = H_G(m) - h_{G;j}(m),\ m = 1, \ldots, N_{\text{bins}}$.

For row j of G, loop through all columns k of the matrix:

For each k, pick a random column l and, with probability p_{swap}, swap the matrix elements in the two locations: $Gj,k \leftrightarrow Gj,l$.

Let G' denote the resulting matrix, compute the new contribution $h'_{G';j}$ of row j to the histogram, and form $H_{G'}$:
$H_{G'}(m) = H'_G(m) + h_{G';j}(m),\ m = 1, \ldots, N_{\text{bins}}$.

Form the difference $\Delta(H_{G'}, H_D)$ according to the formula above.

if $\Delta(H_{G'}, H_D) < \Delta(H_G, H_D)$ **then**

Accept G': Set $G = G'; \Delta = \Delta(H_{G'}, H_D)$

else

Reject G' and thus retain G;

Until $\Delta < \Delta_c$.

Figure 2: The generative algorithm for obtaining artificial data with a given mutual information structure.

Table 2: The reduced Perou *et al.* data set, containing 1082 genes and 16 samples.

Sample	M^i_-	M^i_+	M^i
1	19	664	683
2	67	150	217
3	72	197	269
4	80	247	327
5	97	393	490
6	40	96	136
7	100	202	302
8	115	105	220
9	72	220	292
10	115	234	349
11	85	428	513
12	72	640	712
13	64	451	515
14	58	173	231
15	90	99	189
16	65	260	325

various ways of improving the algorithm, for instance by introducing adaptive control of the time variation of p_{swap}. However, even in its present simple state, the algorithm runs rather fast, and typical running times for a data set with $\approx$ 1000 genes and $\approx$ 15 samples are around 15–20 minutes on a computer equipped with a 550 MHz PIII processor.

Results

Generative model based analysis of breast cancer associated cDNA microarray measurements

In order to assess the relevance of generative models for estimating the frequency of chance separators, we have analyzed the breast cancer associated gene expression matrix published by Perou *et al.* (1999). This publicly available data set contains cDNA microarray based relative expression levels of about 5,600 genes for a number of both normal and neoplastic breast epithelial samples. For our analysis we have used only gene expression measurements derived from either breast cancer cell lines or primary breast tumors, 16 samples altogether. We have retained only those genes in our analysis that showed an at least 3.5–fold up– or down–regulation in at least two samples. Using these threshold values we have transformed the original data set into a 1082x16 ternary data matrix. A summary of this data set is given in Table 2.

The chance appearance of consistently mis-regulated genes, i.e. $K = 1$ separators, constitutes a special case which will be treated elsewhere (Wahde *et al.*, 2001). Here we are focusing on $K = 2$ separators. Applying Eq. 4, it is found that, for this data set, the expected number of separators assuming random and independent selection (i.e. using the simple randomization method) is 8.6. As a comparison, numerical simulations yield an estimate of 8.5 ± 7.7 separators (average of results obtained with 1,000 randomized data matrices).

However, the actual number of separators, obtained from the real data set, equals 16,997. Clearly, a comparison with the randomized data matrix would indicate that this is a very significant number indeed. Comparing, however, with the results obtained using the generative algorithm ($\sim$ 40 independent simulations), the result is very different. In this case, the average number of expected separators equals $25{,}417 \pm 947$.

The high number of expected separators indicate that the 16 samples contained in this data set are not enough to validate the presence of a real $K = 2$ separator. This was obviously not the purpose of our current analysis. At this initial level of analysis we needed an appropriate data set in order to estimate the impact of generative models.

The distribution of mutual information provides a useful visual aid to assess the overall data structure of gene expression matrices. Panel A of Fig. 1 shows the distribution of pair–wise mutual information content of the ternary data set derived from large–scale, cDNA microarray based gene expression measurements of breast cancer samples (Perou *et al.*, 1999). Each vertical bar shows the fraction of the gene pairs whose pairwise mutual information falls within the corresponding interval, of width 0.01. The randomized version of the same data set is shown in panel B, and panel C shows a representative simulated data matrix created by the generative algorithm defined in Fig. 2. Genes that are co–regulated in cancer will display a high mutual information content. Randomization will destroy the effect of co–regulation on the data set and gene pairs with high mutual information content are unlikely to be present. Therefore, the distribution of mutual information will not contain spikes at high mutual information values as demonstrated by panel B versus panel A. The generative algorithm, however, recreates the basic data structure of the gene expression matrix. Therefore, its mutual information distribution will be more similar to that of the original data (panel A). A histogram of the distribution of the number of separators obtained from the generative model is shown in Fig. 3.

Further analysis

An analysis similar to the one reported above was performed for two other data sets as well.

Analysis of a gene expression matrix derived from alveolar rhabdomyosarcoma samples We have also analyzed the gene expression data published by Khan *et al.* (1998). This data set consists of 13 samples altogether, seven of them alveolar rhabdomyosarcoma samples and the rest commonly used human cancer cell lines. The data matrix contained ternary expression information for 1248 genes.

The actual number of separators for this data set was found to be 16,124. Using Eq. 4, an estimate of $0.017 << 1$ separators was obtained, again much lower than the actual value. Using instead the generative algorithm, an average of $17{,}252 \pm 133$ separators were obtained.

Analysis of colon cancer associated gene expression measurements DNA–oligomer chip based gene expression

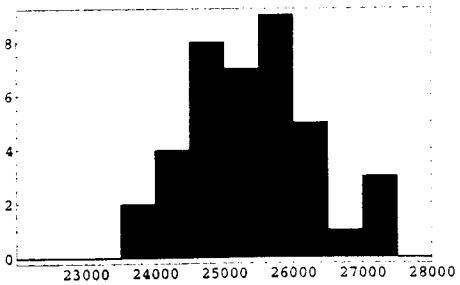

Figure 3: Histogram of the expected number of separators obtained using generative models for the reduced Perou *et al.* data set.

measurements were published on 2,000 genes in 22 patient matched neoplastic and normal colon samples by Alon *et al.* (1999).

According to Eq. 4, increasing the sample number (in this case to 22) decreases the expected number of separators appearing by chance. Indeed, applying this equation, the expected number of separators assuming random and independent selection is found to be $2.3 \times 10^{-12} <<< 1$. On the other hand, the actual number of separators with $K = 2$ was equal to 1 for this data set, suggesting that this separator might play a role in colon cancer. This assumption, however, must be reevaluated after applying the mutual information based generative models. If the essential structure of the colon cancer associated gene expression matrix is retained then the expected number of separators is increased by twelve orders of magnitude to 3.7 ± 1.4. This result puts into question the significance of the separator found in the data. This doubt was reinforced by the fact that neither gene involved in the separator has any documented involvement with any forms of human cancer.

Discussion

Successful analysis of cancer associated gene expression matrices will require a profound understanding of the data structure. In this paper, we have pointed out that statistical analysis ignoring the data structure characteristic of biology can be rather misleading, producing errors of several orders of magnitude. Here, we have introduced generative models that will simulate a large number of random gene expression matrices while retaining the empirically detected level of gene co-regulation and the number of changeable genes.

We note that two of the data sets used here contained rather few samples, and thus a large number of separators was found for both data sets. With more samples, the vast majority of the false separators would disappear, leaving us (at best) with a few separators, as was found for the last data set (Alon *et al.*, 1999). It is interesting to note that, despite the large variation (between the data sets) in the number of separators, artificial data matrices based on simple randomization underestimate the number of separators by several orders of magnitude, whereas artificial matrices obtained from the generative algorithm tend to overestimate the number of separators, but by a much smaller amount. However, this result could be a chance occurrence, and it certainly needs to be investigated further, using a larger ($\sim 10^3$) number of artificial data matrices than the 10–40 or so that were used here. Such a validation will be the next step of our analysis.

If the analysis would indeed confirm that the number of separators expected on the basis of the results form the generative algorithm is larger than the number of separators in the actual data, this would indicate that the data sets studied do contain interesting information worthy of further study.

Determining the exact impact of generative models will require a thorough analysis. For example, clustering algorithms are routinely used to identify significant patterns in cancer associated gene expression matrices. The reliability of these results is routinely evaluated by running the same algorithm on a randomized gene expression matrix (Alon *et al.*, 1999). The validity of this approach is highly questionable in light of our initial results. Generative models can easily be extended to continuous data by e.g. replacing mutual information analysis with the absolute value of the Pearson correlation coefficient. Then these "continuous generative models" could serve as reference points to estimate the chance appearance of clusters.

Generative model based analysis of discrete gene expression matrices, however, has one major advantage over continuous models. It can easily incorporate the often used qualitative parameters such as the histological phenotype of a tumor. This would suggest that discrete and continuous generative models ought to be developed in parallel in order to accommodate the different types of data sets produced by cancer research.

Another important improvement, for which work is under way, is to optimize the simulation program in order to accommodate larger gene expression matrices that will require many simulations in order to provide certainty that a given separator did not appear by chance at a given confidence level.

Furthermore, it would be of interest to obtain theoretical and simulation based estimates of the minimum size of gene expression matrices that will have a low frequency of accidental separators. This will, in turn, set the guidelines for selecting the correct sample size that will allow powerful statistical analysis.

Conclusion

We have shown that an uncritical application of a null hypothesis based on artificial data obtained through simple randomization will underestimate, by several orders of magnitude, the number of separators found in gene expression matrices.

In order to obtain a more useful null hypothesis, we have introduced a generative algorithm which produces artificial data matrices that retain the pairwise mutual information structure of the original data. We have shown that, in the light of the results obtained using the generative algorithm, the separators found in the data sets used here may be chance occurrences, rather than actual indicators of neoplasticity.

References

Alon U., Barkai N., Notterman D.A., Gish 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 USA* 96(12):6745–50

Bäck T., Hoffmeister F., and Schwefel, H.–P. 1991. A survey of evolution strategies. In: Belew, R.K., editor, *Proceedings of the Fourth International Conference on Genetic Algorithms and their Applications*, 2–9, San Diego, California, USA: Morgan Kaufmann Publishers.

Chen, Y., Dougherty, E.R., and Bittner, M.L. 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* 2:364–374.

Galitski, T., Saldanha, A.J., Styles, C.A., Lander, E.S., and Fink, G.R. 1999. Ploidy regulation of gene expression. *Science* 285:251–254.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., 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(5439):531–7.

Kauffman, S. *The origins of order*. 1993. Oxford: Oxford University Press.

Khan J., Simon R., Bittner M., Chen Y., Leighton SB., Pohida T., Smith P.D., Jiang Y., Gooden G.C., Trent J.M., and Meltzer P. 1998. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* 58(22):5009–13

Perou C.M., Jeffrey S.S., van de Rijn M., Rees C.A., Eisen M.B., Ross D.T., Pergamenschikov A., Williams C.F., Zhu S.X., Lee J.C., Lashkari D., Shalon D., Brown P.O., and Botstein D. 1999. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA* 96(16):9212–7.

Rasnick, D., and Duesberg, P.H. 1999. How aneuploidy affects metabolic control and causes cancer. *Biochem J* 340:621–630.

Szallasi, Z. 1998. Gene expression patterns and cancer. *Nature Biotech* 16:1292–1293.

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. Biocomp.* 3:66–76.

Wahde, M. Klus, G.T., Chen, Y. , Bittner, M.L., and Szallasi, Z. 2001. Assessing the significance of consistently mis-regulated genes in cancer associated gene expression matrices. (submitted to *Pac. Symp. Biocomp.*).

Wahde, M and Szallasi, Z. 2000. The diversity of normal gene expression patterns can be exploited to increase the power of the statistical analysis of cancer associated gene expression matrices. (manuscript in preparation)