AD_____

Award Number: DAMD17-01-1-0532

TITLE: Generative Model Based Statistical Analysis of Gene Expression Patterns in Breast Cancer

PRINCIPAL INVESTIGATOR: Zoltan Szallasi, M.D.

CONTRACTING ORGANIZATION: Henry M. Jackson Foundation for the Advancement of Military Medicine
Rockville, Maryland 20852

REPORT DATE: August 2002

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 074-0188*



| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | August 2002 | Annual (1 Aug 01 - 31 Jul 02) |

**4. TITLE AND SUBTITLE**
Generative Model Based Statistical Analysis of Gene Expression Patterns in Breast Cancer

**5. FUNDING NUMBERS**
DAMD17-01-1-0532

**6. AUTHOR(S)**
Zoltan Szallasi, M.D.

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Henry M. Jackson Foundation for the Advancement of Military Medicine
Rockville, Maryland 20852
**E-Mail:** zszallasi@chip.org

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES**

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for Public Release; Distribution Unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** ***(Maximum 200 Words)***

Microarray analysis provides an efficient unbiased strategy to identify differentially expressed genes in breast cancer.The statistical analysis of large-scale gene expression studies, however, imposes several serious novel challenges. Most importantly, gene expression arrays have a well-defined internal data structure dictated by the genetic network of the living cell. We showed that ignoring this data structure leads to errors of several orders of magnitudes in the statistical analysis. The consequence of this is either producing false leads for experimenters or eliminating truly important leads in cancer research. We have introduced two methods to overcome this problem. The first is based on generative models that produce random data sets while retaining the overall level of gene co-regulation as reflected in the distribution of pair-wise co-regulation measures. The second method is an information theoretic approach based on the theory of RxC contingency tables. This latter method also deals with the lack of replicates often encountered in cancer genomics. These methods determine the probability that a given feature, such as a cluster or separator, will appear by chance in a gene expression array.

**14. SUBJECT TERMS**
statistical analysis, gene expression matrix, generative models, classifying breast cancer, RxC contingency table

**15. NUMBER OF PAGES**
30

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | Unlimited |

## Table of Contents

INTRODUCTION:

Microarray analysis provides an efficient unbiased strategy to identify differentially expressed genes in breast cancer.The statistical analysis of large-scale gene expression studies, however, imposes several serious novel challenges. Most importantly, gene expression arrays have a well-defined internal data structure dictated by the genetic network of the living cell. Standard analytical tools often ignore this structure and, as we shown recently (1), this may lead to errors of several orders of magnitudes in the statistical analysis. The consequence of this is obvious: either the statistical analysis might be too "lenient" producing false leads wasting experimental effort, or the analysis might be too strict eliminating truly important leads in cancer research. The solution is creating statistical analytical tools that will take into consideration the internal data structure of cancer associated gene expression measurements. This will in turn determine the probability that a given feature, such as a cluster or separator, will appear by chance in a gene expression array.

BODY:

First, we developed tools to characterize the internal data structure of cancer associated gene epxression matrices. In order to ensure fast implementation, in our first approach we used discretized data. (We are currently modifying our tools for continuous data. See below) We developed a program to calculate the pair-wise mutual information distribution of genes or samples and also to determine significant relevance networks. (1,3) We showed, that the mutual information distribution of the real data and randomized data sets are very different indeed (1). We introduced the concept of a "separator" which is a set of genes coupled by an appropriate set of rules that can disinguish between two phenotypes (e.g. cancer and normal). We developed a program that could search for separators exhaustively in a gene expression matrix up to complexity level of 5 ( i.e. the combination of 5 genes). We determined both theoretically and by Monte-Carlo simulations the expected number of separators in a completely random data set (1). Then, we created a simulation-based tool that generates random data sets while retaining the overall pair-wise mutual information distribution that is detected in the actual breast cancer associated measurements (the concept is described in details in publication 1). After creating and analyzing a large number of such data sets we could estimate the probability of chance appearance of separators in both the completely random data sets and the one that retained the overall data structure of the original gene epxression matrix. We found, that ignoring this data structure, the probability of chance appearance of separators can be underestimated by orders of magnitudes, leading to a large number of false leads for the expreimenters.

Discretizing gene expression data (i.e. working with up- or down-regulation, or no change instead of the actual level of gene expression ) is a fequently used method in microarray analysis. After establishing the significance of these calls, one can ignore the effect of noise of measurements in later steps of numerical analysis.
However, the risks of discretization for statistical analysis are also well known, therefore we decided to further develop our statistical methods for the analysis of continuous data. We set out on two theoretically different paths, with the second approach providing additional benefits considering the unique characteristics of microararay based data.

First we created an evolutionary-algorithm based tool for the molecular classification of cancer. So far, we have implemented a version of the algorithm that is limited to finding classifiers consisting of only four (or fewer) genes. For this initial version we have also set some limitations on the mathematical operators that relate the four genes. Our method has been applied to a breast cancer derived data set recently published by van't Veer et al (5). This contains the relative expression level of 25,000 genes in 98 breast cancers. This data set was used by the authors to discriminate between different patients who developed distant metastases within 5 years and patients who remained disease free after a period of 5 years. They determined a set of seventy genes that individually showed best correlation with metastasis-free survival and concluded that based on the combination of these genes one can provide a classification with 83% accuracy. Our evolutionary based algorithm produced a four-gene classifier that could predict metastasis free survival with 91% accuracy (6) . This very promising result achieved with relatively simple computational tools is a strong incentive for further development of our method. We are currently developing an algorithm that will rank the possible classifiers by their noise tolerance. We are also implementing a test that will determine whether a given classifier could have arisen in a given gene expression matrix purely by chance. We are searching for the simplest, noise tolerant classifiers that are statistically significant. The impact of our approach to extract the simplest classifiers is obvious for molecular diagnostics. Fewer genes can be more accurately and more sensitively measured by e.g. QRTPCR than a set of genes on the order of one hundred that require parallelized methods such as microarray technology.

Our second approach deals with two unique issues in the statistical analysis of microarray measurements: first, retaining the internal data structure as described above; second, cancer associated gene expression matrices often produce a single set of measuremens per tumor sample. Results in these experiments, such as disease classifiers or novel tumor sub-classes, manifest as a sub-array of the expression levels of a selected subset of genes in a selected subset of samples Having only single measurements requires the introduction of information theoretical approaches leading to the question: What is the likelihood, that such a sub-array, which we call a "feature", carries non-zero information? We gave an ab initio derivation of probabilities for features in microarray expression measurements and precise algorithms to quantify the information content of such features. This is accomplished by defining precise ``null hypotheses", formulated in the context of appropriate disproportion measures, which are quantitative measures of the association between rows and columns in an array, and within the context of appropriate ensembles associated with the experimental microarray expression data set. A given null hypothesis for a selected feature asserts that the value of disproportion measured experimentally for that feature could have occurred by chance in the ensemble considered. Therefore, when the probability of validity of the null hypothesis is small, the feature is unlikely to have appeared by chance, and therefore carries non-trivial information. Our method is robust, unbiased and makes no assumptions about underlying uncertainties in the expression data set. This method is a novel tool for validating diagnostic marker genes in gene expression matrices and is also helpful in determining the number of samples required to validate results extracted from microarray measurements. These tools will be made available to the research community soon.

KEY RESEARCH ACCOMPLISHMENTS:

1) We have shown that cancer associated gene expression matrices have a well-defined internal data structure, as reflected in e.g. the mutual information distribution of genes. We have demonstrated, that if this data structure is ignored then the probability of the chance appearance of separators in these data sets can be underestimated by several orders of magnitudes.

2) In order to avoid this error we have created a simulation based tool that can generate random gene expression matrices that retain the internal data structure as reflected in the overall pair-wise mutual information distribution of genes.

3) We further developed our tool to handle continuous data as well. We have introduced an information theoretic approach for the statistical analysis of cancer associated gene expression matrices. This deals with the unique data structure of these measurements and with the lack of replicates at the same time

REPORTABLE OUTCOMES:

- Algorithm to extract gene expression separators from discretized data sets (1).
- Estimating the internal data structure of cancer associated gene expression measurements by mutual information distribution. (1)
- Estimating the probability of chance separators in randomized and real gene expression matrices (1,2).
- Introduction of information theoretical approaches for the statistical analysis of gene expression matrices (4).

CONCLUSIONS:

Ignoring the internal data structure of cancer associated gene expression matrices by using completely randomized data sets as a statistical control may severely mis-estimate the statistical significance of microarray based results. We are introducing two methods to overcome this problem: the first is generative models that produce random data sets while retaining the overall level of gene co-regulation as reflected in the distribution of pair-wise co-regulation measures. The second method is an information theoretic approach based on the theory of RxC contingency tables. This latter method also deals with the lack of replicates often encountered in cancer genomics.

REFERENCES:

1) Wahde, M. and **Szallasi Z.** Generative model based analysis of cancer associated gene expression matrices. Proceedings of the First International Conference on Systems Biology:39-45. 2001.

2) Wahde, M., Klus, G., Bittner, M., Chen, Y and **Szallasi, Z.** Assessing the significance of consistently mis-regulated genes in cancer associated gene expression matrices. Bioinformatics 18:389-394, 2002.

3) Klus, G.T., Song, A., Schick, A., Wahde, M. and **Szallasi, Z.** Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer-associated

differential gene expression patterns. Pacific Symposium on Biocomputing. , 6:42-51, 2001.

4) Periwal, V., Miller, R.E. and **Szallasi, Z.** The Importance of Being Improbable: An Information theoretical approach to validate statistically significant results in gene expression measurements ( in preparation)

5) van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**:530-366

6) Wahde, M. and **Szallasi , Z.** Evolutionary algorithm based classification of cancer associated gene expression matrices. (in preparation)

# MUTUAL INFORMATION ANALYSIS AS A TOOL TO ASSESS THE ROLE OF ANEUPLOIDY IN THE GENERATION OF CANCER-ASSOCIATED DIFFERENTIAL GENE EXPRESSION PATTERNS

GREGORY T. KLUS[@], ANDREW SONG[@], ARI SCHICK[@], MATTIAS WAHDE[%] and ZOLTAN SZALLASI[@]

*[@]Department of Pharmacology, Uniformed Services University of the Health Sciences,Bethesda, MD, [%]Div. of Mechatronics,Chalmers University of Technology, Göteborg, Sweden*

(reprint requests at *zszallas@mxc.usuhs.mil)*

Most human tumors are characterized by: (1) an aberrant set of chromosomes, a state termed aneuploidy; (2) an aberrant gene expression pattern; and (3) an aberrant phenotype of uncontrolled growth. One of the goals of cancer research is to establish causative relationships between these three important characteristics. In this paper we were searching for evidence that aneuploidy is a major cause of differential gene expression. We describe how mutual information analysis of cancer-associated gene expression patterns could be exploited to answer this question. In addition to providing general guidelines, we have applied the proposed analysis to a recently published breast cancer-associated gene expression matrix. The results derived from this particular data set provided preliminary evidence that mutual information analysis may become a useful tool to investigate the link between differential gene expression and aneuploidy.

Most human tumors display a set of well-defined aberrations at different levels of cellular biology and biochemistry. These include numeric chromosomal imbalance, termed aneuploidy[1], mutations in various genes, and an abnormal gene expression pattern[2]. One of the main aims of cancer biology is to find the causative relationship between these aberrations. Beyond scientific curiosity, understanding the link between these changes detected in tumors may have a profound impact on cancer therapy as well. If the abnormal gene expression patterns found in tumors were in fact a direct result of aneuploidy, then reversal of aneuploidy might be able to return tumor cells to a more normal gene expression pattern and phenotype, and therapies based on this approach should be investigated.

With the availability of data from the Human Genome Project specifying the various genes on each chromosome, it should now be rather straightforward to establish whether or not an extra chromosome or the loss of a chromosome is reflected in higher or lower expression levels of the genes present on that chromosome. For example, there are cases of pediatric acute lymphoblastic leukemia in which the sole karyotypic change is chromosome 5 trisomy[3]. In these cases the relative expression levels of the genes localized on this chromosome should be increased and this could be readily measured. However, the karyotype of most tumors is significantly more complex and the ploidy regulation of gene expression is likely superimposed by other regulatory mechanisms. Therefore, proving that differential gene expression patterns detected in cancer are generally induced by aneuploidy will probably involve a more complicated analysis of large-scale gene expression and karyotype databases.

The aim of the current paper is to describe a mutual information-based analytical framework for such an analysis, and to perform the first such analysis on a publicly available data set of breast cancer-associated gene-expression changes.

**The causes of differential gene expression in cancer**: Differential gene expression patterns in cancer result from the superimposition of the following three mechanisms:

1) Extra or missing chromosomes or chromosome regions (segmental aneuploidy). It is obvious that the often-detected complete loss of a given chromosomal region from a cell is reflected in the complete down-regulation of the genes present in that region. It is also well-known that increased copy number of a gene, called DNA amplification or the multiplication of a chromosomal region directly causes up-regulation of gene expression (see for example[4,5])

2) Many oncogenes act as transcription factors themselves or have a well-characterized direct effect on other downstream transcription factors. When these oncogenes (e.g., myc, src and ras) are overexpressed or mutated, they directly or indirectly change the expression level of several other genes[6].

3) The genetic network of a cell with a stable phenotype is self-consistent. In other words, the expression level of each gene is consistent with the expression level of its regulatory inputs. The very existence of cancer-associated differential gene expression proves that the genetic network of a given cell has several alternative stable states. These states are often called attractors in genetic network theory[7], and during malignant transformation the cell is induced to undergo attractor transition. It was also hypothesized, although never proved experimentally, that the cells can reach these alternative attractors after major perturbations of the genetic network, without the continued presence of oncogenes or aneuploidy. This idea is partially supported by the so called hit and run mechanism, when after malignant transformation the causative oncogene (e.g., ras) is lost but the cell still remains in its neoplastic state[8,9]. (It should also be noted that there are examples of reversible malignant transformation, when the cells revert to their non-malignant state after the overexpression of the causative oncogenes has been turned off[10,11].)

**General analytical framework in order to establish aneuploidy as a major mechanism inducing cancer-associated gene expression patterns:** If aneuploidy is its main driving mechanism, then differential gene expression in cancer will be induced as follows: First a group of genes will be up- or down-regulated due to chromosomal gain or loss. Then this aneuploidy induced gene expression pattern will be adjusted by the regulatory functions of the genetic network present in the cell, keeping the network consistent with the gene regulatory rules.

This hypothesis assumes that the genes present on the same chromosome or chromosome region will be often mis-regulated in the same tumor samples, showing a certain degree of co-regulation in gene expression measurements performed on a sufficiently large number of cancer samples. The level of co-regulation can be readily quantitated by simple means such as calculating the

Pearson correlation coefficient in continuous gene expression measurements[12]. In this paper, however, we propose to use mutual information instead of correlation coefficient (mutual information can be considered as a discretized form of the absolute value of correlation coefficients[13]) for two reasons. First the precision of massively parallel gene expression measurements is limited. Second, the degree of up- or down-regulation which can be expected to result from aneuploidy is not known. Thus, currently it is more informative to trinarize the data, classifying each gene as either unchanged or up- or down-regulated, rather than attempt to weight it with the ratios of mis-regulation. Trinerization can be readily performed after self-normalization of large-scale gene expression matrices as described by Chen et al[14].

**Proposed analytical framework:**

1. Take a cancer-associated gene expression matrix that was derived from a series of tumor samples of the same type (e.g. a set of primary mammary carcinomas) as population- and time-averaged gene expression data. Convert these data into a ternary matrix at an appropriate confidence level.
2. Calculate pair-wise mutual information for all gene pairs and create relevance networks of co-regulated genes with a mutual information level that is above the highest level detected in the gene expression matrix after randomization (i.e. above a threshold mutual information that can be still due to chance.)
3. Determine the chromosomal localization of the genes of the relevance network and compare it to the chromosomal distribution due to chance. This is determined by simulations assuming that co-regulated genes are randomly assigned to chromosomes.
4. If there are any relevance networks that show an unexpected clustering of genes located on the same chromosome, compare them to aberrations reported for that chromosome.

We will provide detailed description of the steps of this algorithm below, using a concrete breast cancer-associated gene expression matrix.

A **complete analysis** will require several complementary data sets:

1) A large body of gene expression measurements on a given type of cancer. The size of this data matrix is defined by the possible number of chromosome combinations or karyotypes associated with that type of cancer.

2) A catalog of the possible karyotypes of a given cancer. It is well established, that certain gains or losses of chromosomal regions or of whole chromosomes are frequently observed in a certain type of cancer, whereas others never occur[15]. The potential number of major karyotypes is an important reference point in this analysis: if there is a high number of potential configurations of aneuploidy then the number of required gene expression measurements will be proportionally higher.

3) The complete catalog of chromosomal localization of genes involved in the analysis, which will be soon available with the human genome project nearing completion.

A large number of studies on the karyotypes of cancer indicated, that certain chromosomal aberrations are often associated with a certain type of tumor, whereas others are never observed. (See for example[15]). This is also true for mammary tumors[15-18]. In this paper we were looking for relative enrichment of certain chromosomes in high mutual information relevance networks derived from a breast cancer associated gene expression matrix.

**Mutual information analysis of a breast cancer-associated gene expression matrix:** We have analyzed the breast cancer-associated gene expression matrix recently published by Perou *et al.*[2]. This publicly available data set contains cDNA microarray based relative expression levels of 5,584 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 converted the continuous gene expression data into a ternary matrix, using a 2-fold up- or down-regulation as a threshold value. The ternary representation is justified by the current, relatively limited precision of massively parallel gene expression measurements and the fact that we have no estimates about the expected level of up- or down-regulation of gene expression induced by aneuploidy. The exact karyotype of these tumors have not been reported, but it is well known that most sporadic breast tumors have a chromosome set which is far from normal diploid[15-18]. The breast cancer cell lines included in the analysis are also known to have a highly aneuploidic karyotype[19].

In a recent technical paper[20] we have pointed out that the overall quantitative features of cancer-associated gene expression matrices show several consistent characteristics. Namely, the number of mis-regulated genes and the ratio of down-regulated versus up-regulated genes are not arbitrary but remain within a well-defined range for a given type of tumor. This data set had a high level of gene expression diversity. On average, 35% of all quantitated genes were mis-regulated in each sample. The high level of gene expression diversity was reflected in the high level of mutual information content of the data matrix even after randomization. It is also interesting to note, that the breast cancer samples examined here showed significantly more down-regulation than up-regulation of genes. In fact 13 out of 16 samples had more down- than up-regulated genes relative to normal, and in 10 out 16 samples the down-regulated genes outnumbered the up-regulated ones by 3 to 1.

**Mutual information analysis:** We have calculated mutual information for all possible gene pairs as described in Butte *et al.*[13] and Liang *et al.*[21] with appropriate modifications. For simplicity we kept the range of mutual information between 0 and 1 by using base 3 logarithm for the ternary data set. Therefore the entropy of the mis-regulation for a single gene was calculated as follows:

$$(1) \qquad H(A) = -\sum_{i=1}^{3} p(x_i) \log_3(p(x_i))$$

where $p(x_i)$ is the frequency based probability that gene A will take the value of $x_i$ (i=1,..3) out of the three possible states of 0 (no change), 1 (up-regulation) or -1 (down-regulation). The mutual information for gene pairs A and B is defined as

(2) $MI(A,B) = H(A) + H(B) - H(A,B)$

**Randomization of the data matrix:** We needed to establish a threshold mutual information level (recently termed and abbreviated as TMI by Butte et al.[13]) above which we considered two genes being co-regulated. Random distribution of 1's, 0's and -1's in a matrix will lead to a certain level of background MI distribution. This is routinely assessed by randomizing the gene expression matrix and then recalculating the pair-wise MI for all gene pairs. We have performed permutative randomization on the gene expression matrix as described in Wahde and Szallasi[22]. This will randomize 1's, 0's and -1's within each row and will retain the average number of mis-regulated genes in the data matrix. The high number of mis-regulated genes of this data matrix predicted a high level of background MI level. Indeed, as demonstrated on Figure 1, after randomization there were several gene-pairs with a pair-wise MI level of up to 0.75. Therefore we have set TMI at this level.

Mutual information analysis, matrix randomization and graphic representation was implemented in Borland Delphi 3. The computation time for calculating the pair-wise mutual information for the complete 5584x16 matrix is about 3 min.

**Calculating the chance chromosomal distribution of relevance networks:** In an ideal case to prove the involvement of aneuploidy in differential gene expression patterns, one would expect fully connected relevance networks with high mutual information content where all or most genes are localized on the same chromosome. However, these ideal clusters will be "diluted" by the superimposed effect of gene co-regulation and by the fact that certain chromosomal aberrations occur together with higher frequency. On the other hand, if differential gene expression is driven by gene co-regulation with no ploidy effect at all, then one would expect that the genes present in high mutual information clusters, if they exist at all, would be nearly randomly distributed among all chromosomes. This latter assumption has formed the null hypothesis of our statistical analysis. We determined the likely distribution of chromosomal assignments within each relevance network assuming that those genes are randomly localized on chromosomes. Since the exact number of genes on each chromosome has not been determined yet (with the exception of Chr. 21 and 22), we have assumed that the number of genes/chromosome is proportional to the size of the chromosomes measured in megabases. (These data can down-loaded from the web site of National Center for Biotechnology Information at www.ncbi.nlm.nih.gov/.) Human chromosomes vary in size between 263 Mb (Chr. 1) and 47.7 Mb (Chr. 22). Therefore, we assumed that a gene in a relevance network will be assigned with

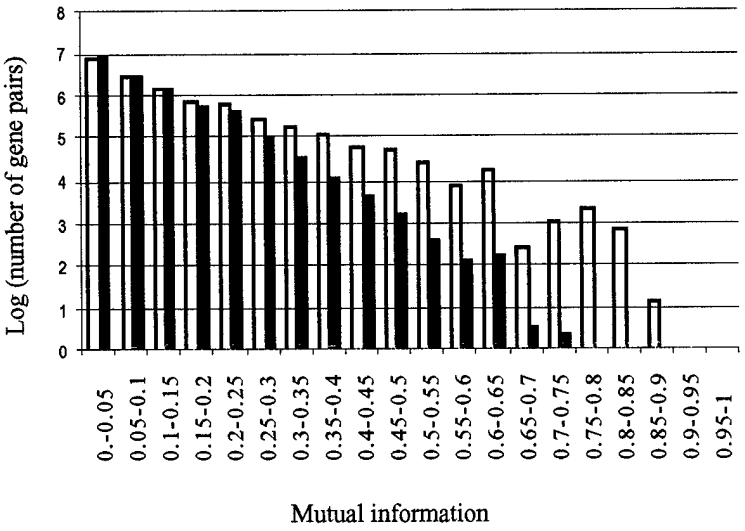


Figure 1 The distribution of mutual information amongst all possible gene-pairs for the actual data set (open columns) and for the average of ten randomized data sets (filled columns). The randomization of the gene expression matrix and the calculation of pair-wise mutual information for all possible gene pairs were performed as described in the text.

about 5-fold higher probability to e.g. Chr. 1 than to Chr. 22. In other words, we assumed that in the absence of ploidy regualation the probability that a given gene is present in a given relevance network will be proportional to the size of the chromosome on which the gene is localized. This assumption will become more accurate as more information becomes available from the human genome project.

We have implemented the following simulation in Matlab: we set the simulated cluster size, i.e. number of genes, to a given detected relevance network of high mutual information (see table 1). Then we have randomly assigned the genes of that cluster to chromosomes in such a way that the probability of assignment was proportional to the size of the chromosome. Finally, we have calculated how frequently we have seen a chromosomal distribution similar to the one observed in the relevance networks derived from the original data set. For each relevance network we ran 1000 simulations and determined whether at 99% confidence level the detected chromosomal distribution is due to chance.

**Summary of findings:** We have identified 65 relevance networks at a TMI level of 0.75. The majority of these were small clusters, namely 35 gene pairs and 16 gene triplets. None of the gene triplets were localized on the same chromosome. Preliminary analysis suggested, that it is likely (>10% chance) that two genes in a

relevance network of three genes will be localized on the same chromosome. Therefore, further examination of these small clusters was not informative. We have identified 14 relevance networks with more than 3 genes. The chromosomal localization of each gene was determined by a sequence-based BLAST search against the human genome data-base maintained by NCBI. (Available at http://www.ncbi.nlm.nih.gov/genome/seq). This has ensured that the chromosomal localization of the actual gene probes were determined even if a given microarray probe carried the wrong gene identification. The chromosomal distribution of the genes of these networks is listed in Table 1. All relevance networks were fully connected at a MI>0.75 level. 13 out of the 14 relevance networks showed chromosomal distributions that could be caused by chance (at 99% confidence level) assuming the random chromosome assignment described above.

Relevance network #3, however, displayed significant "enrichment" of genes originating on three chromosomes. This relevance cluster of 13 genes contained four genes from chromosome 17, three genes from Chr. 1, and two genes from Chr. 11, and the remaining four genes were from different chromosomes. This distribution of chromosomal assignment is unlikely due to chance at a 99% confidence level. It is well documented that chromosomes 1, 11 and 17 belong to the group of chromosomes that show numerical aberration with the highest frequency in breast cancer[15-18]. These chromosomes often show numerical changes together[15-18]. It is also known that loss of heterozygosity involving these chromomes is frequently detected in these tumors, and these chromosomes are more often lost than gained in breast cancer[15-18]. These data showed excellent correlation with the fact that the mis-regulation of genes involved in this relevance network represented mainly down-regulation. (The genes present in this relevance network were down-regulated in 8 tumors, up-regulated in one tumor and unchanged in 7 samples.) In this case, the relevance network gave a very good indication of the abnormal behavior of chromosomes associated with it.

**Discussion:** In this paper we have introduced mutual information analysis as a tool to establish a causative link between aneuploidy and differential gene expression in cancer. The limited sample number of the available gene expression data in breast cancer and the lack of a comprehensive database of karyotypes has obviously limited our analytical efforts at the moment. Nevertheless, in one case our analysis turned up a large relevance network of high mutual information in which the genes' chromosomal assignment was non-random. Furthermore, the three chromosomes highly represented in this relevance network (Chr. 1, 11 and 17) have been reported to show coordinated numerical aberrations in breast cancer[15-18]. These chromosomes are often lost which corresponds well with the frequent coordinated down-regulation of these genes in the breast cancer associated gene expression matrix examined.

The fact that only one out of fourteen relevance networks showed signs of involvement of aneuploidy suggests that chromosomal aberrations may play a limited role in the differential gene expression detected in breast tumors. However, the relevance network with non-random chromosomal assignment provide a

preliminary proof of principal and suggest a wider application of mutual information for this type of analysis.

**Abbreviations:** Chr.; Chromosome, TMI: threshold mutual information, MI: mutual information

**Acknowledgment:** The opinions and assertions contained herein are the private opinions of the authors and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences or the U.S. Department of Defense.

| Relevance Network | Chromosomes represented by 1 gene | 2 genes | 3 genes | 4 genes |
|---|---|---|---|---|
| #1 17 genes (2 unknown) | 1,2,3,4,5,8, 12,13,15,16 | 17 | 10 | |
| #2 15 genes (1 unknown) | 1,2,7,8,14, X | 2,9,18,19 | | |
| #3 13 genes | 2,4,5,10 | 11 | 1 | 17 |
| #4 11 genes | 4,6,13,14, 19,20,X | 2,17 | | |
| #5 10 genes | 2,3,4,6,10, 12 | 5,15 | | |
| #6 10 genes ( 2 unknown) | 2,11 | 3,4,10 | | |
| #7 9 genes (1 unknown) | 1,2,3,6,8, 10,11,19 | | | |
| #8 8 genes | 3,9,10,15, 17,19 | 1 | | |
| #9 7 genes (2 unknown) | 1,3,5,16,19 | | | |
| #10 6 genes (1 unknown) | 15 | 6,12 | | |
| #11 5 genes (1 unknown) | 1,2,9,12 | | | |
| #12 5 genes | 15 | 3,12 | | |
| #13 5 genes | 1,4,7,21,22 | | | |
| #14 4 genes | 1,2,5,7 | | | |

**Table 1.** List of chromosomal assignments of genes present in relevance networks with high mutual information and with more than 3 genes. See further details in the text.

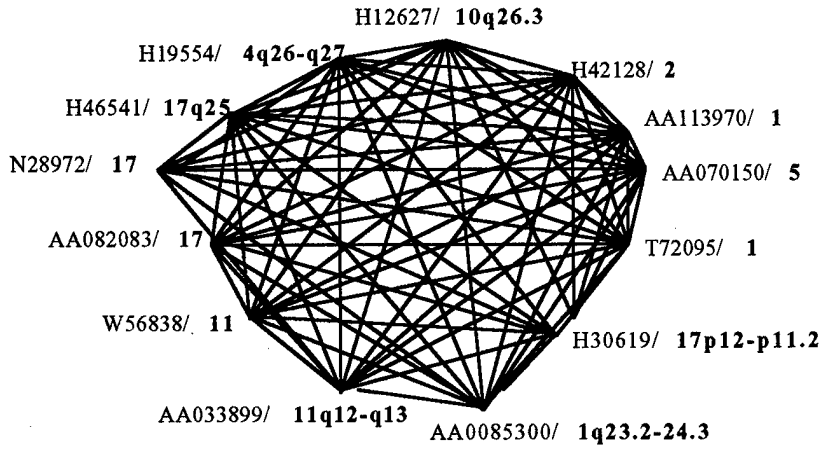


Figure 2. Relevance network #3. The gene accession number and the corresponding chromosomal localization (in bold letters) is listed for each gene.

**References**:

1. Sen, S. Aneuploidy and cancer. *Curr Opin Oncol* 12:82-88 (2000)
2. Perou, C.M., *et al.* Distinctive gene expression patterns inhuman mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* 96:9212-9217. (1999)
3. Sandoval, C., *et al.* Trisomy 5 as a sole cytogenetic abnormality in pediatric acute lymphoblastic leukemia. *Cancer Genet Cytogenet* 118:69-71 (2000)
4. Menard,S,, *et al.*Role of HER2 gene overexpression in breast carcinoma. J *Cell Physiol* 182:150-62 (2000)
5 . Galitski, T., *et al.* Ploidy regulation of gene expression. *Science* 285:251--254. (1999)
6. Coller,, H.A., *et al.* Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc Natl Acad Sci U S A* 97:3260-3265 (2000)
7. Kauffman, S. The origins of order. *Oxford: Oxford University Press.* (1993).
8. Lau, C.C., *et al.* Plasmid-induced "hit-and-run" tumorigenesis in Chinese hamster embryo fibroblast (CHEF) cells. *Proc Natl Acad Sci U S A* 82:2839-2843 (1985)
9. Plattner, R, *et al.* Loss of oncogenic ras expression does not correlate with loss of tumorigenicity in human cells. *Proc Natl Acad Sci U S A* 93:6665-6670 (1996)
10. Felsher, D.W., and Bishop, J.M. Reversible tumorigenesis by MYC in hematopoietic lineages. *Mol Cell* :199-207 (1999)

11. Baasner S., *et al.*Reversible tumorigenesis in mice by conditional expression of the HER2/c-erbB2 receptor tyrosine kinase. Oncogene 13:901-911 (1996)
12. Eisen, M.B., *et al.* Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863-14868 (1998)
13. Butte, A.J. and Kohane, I.S. Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. *Pacific Symposium on Biocomputing* 5:415-426 (2000).
14. Chen, Y., Dougherty, E.R., and Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images.*J. Biomed.Optics* 2:364--374. (1997)
15. Mertens, F., *et al.* Chromosomal imbalance maps of malignant solid tumors: a cytogenetic survey of 3185 neoplasms. *Cancer Res.*57:2765-2780. (1997)
16. Botti C, *et al.* Incidence of chromosomes 1 and 17 aneusomy in breast cancer and adjacent tissue: an interphase cytogenetic study. *J Am Coll Surg* 190:530-539 (2000)
17. Marinhom A.F., Botelho, M., and Schmitt F.C. Evaluation of numerical

# 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 ~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

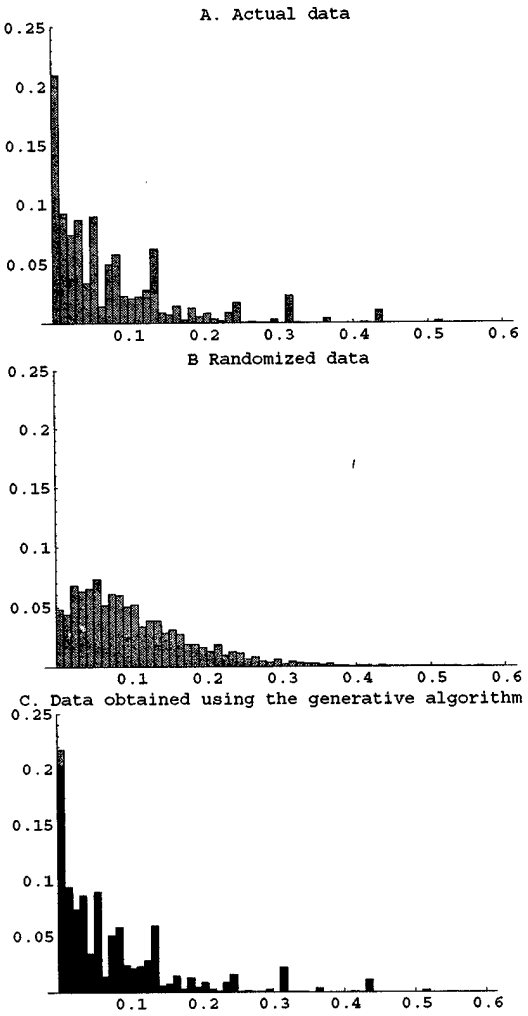


Figure 1: The distribution of pair–wise mutual information content. Panel A: Actual data set from Perou *et al.* (see text); B. Data randomized using simple randomization; C. Data obtained by running the generative algorithm.

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 $N_{\text{bins}}$ 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 $p_{\text{swap}}$, 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 $\Delta_c$. Usually, $\Delta_c$ 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, $p_{\text{swap}}$ is given a large value in the beginning of a run, when the difference between the two histograms is large. The value of $p_{\text{swap}}$ 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 $N_{\text{bins}}$ (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 $\Delta$ 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 $p_{\text{swap}}$, 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 $p_{\text{swap}}$. 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 $8.5 \pm 7.7$ 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 $0.017 \ll 1$ 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

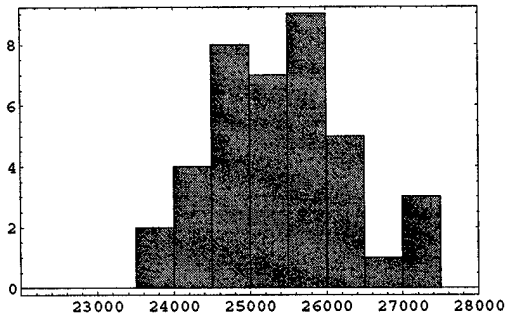


Figure 3: Histogram of the expected number of separators obtained using generative models for the reduced Perou *et al.* data set.

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 $3.7 \pm 1.4$. 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)

BIOINFORMATICS

Vol. 18 no. 3 2002
Pages 389–394

# Assessing the significance of consistently mis-regulated genes in cancer associated gene expression matrices

*Mattias Wahde[1,*], Gregory T. Klus[2], Michael L. Bittner[3], Yidong Chen[3] and Zoltan Szallasi[2,4,*]*

*[1]Division of Mechatronics, Chalmers University of Technology, Göteborg, Sweden, [2]Department of Pharmacology, Uniformed Services University of the Health Sciences, Bethesda, MD, USA, [3]Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA and [4]Children's Hospital Informatics Program, Harvard Medical School, Boston, MA, USA*

Received on April 20, 2001; revised on October 1, 2001; accepted on November 11, 2001

## ABSTRACT

**Motivation:** The simplest level of statistical analysis of cancer associated gene expression matrices is aimed at finding consistently up- or down-regulated genes within a given set of tumor samples. Considering the high level of gene expression diversity detected in cancer, one needs to assess the probability that the consistent mis-regulation of a given gene is due to chance. Furthermore, it is important to determine the required sample number that will ensure the meaningful statistical analysis of massively parallel gene expression measurements.

**Results:** The probability of consistent mis-regulation is calculated in this paper for binarized gene expression data, using combinatorial considerations. For practical purposes, we also provide a set of accurate approximate formulas for determining the same probability in a computationally less intensive way. When the pool of mis-regulatable genes is restricted, the probability of consistent mis-regulation can be overestimated. We show, however, that this effect has little practical consequences for cancer associated gene expression measurements published in the literature. Finally, in order to aid experimental design, we have provided estimates on the required sample number that will ensure that the detected consistent mis-regulation is not due to chance. Our results suggest that less than 20 sufficiently diverse tumor samples may be enough to identify consistently mis-regulated genes in a statistically significant manner.

**Availability:** An implementation using Mathematica™ of the main equation of the paper, (4), is available at www.me.chalmers.se/~mwahde/bioinfo.html.

**Contact:** mwahde@me.chalmers.se, zszallasi@chip.org

*To whom correspondence should be addressed.

## 1 INTRODUCTION

Due to recent technological developments, cancer research is delivering an increasing number of large-scale gene expression matrices associated with a wide variety of neoplastic states. In cDNA microarray measurements changes in gene expression levels are determined relative to an appropriate reference sample such as RNA derived from non-neoplastic tissue or cell lines (see e.g. Perou *et al.*, 1999) or pooled RNA from all tumor samples examined (see e.g. Bittner *et al.*, 2000). Although these measurements produce continuous data, their interpretation, due to a host of experimental and theoretical issues, is far from obvious. Therefore, at the simplest level of analysis it is practical to convert the continuous data into up-, or down-regulation or no change in the expression levels, and then search for consistently up- or down-regulated genes in an appropriately selected subset of samples, e.g. a given type of tumor. (For simplicity, from now on we will use the term 'mis-regulation' instead of up- or down-regulation, whenever the expression of a gene significantly differs from the reference expression level in a given experiment.)

This analysis has required the solution of two non-trivial problems. First, determining up- or down-regulation (or no change) with a given confidence level required the development of appropriate statistical tools that have been described and reviewed elsewhere (Manduchi *et al.*, 2000; Claverie, 1999; Chen *et al.*, 1997). This step can be viewed as a conversion of the continuous data matrix into a discrete matrix which can be either ternary, in which up-, down-regulation and no change are represented by the discrete values of 1, −1, and 0 respectively, or a binary matrix, in which only the fact of change or no change is recorded. The second step of the analysis is the

subject of this paper and determines the probability that any gene is consistently up- or down-regulated by chance in cancer associated gene expression matrices that are usually characterized by a high level of gene expression diversity.

From a biological standpoint the analysis presented in this paper is based on the assumption that there are groups of highly related tumor samples that share the same genetic background in terms of gene expression patterns. Therefore, gene expression changes that are causative or the result of a given type of cancer, are supposed to show a pattern of consistent mis-regulation over a sufficient number of tumor samples and therefore be identifiable by 'guilt by association' analysis. Finding these genes, however, is complicated by the fact that cancer is associated with a large number of changes in gene expression levels. The exact fraction of mis-regulated genes depends on the gene-set contained on the microarray chip. However, measurements performed with a large set (more than 5000) of relatively randomly selected probes, such as the one used by Perou *et al.* suggest that the number of mis-regulated genes may amount to about 10–15% of all genes present. Many of these changes are probably not intimately involved with the development or maintenance of cancer but rather due to the major rearrangement of the genetic network in neoplastic cells which is often associated with aneuploidy (Klus *et al.*, 2001). The considerable level of gene expression diversity, however, raises the question whether the detected consistent mis-regulation is simply due to chance.

In this paper we will provide theoretical and computational guidelines in order to calculate the probability of this event given a certain type of gene expression data set. We will use binarized data in order to introduce some of the combinatorics problems at hand and at the same time we will provide a theoretical estimate about the number of different cancer samples required to perform meaningful statistical analysis. We will briefly point out that the statistical analysis of ternary gene expression data can be derived from the binary analysis, and for all practical purposes it is covered by the equations provided by binary considerations. A more comprehensive analysis of gene expression matrices will search for a group of $K$ mis-regulated genes, the status of which, when coupled by an appropriate rule, will allow the separation of neoplastic and normal samples. Such a group of genes and the appropriate function form a separator (Wahde and Szallasi, 2001). The present paper covers the special case of $K = 1$ separators, whereas higher order ($K = 2$) separators were recently treated in Wahde and Szallasi (2001).

## 2 SYSTEMS AND METHODS

### 2.1 Binary analysis of consistently mis-regulated genes by combinatorics

Binary analysis does not distinguish between the states of up- and down-regulation for a given gene, it only registers the state of mis-regulation. A typical measurement contains $E$ tumor samples, where the number of mis-regulated genes is $M_i$ in the $i$th sample, and the total number of genes expressed across all samples examined is $N$. If the mis-regulated genes are randomly and independently selected then we can assess the significance of finding $K$ consistently mis-regulated genes by solving the following combinatorics problem: let us pick $M_i$ elements randomly and independently out of $N$ elements in $E$ consecutive experiments. How likely is it that at least $K$ elements will be picked in all $E$ experiments? This probability is determined by the following equation (for a brief derivation see Appendix):

$$P(E, k \geqslant K) = 1 - \sum_{i=0}^{K-1} P(E, k), \tag{1}$$

where $P(E, k)$ is the probability that exactly $k$ genes are consistently mis-regulated in $E$ experiments. This probability is determined by the following recursive formula:

$$P(E, k) = \sum_{j=k}^{\min(M_1, M_2, \ldots, M_{E-1})} \frac{\binom{j}{k}\binom{N-j}{M_E-k} P(E-1, j)}{\binom{N}{M_E}}, \tag{2}$$

where $M_E$ is the number of mis-regulated genes in the last ($E$th) experiment, and also

$$P(2, k) = \frac{\binom{M_1}{k}\binom{N-M_1}{M_2-k}}{\binom{N}{M_2}}. \tag{3}$$

In cases where the $M_i$ values are almost equal for $i = 1, \ldots, E$, (2) can be simplified into

$$P(E, k) = \sum_{j=k}^{M_{\text{av}}} \frac{\binom{j}{k}\binom{N-j}{M_{\text{av}}-k} P(E-1, j)}{\binom{N}{M_{\text{av}}}}, \tag{4}$$

where $M_{\text{av}}$ is the average of the $M_i$. Note, however, that if the $M_i$ values vary significantly from sample to sample, (2) should be used.

An implementation of (4) in Mathematica™ is available from the authors at the following web site: www.me.chalmers.se/~mwahde/bioinfo.html.

The computational cost of the formulae above increases rapidly with $E$. In fact, exact calculations for $E > 6$ are impractical because of the long CPU-time required (e.g. already for $E = 6$, $K = 1$, $M = 500$, and $N = 5000$

the computation time is already about 1 h and 35 min on a computer equipped with a 500 MHz PIII processor). Therefore, for practical purposes, we are introducing here an approximative formula that provides results similar to those given by (1), in the case of $N \gg M_{av}$.

$$P(E, k \geqslant K) \approx \frac{\binom{N}{K}\binom{N-K}{M_{av}-K}^{E}}{\binom{N}{M_{av}}^{E}}. \quad (5)$$

The details of the derivation of this equation will not be given here.

Probabilistic approaches provide another useful, somewhat less accurate formula. The probability that a given gene is mis-regulated in a given tumor sample is approximately given by $q = M/N$. The probability that exactly $k$ genes will be mis-regulated in all $E$ experiments can be estimated by the following formula, which is essentially a specialized application of the binomial distribution

$$P(E, k) \approx \binom{N}{k}(q^E)^k(1-q^E)^{N-k}. \quad (6)$$

This equation requires that $K \ll M \ll N$.

Table 1 compares the $P(E, k \geqslant K)$ values as functions of the number of the samples ($E$) derived by the recursive formula (1), using (4), and the approximative equations (5) and (6) for several $k$ values. We have used a ratio of $M/N = 0.1$, which is often observed in cancer associated gene expression matrices.

### 2.2 Ternary analysis of consistently mis-regulated genes by combinatorics

A preliminary analysis (data not shown) indicated that about 50% of all mis-regulated genes show inconsistency in their direction of mis-regulation. These genes show up-regulation in some samples and down-regulation in others within the same tumor type. Therefore, we considered handling up- and down-regulation separately, in order to calculate the probability $P(E, k, i)$ that $k$ genes are consistently mis-regulated by chance with $i$ ($i \leqslant k$) genes being consistently mis-regulated in the same direction (i.e. either up (1) or down (−1)). It is self-evident that there will be fewer cases here than when asking the question how many times exclusively non-0's (without examining the direction of mis-regulation) will be found for exactly $k$ genes. Thus $P(E, k, i) < P(E, k)$.

In most cases, all we want to know is whether our finding is unlikely to be a chance event. If the calculations suggest that it is unlikely to have $k$ mis-regulated genes by chance, then it is even more unlikely that a certain number, $i$, of those genes will be mis-regulated the same direction. Therefore we are justified to avoid the arduous combinatorics calculations on ternary data.

## 3 RESULTS

### 3.1 Estimating the required sample number in order to validate statistically significant consistent mis-regulation

One of the key issues in the experimental design of massively parallel gene expression measurements is determining the required sample number that will ensure the appropriate power of statistical analysis: given a certain sample quality, which includes the number of measurable genes and average gene expression diversity, how many samples do we need to be sure that the consistent mis-regulation of $k$ genes is not due to chance at a given confidence level? Equations (1)–(6) can be exploited in order to answer this question. Currently, a typical cancer associated gene expression measurement contains about 5000 genes, of which 10–15% are mis-regulated in every sample. With these numbers about $E = 8$ samples are sufficient to establish that any (i.e. $K \geqslant 1$) consistent mis-regulation observed is not due to accident. These calculations can be easily updated as experimental data change. However, we would like to point out that for the whole human transcriptome, with about $N = 50\,000$–$100\,000$ different splice variants of genes, with the average 10–15% cancer associated gene expression diversity, about $E = 10$ samples will be sufficient to establish that consistent mis-regulation of a gene is not due to chance with a confidence level of 99.9%.

The breast cancer associated data set published by Perou *et al.* (1999) contains cDNA microarray based relative expression measurements of about 5584 genes for a number of both normal and neoplastic breast epithelial samples. A total of 43 genes are consistently mis-regulated in this data set. Applying (1) we found that the probability that at least 43 genes will be consistently mis-regulated by chance is on the order of $10^{-216}$ and the probability that at least one gene will be mis-regulated is on the order of $10^{-4}$. Thus, it is very unlikely that the consistent mis-regulation of genes observed in these tumor samples is due to chance.

### 3.2 Range of validity for the equations

Equations (1)–(6) were derived assuming that mis-regulated genes are randomly and independently selected, and they therefore loose their validity if this assumption is incorrect. In fact, biological systems display at least two major restrictions on the selection of mis-regulated genes. First, not every gene can be mis-regulated. Gene expression matrices typically contain at least 5–20% genes that are unchanged in any of the neoplastic samples, even if those matrices were derived from a large number of cancer samples, such as the more than 70 lymphoma cases published by Alizadeh *et al.* (2000). (It is obviously

**Table 1.** Values of $\log_{10}(P(E, k \geqslant K))$, i.e. the logarithm of the probability of having at least $k$ consistently mis-regulated genes in $E$ samples, computed using (4)–(6). The number of genes ($N$) was equal to 1000 and the number of mis-regulated genes ($M$) in each sample, was equal to 100. Under these conditions, when $E > 4$, all three equations give the same results to within an error of less than 1%

| | $K = 1$ | | | | $K = 3$ | | |
|---|---|---|---|---|---|---|---|
| $E$ | Equation (4) | Equation (5) | Equation (6) | $E$ | Equation (4) | Equation (5) | Equation (6) |
| 3 | −0.434 08 | −0.195 64 | 0.000 00 | 3 | – | −2.109 75 | −0.814 97 |
| 4 | −1.043 39 | −1.020 77 | −1.000 00 | 4 | −3.822 76 | −3.855 74 | −3.826 81 |
| 5 | −2.004 34 | −2.002 18 | −2.000 00 | 5 | −6.783 79 | −6.841 34 | −6.838 63 |
| 6 | −3.000 43 | −3.002 18 | −3.000 00 | 6 | −9.779 89 | −9.850 78 | −9.850 47 |
| 7 | −4.000 04 | −4.004 36 | −4.000 00 | 7 | −12.779 5 | −12.862 3 | −12.862 3 |
| 8 | −5.000 00 | −5.000 00 | −5.000 00 | 8 | −15.779 5 | −15.874 2 | −15.874 2 |
| 9 | −6.000 00 | −6.000 00 | −6.000 00 | 9 | −18.779 5 | −18.886 1 | −18.886 1 |
| 10 | −7.000 00 | −7.000 00 | −7.000 00 | 10 | −21.779 5 | −21.897 9 | −21.897 9 |

more likely that a higher percentage of changeable genes will display mis-regulation when a large number of samples is examined.) Second, mis-regulated genes are not independently selected as reflected in the high level of pair-wise mutual information content displayed in cancer associated gene expression matrices (Klus *et al.*, 2001; Butte and Kohane, 2000). Ignoring these restrictions can lead to an underestimation of the chance appearance of consistently mis-regulated genes, therefore attaching an erroneously high significance to these observations.

In the following section we will examine the effect of the first of the restrictions listed above on calculating the statistical significance of consistently mis-regulated genes.

### 3.3 Determining the pool of 'mis-regulatable' genes

The fact that some genes remain unchanged in all of the tumor samples will obviously lead to a smaller $N$ in the equations above. Therefore, for more accurate calculations it should be established whether the unchanged genes are never mis-regulated in cancer or whether they can be mis-regulated but the sample number of the gene expression matrix was too small to provide a chance for all possible changes to be displayed. The number of mis-regulatable (or changeable) genes can be estimated using conditional probabilities as follows: let us designate the number of changeable genes as $N_{\text{eff}}$, (the total number of measured genes is $N$). Assuming random and independent selection of the mis-regulated genes, the probability that a gene will remain unchanged across all $E$ cell lines can be written

$$P(\text{U}) = 1 \times P(\text{UC}) + P(\text{UC}|\text{CH}) \times P(\text{CH}), \qquad (7)$$

where $P(\text{UC})$ is the probability that the gene is unchangeable, $P(\text{UC}|\text{CH})$ the probability that the gene does not change in any cell line given that it is changeable, and $P(\text{CH})$ the probability that the gene is changeable. Based on frequencies, these probabilities can be estimated, in the same order, as $(N - N_{\text{eff}})/N$, $\prod_{i=1}^{E}(1 - M_i/N_{\text{eff}})$, and $N_{\text{eff}}/N$. Inserting these values into (7), the expected number of unchanged genes is obtained as

$$N_{\text{U}} = N - N_{\text{eff}}\left[1 - \prod_{i=1}^{E}\left(1 - \frac{M_i}{N_{\text{eff}}}\right)\right]. \qquad (8)$$

We have applied (8) to several published large-scale gene expression matrices. Figure 1 is a representative sample of our results based on breast cancer associated gene expression matrices published by Perou *et al.* (1999). For our analysis we have used only gene expression measurements derived from either breast cancer cell lines or primary breast tumors, 16 samples altogether. In this case, the best fit of (8) to the experimental data is obtained if the number of mis-regulatable genes is set to around 5100.

Equations (1)–(4) suggest that the probability of having at least $k$ consistently mis-regulated genes in a given data set will depend on the effective number of changeable genes. In order to estimate this effect we have calculated the probability for several $k$ values as a function of the ratio of $N_{\text{eff}}$ relative to $N$. The results, shown in Figure 2, indicate that, if $N$ is used instead of the actual $N_{\text{eff}}$, $P(E, k \geqslant K)$ can be underestimated by up to several orders of magnitude in the case of larger $K$ values. The typical range of $N_{\text{eff}}$ is between $0.8N$ and $0.95N$, with for example $0.94N$ for the data set derived from Perou *et al.* (1999). With these values of $N_{\text{eff}}$, underestimation occurs at very low $P(E, k \geqslant K)$ values, creating little practical consequences. Nevertheless, the correct $N_{\text{eff}}$ can be easily estimated by the approach demonstrated in Figure 1. It is evident from (8) that the number of unchanged genes is asymptotically approaching the value of $(N - N_{\text{eff}})$ as $E$ increases. We can readily determine the value of $E$ at which the difference between the expected number of unchanged genes and $(N - N_{\text{eff}})$ drops below a certain threshold value. Using (8), and replacing the individual $M_i$ values with an average $M$ value a simple formula for this

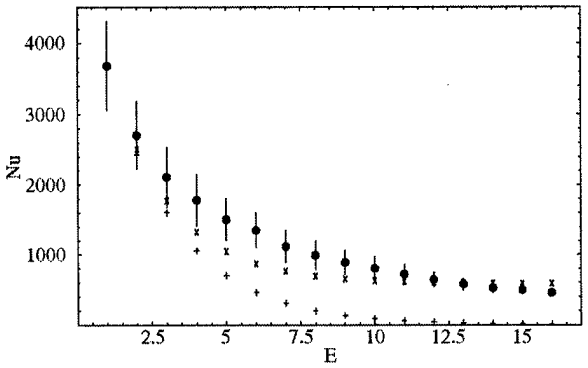


**Fig. 1.** The number of unchanged genes as a function of the number of samples for the Perou *et al.* (1999) data set. The curves show the the expected number of unchanged genes based on (8), assuming that all genes can be changed (lower curve, + symbols) or that only 5100 genes are changeable (X symbols). The dots with error bars show the results from the experimental data. The error bars stem from the fact that, for the data point corresponding to $e$ samples, there are $\binom{E}{e}$ ways of selecting the samples.

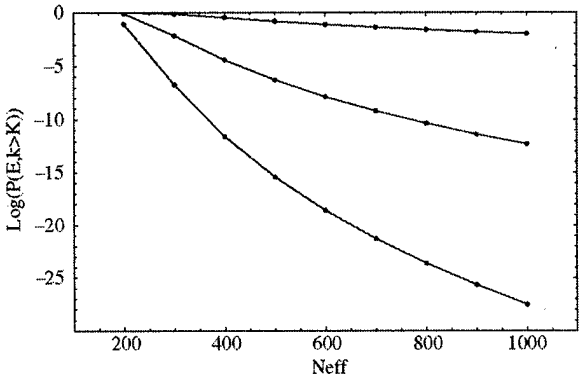


**Fig. 2.** $\log_{10}(P(E, k \geqslant K))$ as a function of $N_{\text{eff}}$ for $M = 100$, $E = 5$, and $K = 1$ (top curve), 5, and 10 (bottom curve).

calculation is obtained

$$\frac{N_{\text{eff}}\left(1 - \frac{M}{N_{\text{eff}}}\right)^E}{N - N_{\text{eff}}} \leqslant \epsilon. \tag{9}$$

Solving for $E$, one obtains

$$E = \frac{\log\left(\epsilon\left(\frac{N}{N_{\text{eff}}} - 1\right)\right)}{\log\left(1 - \frac{M}{N_{\text{eff}}}\right)}. \tag{10}$$

Knowing that $N_{\text{eff}}$ is of order $0.8N$ to $0.95N$, the required number of samples can be estimated. For example, for the Perou *et al.* data set discussed above, with $N = 5584$ and $M_{\text{av}} = 1902$ the required number of samples is between $E = 11$ and $E = 17$ for $\epsilon = 0.01$. However, with a lower ratio of $M/N$, the required number of experiments can be significantly higher.

## 4 DISCUSSION

The analysis of massively parallel gene expression measurements in cancer will be performed at different levels of complexity. In order 'to pick the low hanging fruit' first, it seems feasible to perform a simple form of 'guilt by association analysis' and identify consistently mis-regulated genes in neoplastic samples. The significant diversity of cancer associated gene expression patterns, however, necessitates the use of appropriate statistical analysis. Successful statistical analysis will require understanding the structure of the data, creating the corresponding null hypothesis and performing the appropriate calculations. It is often true, as in the case of this paper, that accepting a simpler data structure (e.g. random and independent selection of mis-regulated genes) yields significantly easier calculations. It is one of the central issues of bioinformatics to find the correct balance between the complexity of data structure and the corresponding difficulties of calculations. Finding this balance will provide biologists with the simplest statistical calculations that provide satisfactory results. We have followed these guidelines in this paper while addressing the issue of consistently mis-regulated genes. Assuming that mis-regulated genes in cancer are randomly and independently selected leads to straightforward combinatorial calculations and easy to use approximative formulae for the case of $K \ll M \ll N$, which holds for all cancer associated gene expression matrices published so far.

These calculations yielded the interesting and practical result, that about 10 sufficiently diverse tumor samples are enough to identify consistently mis-regulated genes in a statistically significant manner, even if the complete human transcriptome is probed. We are well aware of the fact that cancer associated gene expression patterns are produced by the rearrangement of complex genetic networks. Therefore, the assumption of random and independent selection of mis-regulated genes is oversimplified. There are two obvious restrictions on the data structure of these matrices. First, not every gene can be mis-regulated. Second, genes are mis-regulated in a coordinated fashion (see e.g. Wahde and Szallasi, 2001). Here we have examined the effect of the first restriction on statistical analysis. Since the pool of mis-regulatable genes can be well estimated with a relatively limited number of samples (less than 20), statistical calculations can be readily adjusted accordingly. This is probably worth doing even if we found a relatively limited effect of mis-estimating the number of mis-regulatable genes.

We have recently addressed the effect of coordinated mis-regulation for the statistical analysis of $K = 2$ separators. In order to overcome complicated calculations we have introduced a simulative process, called generative models, to estimate the chance appearance of these higher order separators (Wahde and Szallasi, 2001). Strikingly, we found that the results of statistical analysis can be off by many orders of magnitude when the coordinated mis-regulation of genes was ignored. We are currently modifying the generative model in order to accommodate the analysis of $K = 1$ separators i.e. consistently mis-regulated genes as well.

## ACKNOWLEDGEMENT

The authors would like to thank Jake P. Solomon for stimulating e-mail exchanges.

## APPENDIX

A brief derivation of (2) and (3): consider first the case of $E = 2$ samples with $N$ genes each. In the first sample, $M_1$ genes are mis-regulated. Assuming random and independent selection, the probability of having exactly $k$ genes consistently mis-regulated (hereafter denoted CM), i.e. mis-regulated in both samples, can easily be computed by noting that, for the second sample, there are $\binom{M_1}{k}$ ways of selecting the $k$ genes that were mis-regulated in the first sample (to obtain $k$ CM genes), and the remaining $M_2 - k$ mis-regulated genes can then be selected in $\binom{N-M_1}{M_2-k}$ ways.

The total number of ways of selecting $M_2$ genes out of $N$ is $\binom{N}{M_2}$, and thus (3) is derived. Consider now the case of 3 samples and assume, to begin with, that $j$ genes were CM in the first two samples. In order to obtain exactly $k$ CM genes in the three samples, $k$ of the mis-regulated genes in sample three must be selected from the $j$ genes that were CM in the first two samples. This can be done in $\binom{j}{k}$ ways. The remaining $M_3 - k$ mis-regulated genes in the third sample must be selected from the other $N - j$ genes, which can be done in $\binom{N-j}{M_3-k}$ ways.

The selection of $M_3$ genes among $N$ can be done in $\binom{N}{M_3}$ ways, and so, the probability of having $k$ CM genes, given $j$ CM genes after two samples would be

$$p(k|j) = \frac{\binom{j}{k}\binom{N-j}{M_3-k}}{\binom{N}{M_3}}. \qquad \text{(A.1)}$$

Now, the number of CM genes in the two first samples can range from 0 to $\min(M_1, M_2)$. If it is smaller than $k$ then, clearly, the probability of obtaining $k$ CM genes after three samples is zero. Thus, the probability of having $k$ CM genes after three samples will consist of a sum ranging from $j = k$ to $j = \min(M_1, M_2)$, in which the individual terms will consist of the product of $p(k|j)$ (A.1) and $p(2, j)$ (3):

$$p(3, k) = \sum_{j=k}^{\min(M_1,M_2)} p(k|j)p(2, j). \qquad \text{(A.2)}$$

Note, that this is identical to (2) for $E = 3$. It is easy to generalize this equation to any number E of samples, and so (2) follows.

## REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–11.

Bittner,M. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, **5**, 415–426.

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.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.*, **8**, 1821–1832.

Klus,G.T., Song,A., Schick,A., Wahde,M. and Szallasi,Z. (2001) Mutual information analysis as a tool to assess the role of aneuploidy in the generation of cancer associated differential gene expression patterns. *Pac. Symp. Biocomput.*, **6**, 42–51.

Manduchi,E., Grant,G.R., McKenzie,S.E., Overton,G.C., Surrey,S. and Stoeckert,C.J. (2000) Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, **16**, 685–698.

Perou,C.M. *et al.* (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA*, **96**, 9212–9217.

Wahde,M. and Szallasi,Z. (2001) Generative model based analysis of cancer associated gene expression matrices. In Kitano,H. (ed.), *Proceedings of the 1st International Conference on Systems Biology*, pp. 39–45.