



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

Mattias Wahde<sup>1,\*</sup>, Gregory T. Klus<sup>2</sup>, Michael L. Bittner<sup>3</sup>,  
Yidong Chen<sup>3</sup> and Zoltan Szallasi<sup>2, 4,\*</sup>

<sup>1</sup>Division of Mechatronics, Chalmers University of Technology, Göteborg, Sweden, <sup>2</sup>Department of Pharmacology, Uniformed Services University of the Health Sciences, Bethesda, MD, USA, <sup>3</sup>Cancer Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA and <sup>4</sup>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](http://www.me.chalmers.se/~mwahde/bioinfo.html).

**Contact:** [mwahde@me.chalmers.se](mailto:mwahde@me.chalmers.se), [zszallasi@chip.org](mailto:zszallasi@chip.org)

### 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

\*To whom correspondence should be addressed.

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):

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

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:

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

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

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

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

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

where  $M_{av}$  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](http://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  $N \gg M_{av}$ .

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

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  $P(E, k \geq K)$  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  $i$  ( $i \leq k$ ) 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.  $K \geq 1$ ) 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 lose 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  $\log_{10}(P(E, k \geq K))$ , 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	-0.43408	-0.19564	0.00000	3	-	-2.10975	-0.81497
4	-1.04339	-1.02077	-1.00000	4	-3.82276	-3.85574	-3.82681
5	-2.00434	-2.00218	-2.00000	5	-6.78379	-6.84134	-6.83863
6	-3.00043	-3.00218	-3.00000	6	-9.77989	-9.85078	-9.85047
7	-4.00004	-4.00436	-4.00000	7	-12.7795	-12.8623	-12.8623
8	-5.00000	-5.00000	-5.00000	8	-15.7795	-15.8742	-15.8742
9	-6.00000	-6.00000	-6.00000	9	-18.7795	-18.8861	-18.8861
10	-7.00000	-7.00000	-7.00000	10	-21.7795	-21.8979	-21.8979

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

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

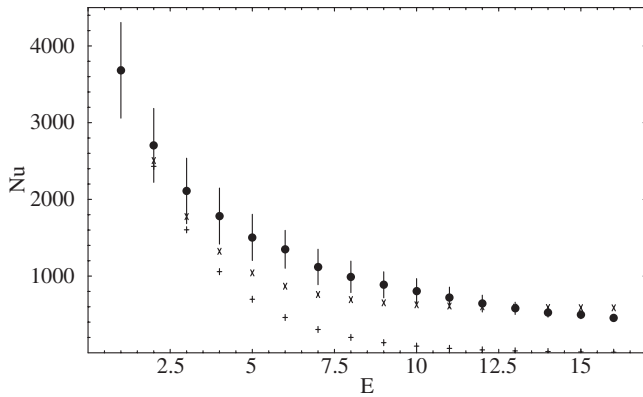
where  $P(\text{UC})$  is the probability that the gene is unchangeable,  $P(\text{UC}|\text{CH})$  the probability that the gene does not change in any cell line given that it is changeable, and  $P(\text{CH})$  the probability that the gene is changeable. Based on frequencies, these probabilities can 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

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

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  $N_{\text{eff}}$ ,  $P(E, k \geq K)$  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  $P(E, k \geq K)$  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 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.

calculation is obtained

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

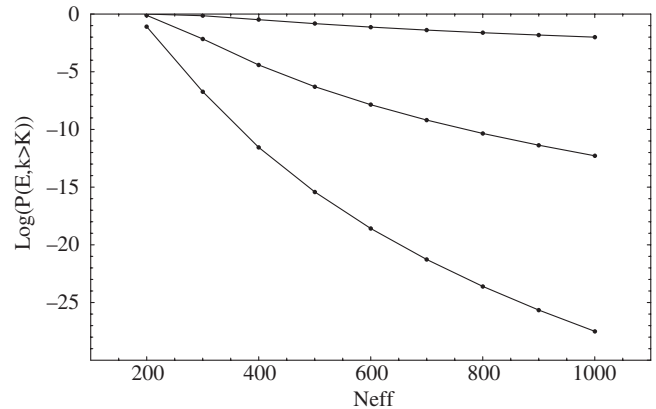
Solving for  $E$ , one obtains

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

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



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

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

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

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):

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

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**, 503–511.
- 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 Stoekert, 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.