Probabilistic estimation of microarray data reliability and underlying gene expression

S Bilke, T Breslin and Mikael Sigvardsson

Post Print

N.B.: When citing this work, cite the original article.

Original Publication:

S Bilke, T Breslin and Mikael Sigvardsson, Probabilistic estimation of microarray data reliability and underlying gene expression, 2003, BMC Bioinformatics, (4), 40.
http://dx.doi.org/10.1186/1471-2105-4-40
Copyright: BioMed Central
http://www.biomedcentral.com/

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-74462

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Probabilistic estimation of microarray data reliability and underlying gene expression

Sven Bilke*[1], Thomas Breslin[1] and Mikael Sigvardsson[2]

Address: [1]Complex Systems Division, Department of Theoretical Physics, University of Lund, Sölvegatan 14A, SE-223 62 Lund, Sweden and [2]The Laboratory for Cell Differentiation Studies, Department for Stem Cell Biology, BMC B12, SE-22185 Lund, Sweden

Email: Sven Bilke* - sven@thep.lu.se; Thomas Breslin - thomas@thep.lu.se; Mikael Sigvardsson - Mikael.Sigvardsson@stemcell.lu.se

* Corresponding author

Published: 10 September 2003

BMC Bioinformatics 2003, **4**:40

Received: 16 January 2003

Accepted: 10 September 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/40

© 2003 Bilke et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: The availability of high throughput methods for measurement of mRNA concentrations makes the reliability of conclusions drawn from the data and global quality control of samples and hybridization important issues. We address these issues by an information theoretic approach, applied to discretized expression values in replicated gene expression data.

Results: Our approach yields a quantitative measure of two important parameter classes: First, the probability $P(\sigma|S)$ that a gene is in the biological state σ in a certain variety, given its observed expression S in the samples of that variety. Second, sample specific error probabilities which serve as consistency indicators of the measured samples of each variety. The method and its limitations are tested on gene expression data for developing murine B-cells and a *t*-test is used as reference. On a set of known genes it performs better than the *t*-test despite the crude discretization into only two expression levels. The consistency indicators, i.e. the error probabilities, correlate well with variations in the biological material and thus prove efficient.

Conclusions: The proposed method is effective in determining differential gene expression and sample reliability in replicated microarray data. Already at two discrete expression levels in each sample, it gives a good explanation of the data and is comparable to standard techniques.

Background

A broad variety of algorithms has been developed and used to extract biologically relevant information from gene expression data. Among others commonly used are visual inspection [1], hierarchical and k-means clustering [2], self organizing maps [3,4] and singular value decomposition [5,6]. These methods aim mainly at identifying predominant patterns and thus groups of "cooperating" genes based on the assumption that related genes have similar expression patterns.

Compared to the amount of work devoted to efficient methods to extract information from the data, somewhat less attention has been paid to the question of the reliability of the generated results. The ANOVA analysis [7] allows estimation, and thus elimination, of some systematic error sources. Bootstrapping cluster analysis estimates the stability of cluster assignments [8] based on artificial data-sets generated with ANOVA coefficients. Some authors also considered the question of how well a certain oligo [10] is suited to measure the mRNA expression level of the related gene.

Some work has gone towards the ambitious task of learning topological properties or qualitative features of the genetic regulatory network from expression profiles, see

e.g. [11]. A major limiting factor in these attempts is the comparative sparseness of available data. It is therefore reasonable to consider reduced models, for example a Boolean representation of the gene activity. It is known that many biological properties, for instance stability and hysteresis, can be modeled by the dynamics of such reduced models [12–14].

In this work we investigate the possibility of reducing complexity of gene expression data by discretizing the expression levels. The approach we present enables a new way of extracting biologically relevant information from the data in the following way: A biological variety, i.e. a biological system defined by the investigator, is represented by several samples which are subjected to gene expression analysis. If gene expression levels are discretized into n values, and the variety is represented by m samples, the number of observable expression states for a gene are limited to n^m. These observed states S are modeled as being derived from a smaller number of underlying, biological states σ, through a measurement process. Rather than making static assignments $S \rightarrow \sigma$ we calculate conditional probabilities $P(\sigma|S)$. The number of possible expression profiles for a gene over a set of varieties is limited and the probability of each expression profile is easily calculated. Since the model we use considers both the underlying biology and the measurement process it also generates a measure of sample coherence in each biological variety.

We demonstrate the feasibility of this approach for a *binary* discretization of gene expression. For the discretization step we use the absent/present classification provided by the Affymetrix software [9]. The outcome of our method on a data set covering gene expression in developing murine B-cells is compared to the results of a standard analysis. We show that even with the crude discretization into only two expression levels the method is competitive to statistical methods based on continuous expression levels.

Methods

The Model

A major step in the analysis of gene expression data is to separate the biological content of the data from measurement and sample specific errors. In other words given an observation, i.e. the expression values of a gene in several samples representing the same biological variety, one wants to conclude on the biological state σ, which generated the observation. This can be expressed as a conditional probability,

$$P(\sigma|S) \quad (1)$$

, that a gene is in a certain biological state σ given the corresponding observed state S. In the application on which we demonstrate the method we consider three different biological varieties: pro, pre, and mature B-cells. The samples in each variety are different cell lines arrested at the corresponding stage of development.

In this work we take an information theoretic point of view to estimate this probability: The information of interest, the state σ, is "transmitted" in a noisy measurement process and potentially distorted (Figure 1). Using Bayes' theorem, the desired conditional probability Eq. (1) can be expressed as:

$$P(\sigma|S) = \frac{P(S|\sigma)P_\sigma}{P_S}. \quad (2)$$

On the right hand side of this equation, $P(S|\sigma)$ is the probability to observe state S if the underlying biological state is σ. In a sense, $P(S|\sigma)$ describes the noise characteristic of the measurement process. In the following we will show how this conditional probability, and the other probabilities on the r.h.s. of Eq. (2) can be estimated.

Given a set of m samples representing the same biological variety, differences in the expression level of a gene between the samples can arise from two independent sources:

1. *Random variation* within the variety. This may be caused by temporal differences in response to the stimuli, slightly different environmental conditions, genotypic differences between samples, etc.

2. *Sample specific errors*. These are mainly caused by the measurement process, e.g. differences in the treatment of the mRNA, scratched arrays, and so on. However, outlier samples, cultured under considerably different conditions, also contribute to sample specific errors.

A separation of these two contributions is possible only with an appropriate model for the variation of gene expression between the samples. In the choice of model, one has considerable freedom within the bounds set by biological plausibility. A limiting factor on the biological model comes from the type and amount of available data. The data used in this work contains only four samples for each variety. For the model we propose this is the *minimum* number of samples required to estimate the model parameters.

In the discretization of gene expression levels, we use only two discrete values, 0 and 1, for the expression of a gene in a sample. This means that the number of observable states, S, in a variety consisting of m samples is 2^m. With

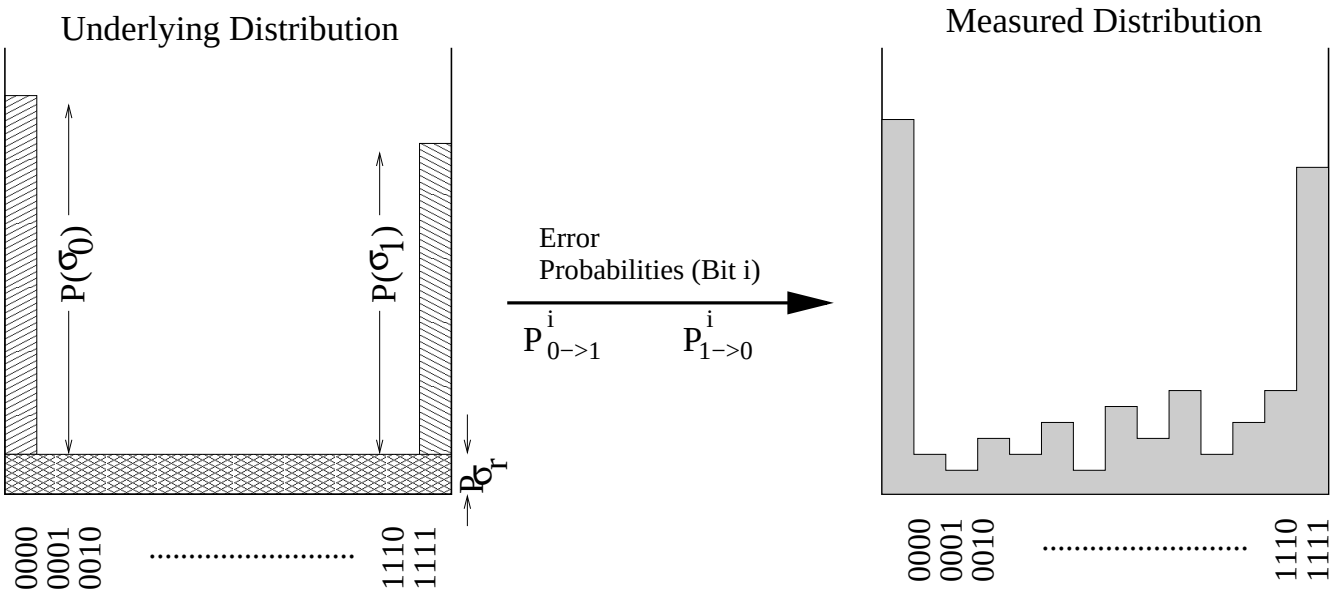

Figure 1

Schematic diagram illustrating the transition from underlying to observed distributions of states, in the case of m = 4 samples. The underlying distribution on the left hand side can be described by the probabilities for each underlying state, $P(\sigma_1)$, $P(\sigma_0)$, and $P(\sigma_r)$ (see text). This distribution is then distorted by sample specific errors, $P^i_{0\to1}$ and $P^i_{1\to0}$, resulting in an experimentally observed distribution, depicted on the right hand side.

no measurement errors we could immediately conclude on the underlying biological state σ: the two cases, where all observations agree $S = (1,...,1)$ and $S = (0,...,0)$ can be mapped to the biological states σ_1 and σ_0 respectively, which describe "pure" states without variation. The remaining N - 2 observable states S, where the individual measurements disagree, correspond to biological states σ with random variation. For the application in our biological study with supposedly *identical* biological systems contributing to the observable states S, the exact pattern leading to contradicting observations does not carry any information, as long as we assume that there are no sample specific errors. Therefore, we subsummize all N - 2 possible observations as one biological state σ_r with a random variation.

The biological rationale for this model is given by the following example: If one considers a biological variety such as cells in the retina of the eye, then a certain number of crucial genes ought to be expressed in all samples. Such genes might include rhodopsin, a molecule that responds to light. In contrast, genes such as the hemoglobin family, which are typical of erythrocytes, ought not to be expressed in the retina. A third class of genes could be considered as independent of the system in the sense that their expression is not directly related to the biological system. Such genes may vary in expression both due to environmental and genetic differences between the samples.

The model discussed so far is depicted graphically in the left part of Figure 1, where a possible distribution of the relative frequencies of the three biological states is depicted, for the case of m - 4 samples. The distribution can be described by three numbers: the probabilities $P(\sigma_1)$ and $P(\sigma_0)$, which contribute to the frequencies of the states $S = (1,1,1,1)$ and $S = (0,0,0,0)$, and $P(\sigma_r)$ which contributes to both the frequency of mixed states and the two states above. Describing the mixed states with only one parameter $P(\sigma_r)$ implies that the biological variation is modeled evenly and identically distributed independently for each sample. In a second step, the measurement process with possible sample specific errors is modeled as statistically independent between samples. For each sample i, we define two parameters, $P^i_{0\to1}$ and $P^i_{1\to0}$, denoted sample specific *error probabilities*.

To introduce the full formalism of our current model we start by considering a simple example, again for m - 4 samples. An observed state S, $S \equiv (S_1, S_2, S_3, S_4) = (1,0,1,0)$,

may be generated by the gene being in state σ_1 with the probability:

$$P(\sigma_1)\left(1-P^1_{1\to 0}\right)\left(P^2_{1\to 0}\right)\left(1-P^3_{1\to 0}\right)\left(1-P^4_{1\to 0}\right),$$

or it may be generated by the gene being in state σ_0 with the probability:

$$P(\sigma_0)\left(P^1_{0\to 1}\right)\left(1-P^2_{0\to 1}\right)\left(P^3_{0\to 1}\right)\left(1-P^4_{0\to 1}\right),$$

or it may be generated by the gene being in state σ_r with the probability:

$$P(\sigma_r)\times\frac{1}{2}\left[\left(P^1_{0\to 1}\right)+\left(1-P^1_{1\to 0}\right)\right]\frac{1}{2}\left[\left(1-P^2_{0\to 1}\right)+\left(P^2_{1\to 0}\right)\right]\times\frac{1}{2}\left[\left(P^3_{0\to 1}\right)+\left(1-P^3_{1\to 0}\right)\right]\frac{1}{2}\left[\left(1-P^4_{0\to 1}\right)+\left(P^4_{1\to 0}\right)\right].$$

With the briefer notation,

$$\rho^i_{1\to 0} \equiv P^i_{1\to 0}\delta_{S_i,0}+\left(1-P^i_{1\to 0}\right)\delta_{Si,1}$$

$$\rho^i_{0\to 1} \equiv P^i_{0\to 1}\delta_{S_i,1}+\left(1-P^i_{0\to 1}\right)\delta_{Si,0}$$

, where δ refers to the Kronecker delta (i.e. $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization with m samples, as:

$$P(S)=P(\sigma_1)\prod_{i=1}^m \rho^i_{1\to 0}+P(\sigma_0)\prod_{i=1}^m \rho^i_{0\to 1}+P(\sigma_r)\prod_{i=1}^m \frac{1}{2}\left(\rho^i_{1\to 0}+\rho^i_{0\to 1}\right). \quad (3)$$

Altogether the model uses 3 + 2 * m variables. These parameters $P(\sigma_1)$, $P(\sigma_0)$, $P(\sigma_r)$ and $\left\{P^i_{1\to 0}, P^i_{0\to 1}\right\}_{i=1}^m$ are estimated from the observed distribution of states (right side of Figure 1) by Levenberg-Marquardt [15] chi-square minimization of the unweighted error to the theoretical distribution Eq. (3). Using Eq. (2), and the parameters estimated as above, our belief that a gene belongs to the underlying states σ_0, σ_1, σ_r, given the $2^4 = 16$ observable states S, can now be expressed as:

$$P(\sigma_0|S)=\frac{P(S|\sigma_0)P(\sigma_0)}{P(S)}$$

$$P(\sigma_1|S)=\frac{P(S|\sigma_1)P(\sigma_1)}{P(S)}$$

$$P(\sigma_r|S)=\frac{P(S|\sigma_r)P(\sigma_r)}{P(S)}$$

Once the probability that a gene is in a certain biological state $\Sigma^i \in \sigma_1, \sigma_0, \sigma_r$ has been calculated for all varieties $i = 1...\upsilon$, one can calculate the probability that a gene exhibits a certain expression profile over a set of different varieties by taking the product

$$P\left(\sum^1,\ldots,\sum^\upsilon \middle| S^1,\ldots,S^\upsilon\right)=\prod_{i=1}^{\upsilon} P\left(\sum^i \middle| S^i\right). \quad (4)$$

In this way, the probabilistic state analysis also generates a clustering: For a given expression profile over the varieties, e.g. $\sigma^1_0\sigma^2_r\cdots\sigma^i_1$, we may extract those genes for which this expression profile is the most probable. In fact this is a "soft" clustering, in that an expression profile can belong to several clusters simultaneously with different probabilities. Moreover the genes clustered to a biologically interesting expression profile can be ranked by the probability of Eq. (4).

Experimental data preparation

All cells were grown in RPMI medium supplemented with 7.5% fetal calf serum, 10 mM HEPES, 2 mM pyruvate, 50 mM 2-mercaptoethanol and 50 mg gentamicin per ml (complete RPMI media) (all purchased from Life Technologies AB, T = E4by, Sweden) at 37 = B0C and 5% CO2. RNA was prepared using Trizol (GIBCO) and 7.5 = B5g of total RNA was annealed to a T7-oligo T primer by denaturation at 70 = B0C for 10 minutes followed by 10 minutes of incubation of the samples on ice. First strand synthesis was performed for 2 hours at 42 = B0C using 20 U of Superscript Reverse Transcriptase (GIBCO) in buffers and nucleotide mixes according to the manufacturers instructions. This was followed by a second strand synthesis for 2 hours at 16 = B0C, using RNAseH, E coli DNA polymerase I and E coli DNA ligase (all from GIBCO), according to the manufacturers instructions. The obtained double stranded cDNA was then blunted by the addition of 20 U of T4 DNA polymerase and incubation for 5 minutes at 16 = B0C. The material was then purified by Phenol:Cloroform:Isoamyl alcohol extraction followed by precipitation with NH4Ac and Ethanol. The cDNA was then used in an in vitro transcription reaction for 6 h at 37 = B0C using a T7 IVT kit and biotin labeled ribonucloetides. The obtained cRNA was purified from unincorporated nucleotides on a RNAeasy column (Qiagen). The eluted cRNA was then fragmented by incubation of the products for two hours in fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, 150 mM MgOAc). 20 = B5g of the final fragmented cRNA was then hybridized to affymetrix chip U74Av2 (Affymetrix) in 200 = B5l hybridization buffer (100 mM MES-buffer, pH 6.6, 1 M NaCl, 20 mM EDTA, 0.01 Herring sperm DNA (100 = B5 g/ml) and Acetylated BSA (500 = B5 g/ml) in an Affymetrix Gene Chip Hybridization oven 320. The chip was then developed by the addition of FITC-streptavidin followed by washing using an Affymetrix Gene Chip Fluidics Station 400. Scanning was performed using a Hewlett Packard Gene Array Scanner.

Results

To evaluate the method we used both real and synthetic data. The experimental data was generated with Affymetrix microarrays for the study of differentiating murine B-cells at different stages in the differentiation process. In this publication the data is only used to demonstrate the feasibility of the proposed method. The biological implications of this study are published elsewhere [20].

Synthetic data and the effect of correlation

For synthetic data, generated with the model parameters $P(\sigma_0) = 0.45$, $P(\sigma_1) = 0.35$, $P(\sigma_r) = 0.2$ and $P^i_{1\to 0} = P^i_{0\to 1} = 0.02$ for all samples i, parameter estimates are, as expected, given with low errors. The parameter values were chosen as typical values from the estimates on real data, see next section, and the result was verified for sample sizes $m = 4$, $m = 5$, and $m = 6$ (data not shown).

An assumption of simple model used to derive Eq. (3) is that randomly varying genes vary *independently* in the samples of a variety. Hence we investigated how severely this assumption influences the estimation of the model parameters.

To assess the influence of correlations between randomly varying genes we generated a data set consisting of four bits, i.e. samples, with the same parameters as above. In the random patterns a correlation was introduced between the third and fourth bit by changing the value of of the fourth to that of the third with a certain probability. We define this probability as the correlation factor. The correlation was introduced before distorting the patterns with error probabilities. We then plotted the mean error in the estimation of parameters over 500 runs of synthetically generated data for correlation factors in the range $\{0, 0.02,...,0.98\}$.

Figure 2 shows the error in the estimation of the parameters describing the underlying distribution. We notice that even for fully correlated patterns the estimation error is less than 20% of the correct values. The estimation of the probability for biologically varying genes is somewhat worse, for fully correlated patterns the error is almost 50%. For real data one can, however, expect a much smaller correlation. The average error in the estimates of the error probabilities, as seen in Figure 3, shows the expected behavior: The average error grows with the correlation for the uncorrelated samples, while the estimate for the correlated observations is almost unaffected. Intuitively, the model compensates for the correlation by increasing $P(\sigma_1)$ and $P(\sigma_0)$ as well as the error probabilities and lowering $P(\sigma_r)$. For correlation factors above 0.50, due to the compensation effect, the model deteriorates in explaining the data. This can be seen in the sum $P(\sigma_0) + P(\sigma_r) + P(\sigma_1)$ which initially drops from almost 1 to 0.99 as the correlation factor rises from 0 to 0.50 and then from 0.99 to 0.96 for correlation factors in the range 0.50 to 0.98 (data not shown). We hence conclude that it is reasonable not to impose the condition $P(\sigma_0) + P(\sigma_1) + P(\sigma_r) = 1$ in the model, as this sum indicates if samples are strongly correlated in genes whose expression vary around the threshold of discretization.

In summary, for not too large correlations in the biological variance the algorithm gives a good quantitative estimate of the model parameters. In the case of large correlations the qualitative picture given by the estimated parameters is still reliable.

Real data

Differentiating B-cells are characterized by phenotypic markers into different stages of development. Here we chose to study the expressional differences between three such stages; pro, pre and mature B-cells. For each of these three varieties we used four different cell lines arrested at the corresponding stage of development. Measurements we performed with Affymetrix array containing probesets for 12488 genes and ESTs on each sample. The discretization of expression levels was given by the Affymetrix GeneChip absent present calls [9].

Our algorithm was used to estimate the parameters $P(\sigma_1)$, $P(\sigma_0)$ and $P(\sigma_r)$, describing the biological distribution and the error probabilities (see Table 1). Theoretically, one expects the three biological probabilities to sum up to one. In our model, Eq. (3), we do not explicitly impose this condition. Nevertheless, the sum of the independently estimated parameters is close to one. This indicates that our model is a reasonable approximation of the biological system and the measurement process.

The error probabilities from Eq. (3) can be used as a consistency index for the samples in a given variety. In the last variety (mature B-cells) the maximum error probability is notably higher. This effect is likely to be explained by the different anatomical origins of the cell lines representing this group. No such differences exist in the other groups since they all originate in the bone marrow which is the only anatomical site for B cell development in the adult animal [16]. In contrast, the mature B cell can reside in several other sites such as spleen, lymph-nodes and intestine which may affect the gene expression profile in these cells [17,18]. With only four samples, it is not unlikely that these effects show up in the error probabilities and not only in the random variation parameter $P(\sigma_r)$.

Comparison to conventional t-test on known genes

To determine how well biologically relevant information can be extracted from the discretized data, we compare it

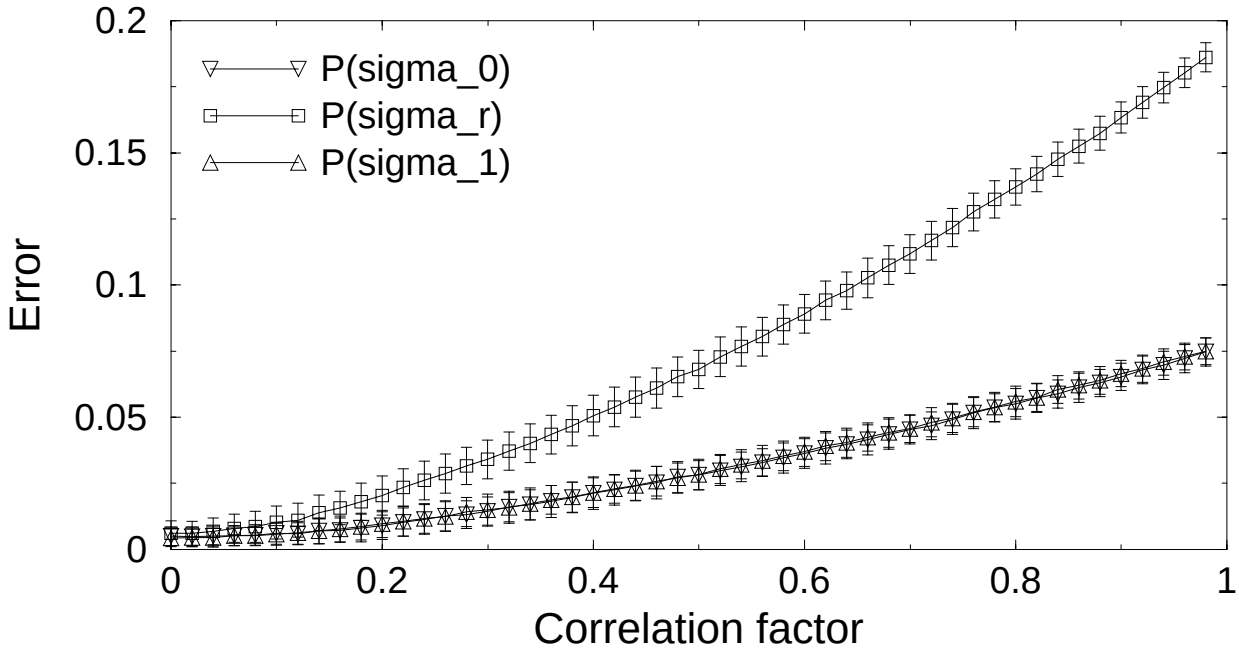

Figure 2
The average error in the estimation of the parameters $P(\sigma_1)$, $P(\sigma_0)$, $P(\sigma_r)$ are given as a function of correlation factor between the third and fourth bit. For correlation factors above 0.2 the error in $P(\sigma_r)$ rises considerably.

with another statistical method based on continuous expression values. We use our method to identify differences in gene expression between two varieties in the following way. A gene that goes up between variety *i* and variety *j* is characterized by the states σ_0^i, σ_1^j or σ_0^i, σ_r^j or σ_r^i, σ_1^j. Hence the belief that a gene goes up is given by the probability:

$$P(\text{up between variety } i \text{ and } j) = P(\sigma_0^i)P(\sigma_1^j) + P(\sigma_0^i)P(\sigma_r^j) + P(\sigma_r^i)P(\sigma_1^j),$$

suppressing the conditional probabilities, $P(\cdot|S)$, for brevity. Similarly, the belief that a gene goes down is given by the probability:

$$P(\text{down between variety } i \text{ and } j) = P(\sigma_1^i)P(\sigma_0^j) + P(\sigma_1^i)P(\sigma_r^j) + P(\sigma_r^i)P(\sigma_0^j).$$

Taking 1 - *P*(up) thus yields the Bayesian *p*-value of a gene going up. To answer the same question when working on continuous expression data one possibility is to employ a one sided two sample *t*-test in the Welch approximation of unknown variances in the varieties. This enables testing, for each gene, whether the mean of expression is higher or lower in one variety than in another. For comparison of these two approaches we selected a set of genes based on their well documented expression pattern and biological functions in the developing B lymphocyte [16,19]. Several of these are functionally linked since they participate directly in somatic DNA rearrangement events occurring specifically at the pre-B cell stage or participate in the regulation of genes involved in this process and thus display restricted expression patterns (pre-B specific). A second set of genes were selected based on their expression in cells that are either committed to the B lineage (B-lineage specific genes, in pre-B and B-cells) or non committed to this developmental pathway (Not in B-lineage, expressed in pro-B cells) [21].

The result of this comparison is presented in Table 2. For 14 out of the 22 genes the two methods completely agree. Out of these 14 only one (Mb1) does not match the expected target profile. For the other 8 genes, where the two methods yield different results, the probabilistic state analysis gives the expected answer in 5 cases, which should be compared to the two cases, where the *t*-test gives the right answer. In one case (rag-1), neither of the two methods gives the expected result.

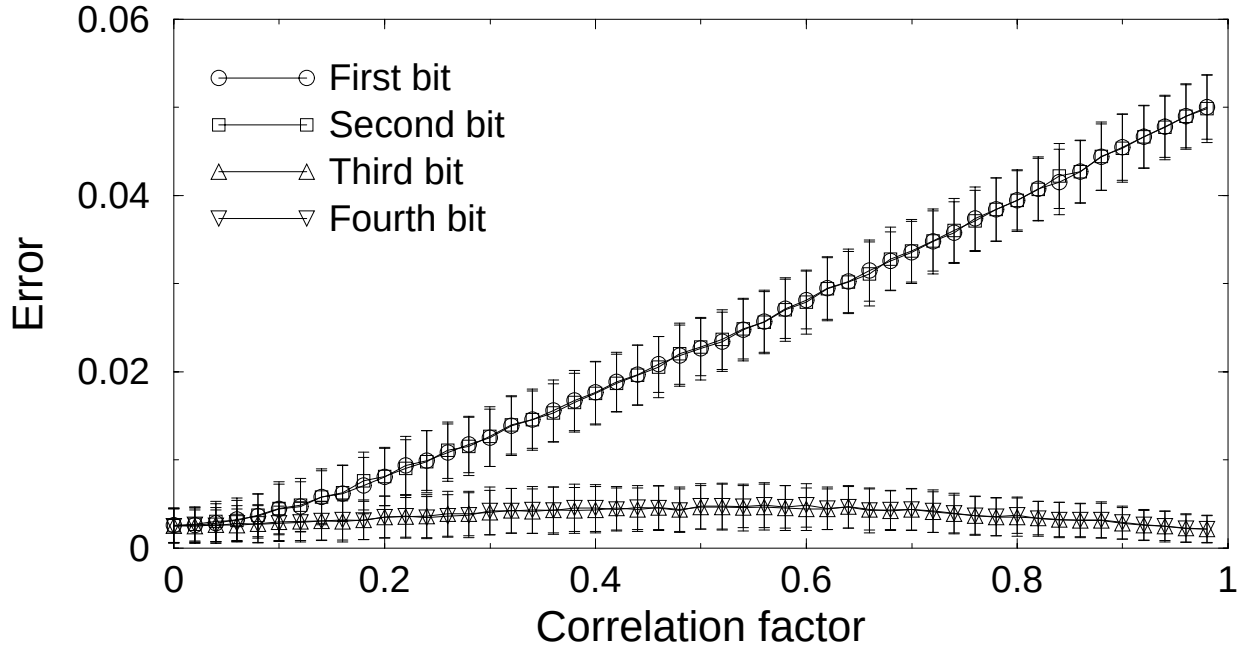

Figure 3

The average error in the estimation of the error probabilities $\left\{P^i_{0\to1}\right\}^4_{i=1}$. For correlation factors above 0.2 $P^1_{1\to0}$ and $P^2_{1\to0}$ are notably raised. Patterns were these bits deviate from the other two are then not considered as random but rather caused by an error. This effect could only be avoided by introducing extra parameters for correlation between bits.

Table 1: Typical paramter values. Summary of the estimated parameter values for the B-cell data. Pe refers to the set of error probabilities, i.e., $\left[P^i_{1\to0}, P^i_{0\to1}\right]^4_{i=1}$.

	$P(\sigma_1)$	$P(\sigma_0)$	$P(\sigma_r)$	Max P_e	Min P_e	Median P_e
Pro B-cells	0.405	0.460	0.135	0.035	0.0002	0.020
Pre B-cells	0.395	0.450	0.155	0.047	0.003	0.028
Full B-cells	0.343	0.471	0.186	0.073	0.0007	0.022

For the subset of genes considered here, our method has an advantage of 5 : 2 in giving the correct (i.e. expected) expression pattern. However, the number of samples is not big enough to draw firm conclusions from this result.

Conclusions

The method we have presented serves several purposes:

1. It gives a measure of the biological variation of the genes' expression in different varieties.

2. It estimates each hybridizations' global error probabilities. These parameters are very useful as they serve as quality/consistency indicators of the samples of each variety.

Table 2: t-test vs. probabilistic analysis of gene expression levels. The three groups I, II and III indicate the expressional changes between *Pro-B to Pre-B, Pre-B to Mature-B*, and *Pro-B to Mature-B* respectively. U stands for accepting the hypothesis up, D for down, and S (stable) if no hypothesis could be accepted on the 95% confidence level.

		Discrete			Continuous			
Gene	Id	I	II	III	I	II	III	Literature
Target profile	-	**S**	**U**	**U**	**S**	**U**	**U**	B-Cell specific
CD20	99446_at	S	U	U	S	U	U	
Spi	93657_at	S	U	U	S	S	S	
Target profile	-	**U**	**D**	**S**	**U**	**D**	**S**	Pre-B specific
Sox-4	160109_at	U	D	S	U	D	S	
lef-1	103628_at	U	D	S	U	S	S	
rag-1	93683_at	U	S	S	S	S	U	
VpreB	92972_at	U	D	S	U	D	S	
Lambda-5	99429_at	U	D	S	U	D	S	
Il-7 receptor	99030_at	U	D	S	U	S	S	
TdT	99030_at	U	D	S	U	S	S	
Target profile	-	**U**	**S**	**U**	**U**	**S**	**U**	B-lineage Specific
Bob-1	93915_at	U	S	U	U	S	U	
CD 19	99945_at	U	S	U	U	D	U	
Blnk	100771_at	U	S	U	U	S	U	
Pax-5	96993_at	U	S	U	U	S	U	
Blk	92359_at	U	S	U	U	S	U	
Mb-1	102778_at	U	D	S	U	D	S	
B29	161012_at	S	S	S	U	S	U	
CD24	100600_at	U	S	U	U	S	U	
Target profile	-	**D**	**S**	**D**	**D**	**S**	**D**	Not in B-lineage
Id-1	100050_at	D	S	D	D	S	D	
Fag-1	97974_at	S	D	D	D	S	D	
Il-3 receptor	94747_at	D	S	D	D	S	D	
CD 63	160493_at	D	S	D	D	S	D	
Gata-2	102789_at	D	S	D	D	D	D	

3. Given the parameters above, it estimates the probability of a gene belonging to each of the three groups σ_0, σ_r and σ_1. These probabilities in turn indicate weather the gene is likely to be below, fluctuating around or above the threshold of discretization.

4. Clustering of genes to expression profiles over a set of different varieties is achieved with Eq. (4). The probability, i.e. belief, that a gene belongs to a certain cluster is exactly quantified.

This novel approach is proven valuable for quantifying both data reliability and underlying gene expression in microarray experiments. Our method has been successfully applied in two different projects [22,20].

Authors' Contributions

SB and TB have contributed to the development of the model and implemented the algorithms. MS has contributed the experimental data and biological expertise.

Acknowledgments

SB and TB are supported by Knut and Alice Wallenberg Foundation through the SWEGENE consortium. MS is supported by the Swedish science council and the A. Österlund foundation. The authors also extend their gratitude to Carsten Peterson, Patrik Edén, Jari Häkkinen and Markus Rignér, for fruitful discussions and careful proofreading of the manuscript.

References

1. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D and Lockhart DJ: **Genome-Wide transcriptional analysis of the mitotic cell cycle.** *Mol Cel* 1998, **2:**65-73.
2. Eisen MB, Spellman PT, Brown PO and Botstein D: **Cluster Analysis of Genome Wide Expression Patterns.** *PNAS* 1998, **95:**14863-68.

3. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dimitrovsky E, Lander E and Golub TR: **Interpreting Gene Expression with Self-Organizing Maps.** *PNAS* 1999, **96:**2907-2912.
4. Törönen P, Kolehmainen M, Wong G and Castren E: **Analysis and Visualization Of Gene Expression Data Using Self-Organizing Maps.** *FEBS Lett* 1999, **451:**142-146.
5. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR and Fedoroff N: **Fundamental patterns underlying gene expression profiles: Simplicity from complexity.** *PNAS* 2000, **97:**8409-8414.
6. Alter O, Brown P and Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *PNAS* 2000, **97:**10101-10106.
7. Kerr MK, Martin M and Churchill GA: **Analysis of Variance for Gene Expression Microarray Data.** *J Comp Biol* 2000, **7:**819-837.
8. Kerr MK and Churchill GA: **Bootstrapping cluster analysis: Assessing the reliability of conclusions from micro-array experiments.** *PNAS* 2001, **98:**9061-8965.
9. **Affymetrix technical notes** [http://www.affymetrix.com]
10. Li F and Stormo GD: **Selection of optimal DNA oligos for gene expression arrays.** *Bioinformatics* 2001, **17:**1067-1076.
11. Friedman N, Linial M, Nachman I and Peér D: **Using Bayesian Networks to Analyze Expression Data.** *J Comp Biol* 2000, **7:**601-620.
12. Kauffman SA: **Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets.** *J Theor Biol* 1969, **22:**437.
13. Kauffman SA: **The Origins of Order.** *Oxford University Press* 1993.
14. Huang S: **Gene expression profiling, genetic networks and cellular states: an integrating concept of tumor genesis and drug discovery.** *J Mol Med* 1999, **77:**469-480.
15. Marquardt DW: *Journal of the Society for Industrial and applied Mathematics* **11:**431-441.
16. Gia P, ten Boekel E, Rolink AG and Melchers F: **B-cell development: a comparison between mouse and man.** *Immunol Today* 1998, **19:**480-5.
17. Rolink AG, Melchers F and Andersson J: **The transition from immature to mature B cells.** *Curr Top Microbiol Immunol* 1999, **246:**39-43.
18. Rolink AG, ten Boekel E, Yamagami T, Ceredig R, Andersson J and Melchers F: **B cell development in the mouse from early progenitors to mature B cells.** *Immunol Lett* 1999, **68:**89-93.
19. Liberg D and Sigvardsson M: **Transcriptional regulation in B cell differentiation.** *Crit Rev Immunol* 1999, **19:**127-53.
20. Tsapogas P, Breslin T, Bilke S, Lagergren A, Mnsson R, Liberg D, Peterson C and Sigvardsson M: **RNA analysis of B-cell lines arrested at defined stages of differentiation allows for an approximation of gene expression patterns during B cell development.** *J Leukoc Biol* .
21. Rolink AG, Schaniel C, Busslinger M, Nutt SL and Melchers F: **Fidelity and infidelity in commitment to B-lymphocyte lineage development.** *Immunol Rev* 2000, **175:**104-11.
22. Högerkorp CM, Bilke S, Breslin T, Ingvarsson S and Borrebaeck C: **CD44-stimulated human B cells express transcripts specifically involved in immunomodulation and inflammation as analyzed by DNA microarrays.** *Blood* 2003, **101:**2307-13.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral