Inference from Low Precision Transcriptome Data Representation

Salih Tuna · Mahesan Niranjan

Received: 3 December 2008 / Revised: 3 December 2008 / Accepted: 19 March 2009 / Published online: 22 April 2009

Abstract Microarray measurements are being widely used to infer gene functions, identify regulatory mechanisms and to predict phenotypes. These measurements are usually made and recorded to high numerical precision (e.g. 0.24601). However, aspects of the underlying biology, including mRNA molecules being highly unstable, being only available in very small copy numbers and the measurements usually being made over a heterogeneous population of cells, ought to make us sceptical about the reproducibility of these measurements and thus the numerical precisions reported. In this paper, we show that over a range of different procedures (classification, cluster analysis, detection of periodically expressed genes and the analysis of developmental time course data), the quality of inference from microarray data does not significantly degrade when the numerical precision is lowered by quantization. A surprising finding, with respect to classification problems, is that much of the discrimination is retained with numerical precision as low as binary (i.e. whether the gene is expressed or not). From this premise we show preliminary results that similarity metrics suitable for binary spaces, namely the Tanimoto metric used in chemoinformatics, can be successfully deployed to improve classification accuaracies of binarized transcriptome data.

Keywords Gene expression · Microarray · Quantization · Inference

1 Introduction

Microarray technology enables the simultaneous measurement of expression levels of thousands of genes in a single experiment. Work on developing diagnostic tools, inferring functions of unknown genes by observing similarity of expression patterns with genes of known function across a range of experiments, the detection of genes operating in response to specific environmental influences such as heat shock (e.g. Causton et al. [1]) and the use of high density arrays (e.g. Stolc et al. [2]) to detect alternative splicing action that offers enhanced functional diversity at the protein level are applications that follow from these high throughput measurements. Transcriptome measurements have been applied to a range of problems [3–8].

From a machine learning perspective, too, the availability of transcriptome data has caught the fascination of several researchers. Apart from the natural curiosity induced by the post-genomic era, more fundamental research challenges have been of interest. An example of this is the high dimension, small sample problem inherent in most transcriptome based classification problems that have been described in the literature. Owing to this, and other reasons, a plethora of machine learning (or statistical inference) methods have been developed and applied to transcriptome measurements.

We take a critical look at a particular aspect of the problem, that of the numerical precision with which mRNA concentration measurements can be utilized

Electronic supplementary material The online version of this article (doi: 10.1007/s11265-009-0363-2) contains supplementary material, which is available to authorized users.

S. Tuna · M. Niranjan (✉)
University of Southampton, Southampton, UK
e-mail: mn@ecs.soton.ac.uk

in making inferences. By numerical precision we refer to the difference between expresion levels quoted as 2.4601 and simply 2. Clearly, the former looks unrealistic and some truncation seems reliable. But how low can this precision be dropped?

Questioning such numerical precision is motivated by several biological considerations. Firstly, microarray hybridisation is carried out against mRNA samples taken from a population of cells from a biological sample of interest [9–11]. Except in few studies, different cells in such a population can potentially be in different states with respect to the expression of their genes. The few exceptions include forcing cells into synchrony for investigating cell cycle behaviour [5] and entraining cultured cells for observing circadian rhythm [12]. While there is apparent broad acceptance of synchronization in the literature, the topic is not without controversy [13]. Secondly, there is only a finite number of mRNA molecules in any particular cell. The total number of mRNA molecules is of the order of 50,000 in a mammalian cell, leading to an average of about ten molecules per cell for a gene of interest [11, 14]. With the underlying biology of such small numbers, pooling of cells and subsequent amplification of the extracted mRNA population can potentially give rise to a variable environment against which the arrays are set to hybridise [10]. Thirdly, in the cellular context from which it is extracted, mRNA is an inherently unstable molecule which is subject to decay at different rates [15, 16] often much faster than the decay of proteins. Thus the process of extracting cellular mRNA will be subject to significant variability prior to subsequent amplification processes. Fourthly mRNAs undergoing translation can be bound to several ribosomal complexes, the effect of which may be to restrict the availability of these molecules [17]. Thus the overall picture is quite different from one in which molecules of very high abundance, free-floating in the medium, are set to hybridise efficiently onto the probes of a microarray.

In a critical appraisal of the reproducibility and reliability of the microarray technology, Draghici et al. [18] conclude as follows: “...the existence and direction of gene expression changes can be reliably detected for the majority of genes. However, accurate measurements of absolute expression levels and the reliable detection of low abundance genes are currently beyond the reach of microarray technology.” Put another way, how similar are the transcriptome measurements of biological samples coming from the same tissue or cell culture? It is generally acknowledged that biological variability is high when compared to the technical variability of hybridizing amplified mRNA against different microarrays [41, 42].

An idea of the variation in biological replicates can be gained looking closely at studies comparing microarray measurements with quantitative PCR (qPCR) measurements. Tomayko et al. [44] compare mRNA measurements of 61 genes using microarray and qPCR. Figure 1b shows their results. When one measures the correlation between the two sources of measurements, taken across all the data, one obtain a correlation of 0.82, very close to what is reported by the authors.[1] However, a better description of the data is that it is bimodal with two outlier data points. When the up and down-regulated sets of genes are considered separately, with the outliers removed, the correlation are 0.36 and 0.47 respectively. This adds further weight to Draghici et al.’s observation that it is the up or down regulation information, rather than the precise expression levels, that is reliable in microarray data.

To overcome all these issues we propose the use of binary gene expression values and take a computational approach to explore the performance of quantized transcriptome measurements. We used Zhou et al.’s [19] binarization method to obtain different levels of quantization, and ask if researchers would have reached different conclusions, had they worked with data represented at lower precisions. We consider five inference problems, which are: (a) inferring gene function from gene expressions by posing a classification problem, using both two colour spotted array data and Affymetrix synthetic oligonucleotide data; (b) classification of phenotypes (medical conditions) from gene expressions; (c) function inference by cluster analysis; (d) detecting periodically expressed genes in the cell cycle and (e) analysing developmental time series. We report in this paper observations on a sample of problems to illustrate the critical question we pose and present some further analyses in on-line supplementary material accompanying this paper.

We stress that this work should not be considered as questioning the capability of microarrays as measurement devices to accurately measure mRNA concentrations. Several spike-in studies, and inter-platform comparisons [43] demonstrate excellent similarities between technical replicates (Fig. 1a). Our contention is that the biological variability of the source of the data is so highly variable that binary precision is best suited for inference.

[1] This data is not available in the public domain. We re-created the expression levels by measuring off an enlarged printout.

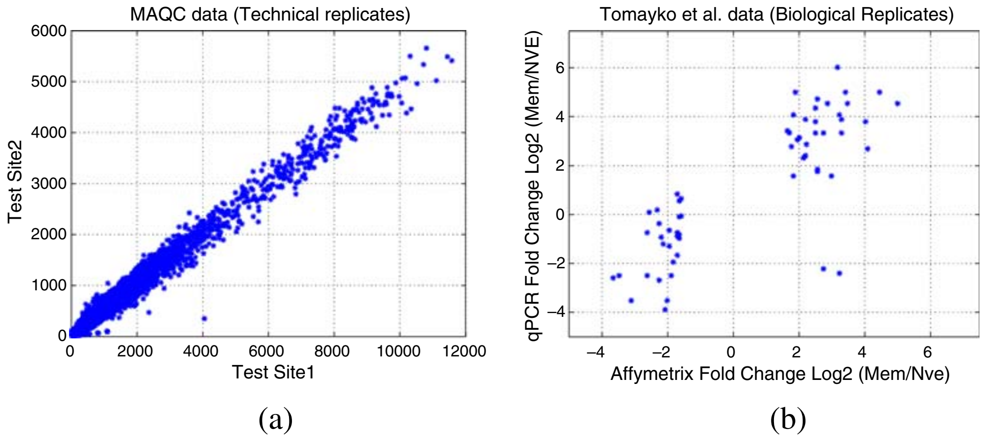

Figure 1 Comparison of reproducibility of mRNA measurements. (**a**) example of technical replicates of two Affymetrix arrays achieving a correlation of 0.99 taken from [43]. (**b**) example of biological variability from a comparison of microarray and qPCR measurements [44].

2 Quantization

Quantization is a topic that has been researched extensively in the context of rate distortion theory for data reduction, where the problem is one of faithful reproduction of data at a constrained data rate. To achieve optimal quantization, statistical properties of the signal and perceptual properties at the receiver (e.g. spectral weighting in speech coding) have to be exploited.

There is some work in the literature on quantizing gene expression data, starting from [20, 21], in which discrete expression levels were used to derive gene interaction networks in a Bayesian setting. They set up a multinomial probability distribution over three discrete levels −1 for under-expressed genes, +1 for over-expressed genes and 0 for the remainder, using threshold values of −0.5 and +0.5 on a logarithmic scale of expressions. While acknowledging that such a discretization procedure might suffer loss of information, Friedman et al. [20] also state that the discrete space has greater flexibility to capture combinatorial dependencies that cannot be extracted from linear Gaussian models operating on the raw data.

Camillo et al. [22] suggest quantization may be a means of reducing the probability of finding random associations between genes, given the fact that often the number of data points is smaller than the number of genes in microarray experiments. Similar to [20], they quantize to three levels, proposing a method to set thresholds based on an expected balance between true positives and false positives. It is argued that this approach yields better thresholds to discretize gene expression data, by comparing the resulting gene interaction networks on simulated data.

Shmulevich et al. [23] argue for binary representation and the use of Hamming distance as a measure of inverse similarity between gene expression signatures. They use a genetic algorithms approach to normalize microarray data, followed by a discretization procedure that looks for sudden changes in sorted gene expressions. The authors argue that different thresholds should be set for individual genes, as opposed to a global threshold for the entire experiment. Using two very small cancer datasets they demonstrate that discrimination between classes exist under Hamming distance between profiles. Surprisingly the paper does not offer a comparison of discrimination using the raw and quantized data. The two examples chosen are very easy classification tasks in which the classes are well separated. While the above work is the closest in literature to the central theme of our paper, our analyses report novel results. We analyse a range of inference problems in classification, clustering, periodicity determination and in the analysis of developmental time series, and systematically quantify what is lost when scaling down from raw down to binary precision. In all these cases we compare what might be achieved at binary levels with the original inference made by authors who published them. Our quantification includes appropriate measures of comparison, for example in quantifying the loss in classification, we use area under the ROC curve.

For quantizing gene expressions, we follow [19], who fit a mixture Gaussian model to log expression levels (Fig. 2). Our justification for choosing the method in [19] is that it is relatively more principled than other approaches to quantization reviewed above. Arbitrary thresholds set by other researchers are not necessarily transferable across different platforms or experiments due to variabilities induced by image processing and normalization, while the method in [19] depends on the underlying probability density of the expression levels and hence the idea is portable to any situation.

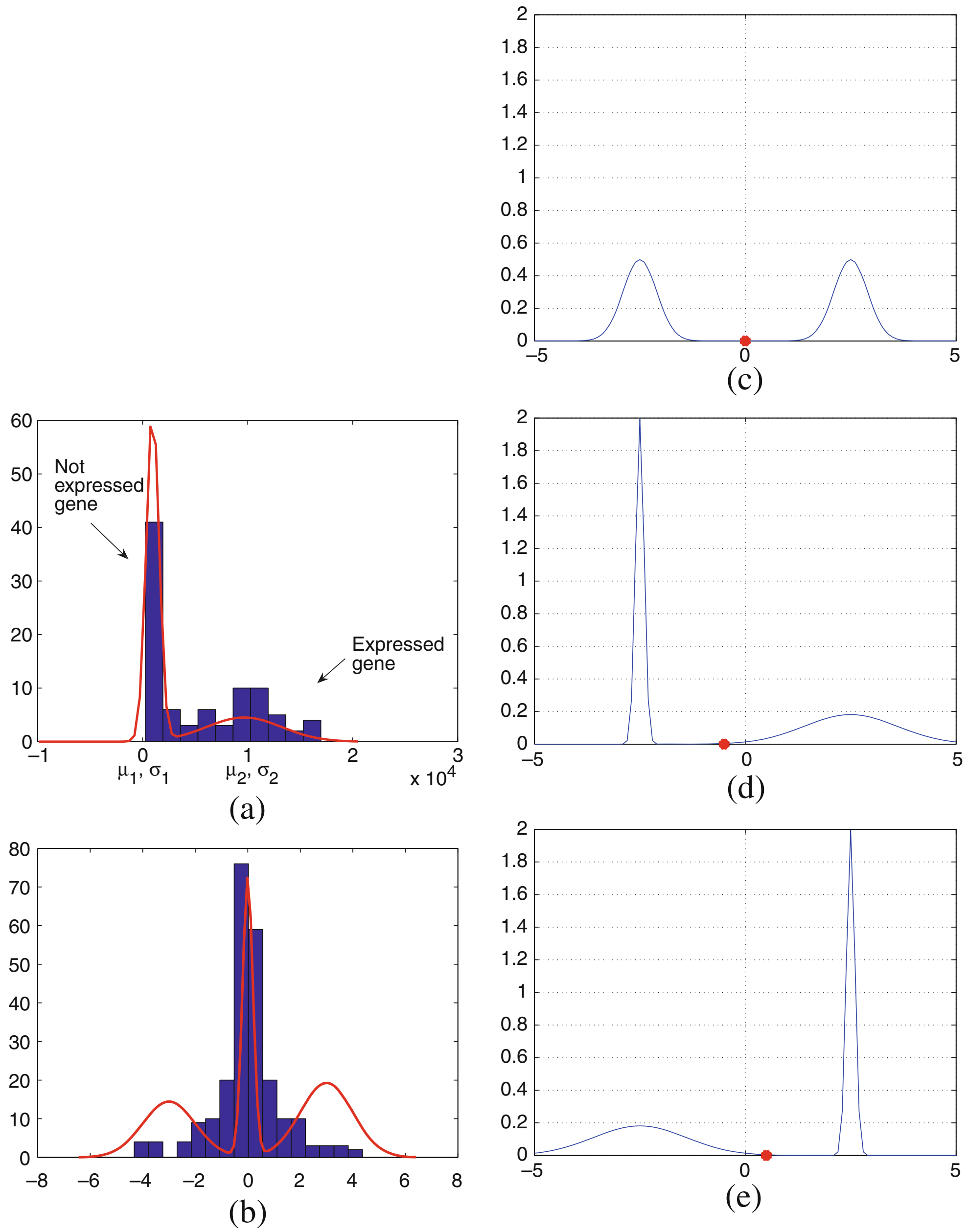

Figure 2 Mixture Gaussian distributions and corresponding histograms of gene expression levels for a subset of data taken from a sample of (**a**) *Affymetrix* gene expression measurements (Causton et al. [1]), and (**b**) from a cDNA experiment (Eisen et al. [29]). (**c**), (**d**) and (**e**) illustrate how a mixture model is used in setting a quantization threshold [19] (see Section 4). When the standard deviations of the expressed genes are the same, the quantization threshold is set at the average of the means. When the variances of the expressed and not expressed genes differ, the threshold is moved to take the variances into account.

3 Simulation Studies

Classification Two types of classification problems are usually posed on transcriptome data. Function prediction, where class labels correspond to genes in particular functional groups, and phenotype prediction where class labels correspond to different diagnostic outcomes in a clinical setting. The latter, particularly in the case of various types of cancer, has been studied widely while examples of the former include [4] and [24].

Using SVMs as classifiers, implemented in the `SVMlight` package [25], and using cross validation to optimise the free parameters, we measured classifier performance by means of areas under receiver operating characteristics curves. In all experiments, we confirmed that cross validation error of the baseline (i.e. unquantized raw data) is identical to what was claimed in each of the original publication.

As examples of function prediction by classification we used two datasets of classifying ribosomal genes from all others, following the work reported in [4]. One of these used cDNA spotted arrays and the other used synthetic oligo arrays (Affymetrix arrays). Missing values in the datasets were simply replaced by zeros. For phenotype classification we considered the well known problem of molecular classification of two types

Table 1 Loss of discriminability in a sample of classification problems when expression data is quantized to three and two levels.

Dataset	Raw data	3 level of quantization	Binary
Golub et al. [26]: Leukhemia data, 5000 genes, 38 ALL vs 37 AML samples	0.92 ± 0.05	0.89 ± 0.06	0.89 ± 0.07
Ramaswamy et al. [27]: Cancer (190) vs non-cancer (66) classification; 7000 genes	0.90 ± 0.03	0.89 ± 0.03	0.90 ± 0.04
Brown et al. [4]: Ribosomal genes (121) vs non-ribosomal (2000) genes in yeast from 79 different hybridizations; cDNA arrays	0.99 ± 0.004	0.99 ± 0.001	0.99 ± 0.001
Causton et al. [1]: Classifying ribosomal genes as above, but with 45 hybridizations using Affymetrix arrays	0.95 ± 0.02	0.96 ± 0.01	0.95 ± 0.01

Averages and standard deviations across 25 random bootstrap partitions of areas under the receiver operating characteristics curve are shown for a sample of problems.

of cancer (ALL/AML) considered in [26] and a similar problem, the GCM problem, considered in [27].

Table 1 shows the result for four of the datasets that we worked with, confirming that even under extreme levels of quantization discriminability between these classes is retained. A similar conclusion can be reached by looking at Fig. 3, showing the ROC curves with original and quantized data for one of the ribosome classification problems [4].

Clustering The most widely used inference tool in transcriptome analysis is clustering. One hypothesizes that genes that are co-regulated or those that have similar biological functions might show similar expression profiles under different hybridisation conditions, and hence may be found in the same clusters. To demonstrate that inferences made from cluster analysis do not degrade with lower precision data, we adopted two computational strategies. We took several published microarray cluster studies and computed the average pairwise correlation of gene profiles, where the pairs were taken from within and across clusters. Figure 4 shows that the the average within cluster correlation is much higher than the average correlation of pairs of genes taken from outside clusters. The discrimination is not affected significantly with quantization of the data down to binary precision.

As a second illustrative example, we analysed the clusters published in [29]. For the ten clusters identified in this work, we computed average pairwise within and cross class correlations for every pair of clusters. The resulting 10 × 10 matrix is shown as an intensity plot in Fig. 5. For two of these, identified as clusters *B* and *C* in the paper, we compared the histograms of within cluster and cross cluster pairwise gene correlations. These two distributions being well separated is an indication of how well-clustered the two groups of genes are, in the space of expression profiles. We attempted to quantify how much the separation between these two distributions degrade when the data is quantized to

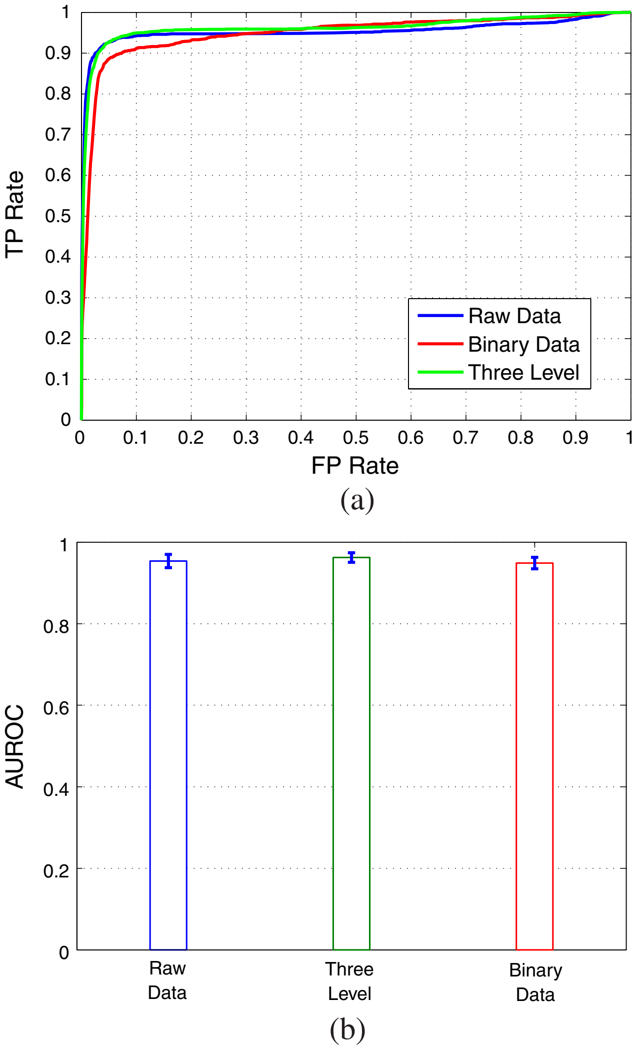

Figure 3 Change in classification performance with progressive quantization of gene expression levels for the problem of discriminating ribosomal yeast genes from data published by Causton et al. [1]. Receiver operating characteristic curves and areas under the curves, averaged over 25 bootstrap partitions of the data, are shown in (**a**) and (**b**) respectively. Error bars over these partitions are also shown in (**b**).

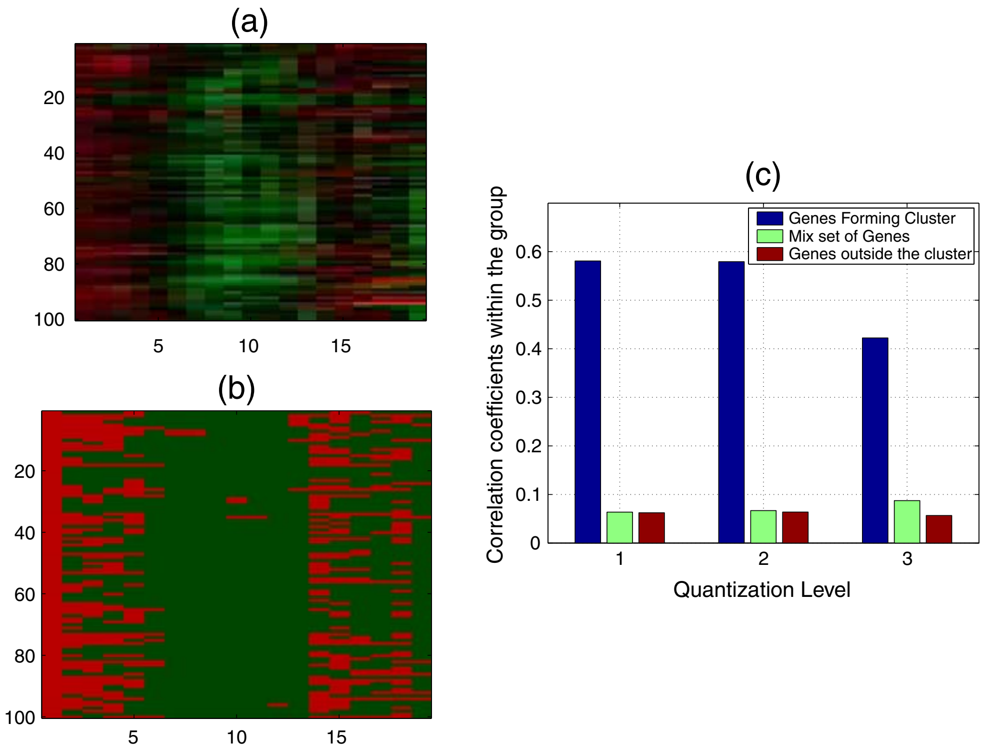

Figure 4 Average within and cross group correlations for a cluster of genes taken from Iyer et al. [28]'s study of human fibroblast response to serum. (**a**) and (**b**) are the expression levels of an identified cluster of 100 genes, with raw and binary-quantized data. (**c**) shows correlations, illustrating that the average within group correlations stay much higher than cross group correlations even under extreme quantizations. Quantization levels 1—raw data, 2—three levels (+1, 0 and −1), 3—binary.

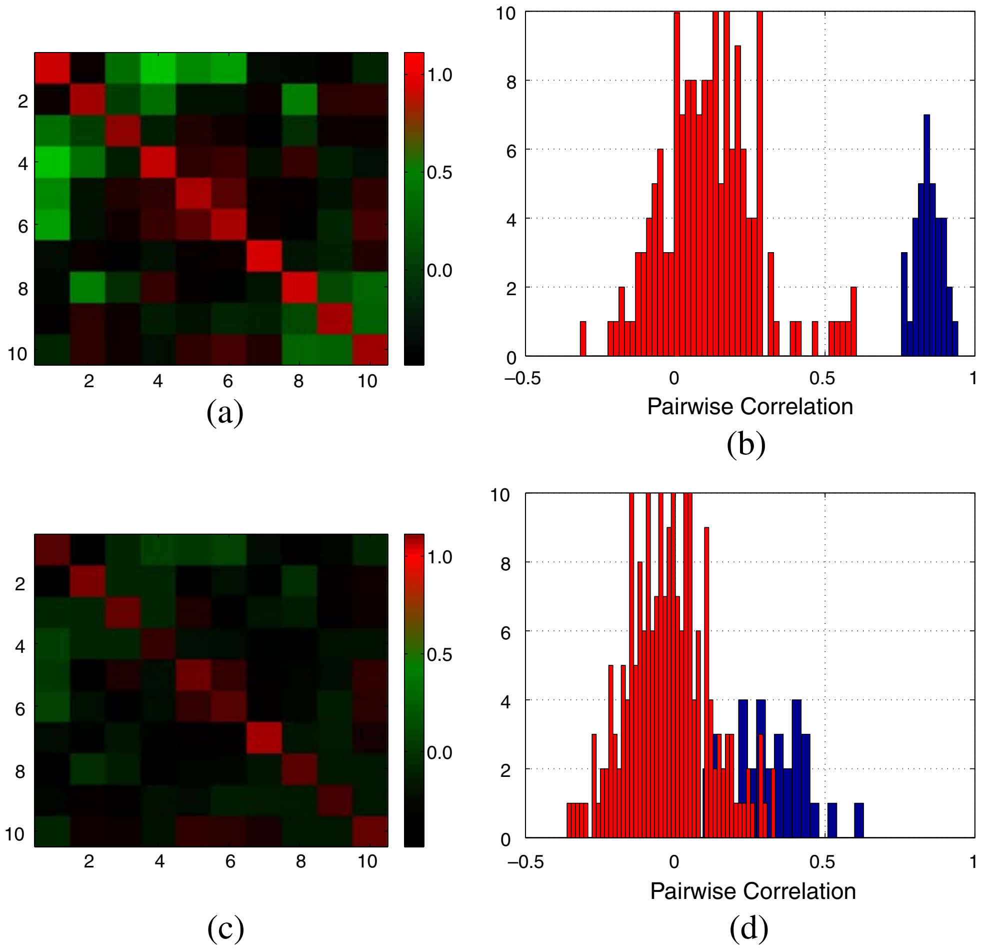

Figure 5 Average pairwise correlations, within and cross-group, of ten clusters taken from Eisen et al. [29], shown as intensity plots. (**a**) and (**c**) are 10 × 10 average correlation matrices computed using the raw expression levels and binary-quantized expression levels respectively. (**b**) shows within group and cross group correlations of genes in clusters identified by labels *B* and *C* in [29] as histograms. (**d**) shows the same histograms when the data is quantized to binary precision.

different levels. A convenient way of quantifying the separation between distributions is the Fisher ratio,

$$P(g, c) = \frac{\text{abs}(\mu_1(\text{g}) - \mu_2(\text{g}))}{\sigma_1(g) + \sigma_2(g)} \quad (1)$$

which has been used in many other similar contexts, for example in selecting discriminant genes, where $[\mu_1(g), \sigma_1(g)]$ and $[\mu_2(g), \sigma_2(g)]$ are the means and standard deviations of the two distributions. We find a reduction in Fisher ratio from 3.85 to 1.25 when the data was quantized from its original to binary precision and some overlap between the distributions is seen in the histograms. How much discriminability has been lost? The Fisher ratio does not give an intuitive picture of this loss. In Fig. 6 we give a simulation to gain a feel for this by comparing the Fisher ratios and areas under receiver operating characteristics curves (AUROC) for randomly chosen one dimensional Gaussian distributions. AUROC, effective in quantifying classifier performance, has a useful statistical meaning: when one is presented with two data points, one from each class of a two class classification problem, the AUROC is the probability the classification system under study correctly ranks them. In the clustering context we have here, given two pairs of genes, one pair from the same cluster and the other pair formed by genes from different clusters, this figure is the probability we would correctly rank the pairs as to their likelihood of having come from the same cluster. Given that, the observed worst case Fisher ratio of 1.25, under extreme quantization, corresponds to an AUROC of 0.95, meaning that only 5% of the genes will be grouped into wrong clusters, had we worked with binarized, rather than the original, data.

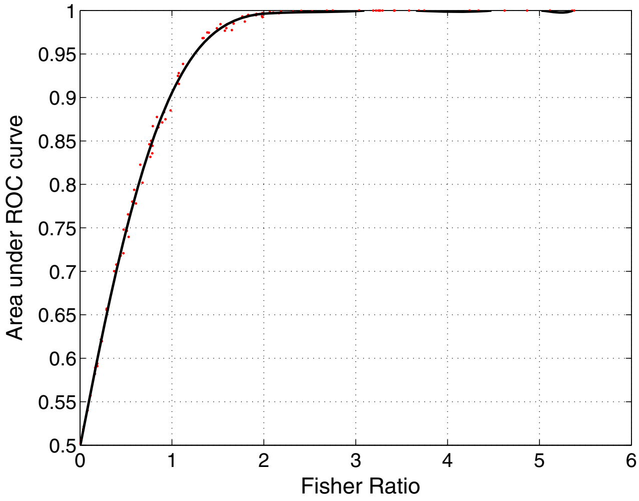

Figure 6 In comparing Fisher ratios with area under receiver operating characteristics curves (AUROC), of interest is how much discrimination is lost when the Fisher ratio between clusters reduces (from 3.85 to 1.25, in the example considered) as a result of quantization. We randomly generated several pairs of one dimensional Gaussian densities and measured the two figures of merit for their separation. The points on the scatter diagram correspond to pairs of Gaussians and the continuous line is an interpolation through them, obtained by curve fitting. Note that at a Fisher ratio of 1.25, AUROC has only reduced to 0.95, demonstrating that significant discriminability is retained between the clusters.

In a second computational strategy, we pooled genes from published clusters and applied K-Means clustering algorithm, at raw and binarized precisions, and compared the resulting cluster memberships. To quantify the overlap between genes in a particular cluster, we used the $F1$ measure, used widely in information retrieval problems (more details in supplementary material). This analysis confirmed significant overlap in membership, results of which are presented as supplemental information.

Periodic Expression We now consider the detection of periodically expressed genes, using a recent yeast cell cycle data from [30]. We took a subset of genes identified as cell cycle regulated, with peak expression in the S phase of the cycle. Our objective is to show how the ability to detect periodicity in the expression of these genes degrades with quantization of the data. We adopted a computational strategy similar to that used in clustering above, and measured the average pairwise correlation amongst three groups of genes: (a) the 99 genes which are known to be periodically expressed, yielding an average correlation measure, averaged across $99 \times 98/2 = 4851$ pairwise correlations; (b) similar average correlation across an arbitrary group of 100 genes taken from the dataset, but not overlapping with those in group (a); and (c) average correlation between the above 99 genes and 99×100 genes whereby we picked 100 genes at random to correlate against the 99 above. For correct detection of periodically expressed genes we would expect group (a) to show higher average correlation than those of groups (b) and (c). Of interest is whether the average correlation amongst group (a) genes continues to be higher than the other two groups under increasing levels of quantization.

We find (Fig. 7) that the correlation difference we measure, i.e. differences between within class correlation of the periodically expressed genes from those for the other two groups (mixture of periodic and aperiodic genes and the random set of genes) do not change. Thus even under such coarse quantization, we would have picked out these genes as expressed in regulation with

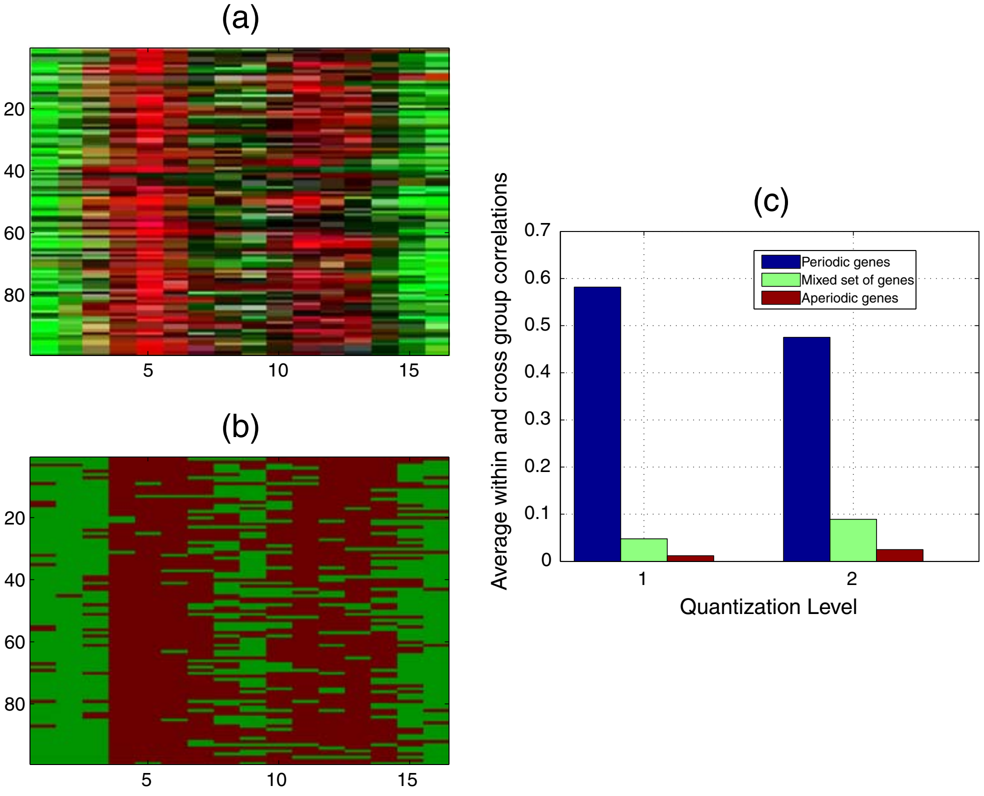

Figure 7 Expression profiles of a subset of periodically expressed genes, (**a**), and binary expression profiles after coarse quantization, (**b**). (**c**) shows the within class average pairwise correlation for three groups of genes considered (see text), showing that the discriminability of the set of periodic genes from the remainder is robust enough to be maintained at low precisions of the expression levels. Quantization levels one and two refer to the use of raw data and binary levels +1, and −1.

the cell cycle in the *S* phase. Here our demonstration of the effect of quantization on periodicity determination is based on correlation, within and across groups of genes identified as periodic. We have not re-computed the Fourier transform with binarized data. This is because the Fourier transform is an expansion in terms of orthogonal basis functions, correlating data against sinusoids of varying frequencies, and thus should produce the same results.

Developmental Time Series In a recent study [31], the authors analyse significant changes in gene expression during embryonic development of the fruit fly *Drosophila melanogaster*. To detect significant changes they use local convolution with two step functions:

- $[+1, +1, +1, +1, -1, -1, -1, -1]$, to detect down-regulation; and
- $[-1, -1, -1, -1, +1, +1, +1, +1]$, to detect upregulation.

The numbers of genes that undergo significant changes in expression give a picture of major regulatory changes during the stages of development. We re-analyzed this data[2] at the original precision and after discretizing it to binary precision. Figure 8 shows this comparison, demonstrating that the numbers of genes detected as significantly up-regulated (or down-regulated) along the developmental time-course of interest is very much the same at the lowest possible precision.

Classification in Binary Spaces Noting that the same inference may be made at lower precision, binary precision in particular, leads to two other questions. Firstly, we see a plethora of sophisticated methods for microarray gene expression analysis. Indeed, it would be hard to find a statistical inference model that has not

[2] As an aside, with respect to the use of raw precision, we noted that we were unable to achieve an exact reproduction of of the results reported in [31], in terms of the numbers of genes detected as increasing / decreasing in expression levels. We believe, this is mainly because of the step function with which local correlation is taken. The step function used in [31] ranges between 0 and 1, while ours takes values −1 and +1. When 0 is the lower figure any variation on the signal is suppressed (i.e. multiplying any number by zero returns a zero). This difference does not affect the argument developed in our paper, however.

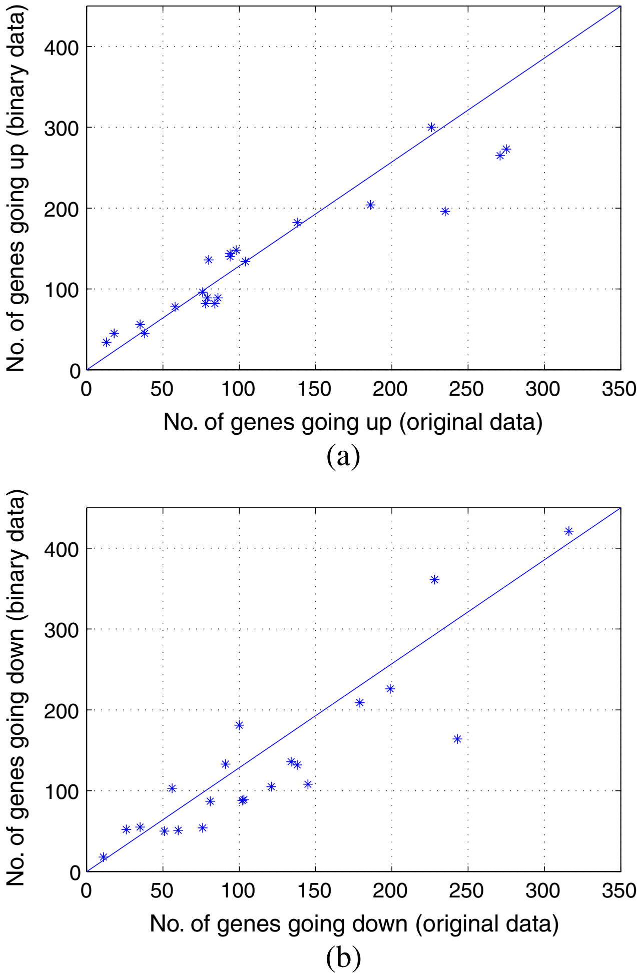

Figure 8 Comparison of the numbers of significantly upregulated, (**a**), and significantly downregulated, (**b**), genes at different stages of development, using raw gene expression measurements and binary quantized expression levels.

been applied to microarray data in recent years. One is led to ask if much of this falls into the category of "cracking a nut with a sledgehammer". In a recent study comparing methods of detecting cell cycle regulated genes [32], it was found that no sophisticated algorithm outperformed the original Fourier transform based method used by Spellman et al. [5]. Secondly, we could expect algorithms that are specifically designed to deal with low precision (mainly binary) data, either for computational or for performance reasons, to perform well in transcriptome based inference problems. An example of this is the use of the Tanimoto similarity in matching chemical fingerprints, and the incorporation of such a distance measure in a kernel classifier, as has been attempted by Trotter [33].

Table 2 shows results of classification experiments comparing the use of Tanimoto kernel with other classifiers (see Section 4). The two microarray studies chosen are ones for which baseline performance is not too close to perfect classification. We have also included results from a chemoinformatics problem for reference [34]. This is a two-class classification study that used 5747 training patterns of chemical fingerprints in 992 binary dimensions, and the task was to classify molecules with drug-like properties. Note the Tanimoto kernel classifier has no free parameters to tune, giving it a distinct advantage where applicable. The results suggest that the observations seen in the chemoinformatics area of the suitability of this metric for high dimensional binary problems translates to gene expression measurements as well. Quantization does slightly reduce the performances but these are recovered by a method suitable for high dimensional binary spaces.

Table 2 Comparison of Tanimoto kernel with linear and RBF kernels in an SVM classifier with quantized input data.

Dataset	Kernel	Parameters	AUROC
Chemoinformatics	Linear	$C = \infty$	0.87 ± 0.02
	RBF	$\sigma = \sqrt{\frac{m}{2}}, C = \infty$	0.88 ± 0.02
	Tanimoto		0.91 ± 0.02
	Barcode		0.79 ± 0.01
Alon et al. [6]	Linear(raw data)	$C = \infty$	0.87 ± 0.10
	Linear	$C = \infty$	0.85 ± 0.06
	RBF	$\sigma = \sqrt{\frac{m}{2}}, C = \infty$	0.86 ± 0.06
	Tanimoto		0.88 ± 0.05
	Barcode		0.82 ± 0.10
Pomeroy et al. [40]	Linear(raw data)	$C = \infty$	0.72 ± 0.22
	Linear	$C = \infty$	0.54 ± 0.46
	RBF	$\sigma = \sqrt{\frac{m}{2}}, C = \infty$	0.59 ± 0.27
	Tanimoto		0.95 ± 0.07
	Barcode		0.94 ± 0.05

RBF width σ is set as a function of the total number of training patterns m. Barcode refers to a distance to template classifier used in [36] where the template is fixed at class means.

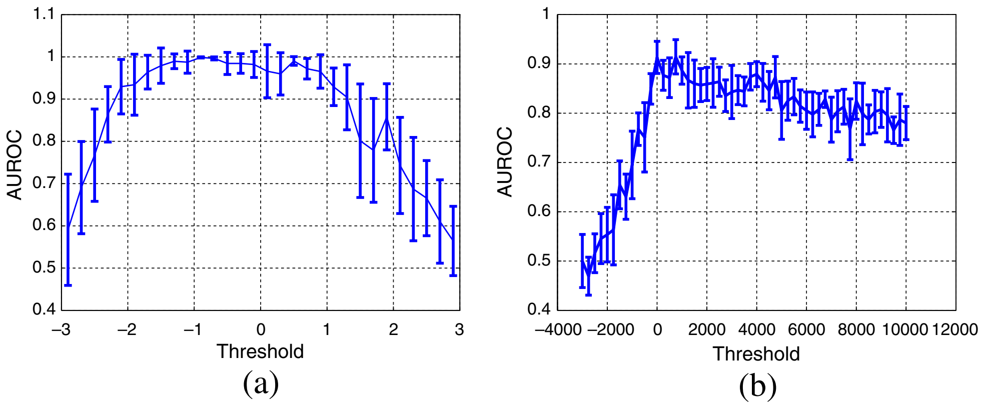

Figure 9 AUROC results when gene expression data is binarized by using a global threshold. There is a wide range of thresholds over which classifier performance is very similar. Note the dynamic ranges are different because (**a**) is on data supplied as log expression levels and (**b**) is from raw expression levels. This distinction is not relevant for the point we make in this paper. Two examples are shown from (**a**) [4] and (**b**) [27].

4 Methods

The mixture Gaussian model for quantization is

$$p(x) = \sum_{j=1}^{M} \lambda_j \mathsf{N}\left(\mu_j, \sigma_j\right)$$

where $p(x)$ is the probability density of gene expression measurement, M, the number of mixture components, and $\mathsf{N}(\mu, \sigma)$ is a Gaussian density of mean μ and standard deviation σ. Fitting such a model is by standard maximum likelihood techniques, and we used the gmm function in NETLAB software http://www.ncrg.aston.ac.uk for this purpose. We used two and three component mixtures mostly, corresponding to $M = 2$ and $M = 3$ in the above equation. After learning parameters of the model, a threshold T is chosen as:

$$T = 0.5 \ \{\mu_1 + \sigma_1 + \mu_2 - \sigma_2\}$$

to achieve binary quantization. For three level quantization, we fit a model of three Gaussian components, ordered them by their means and selected two thresholds between adjacent Gaussians using the above formula.

Note there is some flexibility in designing a binarization scheme: (a) a global threshold obtained by pooling the expression levels of all the genes in all the arrays of an experiment and fitting a mixture Gaussian model to the pooled data; (b) estimate a quantization threshold for each gene across the different hybridizations; or (c) quantize gene expressions on an array-by-array basis. An example of the last of the above is the use of present/absent calls from Affymetrix arrays to determine if a gene is expressed or not, where individual arrays are processed independent of each other. For the main message conveyed in this paper, that low precision representations still carry much of the information needed for inference, the different quantization strategies do not make a big difference. We illustrate this in Fig. 9, by scanning a range of thresholds for global quantization, we are able to observe a wide range of threshold settings for which discriminant performance remains high. Our method of choice, however, is gene-by-gene quantization. This is because different genes show different expression levels in cells. Genes encoding transcription factors, for example, are known to be expressed at very low copy numbers, so a low threshold is required to detect their presence for inference.

Tanimoto distance is given by $c/\,(a + b\, - c)$, where c is the number of genes expressed in common between the two profiles, and a and b are the numbers of genes expressed in each of the two profiles individually. Pairwise similarities between items of data computed by the above formula were used to form elements of the kernel matrix and input to the SVM optimiser. This was implemented within a MATLAB SVM package [35].

5 Conclusion and Discussion

Transcriptome measurements are recorded, reported and archived on a very large scale. Public availability of archived data has triggered extensive research into sophisticated computational algorithms for making functional inferences and detecting disease associated biomarkers from gene expressions. We demonstrate here that the inferences drawn from such data are largely unaffected when the precision of measurement is dropped, often down to binary levels. This observation is consistent with the underlying biology of messenger RNA molecules: that they are unstable, much regulation takes place post-transcriptionally and transcriptome measurements are usually done with a population of cells, averaging out inter-cell variabilities. The seemingly high precisions in measurements reported should be regarded as artefacts of measurement systems, such as image processing and normalizations.

A very recent study by Zilliox et al. [36] carries similar ideas to ours in advancing what the authors call a "bar code" for microarray data, essentially suggesting advantages in binary representations of microarray data. However, several differences between that paper and our work should be noted. Firstly, Zilliox et al. do not make robust direct comparisons between inference drawn from gene expressions at raw precision and expression levels quantized to binary precision. The problems on which the advantage of working with binary precision is demonstrated, namely the tissue prediction problem is different from the original class prediction problems for which the array experiments were designed. Secondly, the performance gains reported by these authors is carried out against a method that cannot be regarded as state-of-the-art. The method of Predictive Analysis of Microarrays (PAM) [37] does not come out as a high performing method in a systematic study conducted in [38]. Thus the comparison in [36] is against a very weak baseline. Thirdly, the points we make about negligible loss of the quality of inference at low precisions covers a wider range of problems, including, for example, the analysis of developmental time-course data. Fourthly, we use performance measures that are appropriate for the problems at hand, for example, the area under the ROC curve, rather than error rates, for classifier performance. Finally, where we demonstrate that the binary representation may also lead to performance gains, we explicitly take advantage of properties of high dimensional binary spaces, by use of a specific distance measure, namely the Tanimoto distance, integrated in a kernel discriminant (or Support Vector machine, SVM) framework. The comparison suggests that the Euclidean distance to a quantized class mean vector (or bar code) actually *does not* carry the performance advantage claimed in [36].

As a closing remark, we stress that the limitation we observe in this paper is about how precision of representing gene expression influences the inference one can make from transcriptome data, and *not* about the possibility or practicality of making such measurements themselves. Studies have shown, for example with spike-in data of known concentrations, that microarray measurements can be made over a wide range of concentrations and our work should not be seen as questioning their validity. Our claim, instead, is that though the measurements themselves may be precise, the inferences drawn from them do not change at lower precisions. We have shown this via a range of examples in this paper, sufficient to illustrate the point, but have now begun a systematic study of all inference problems that have been posed on the data archived at the European Bioinformatics Institute's ArrayExpress repository. Results described in [39] are also of interest in the context of our findings, where the authors report that binarization of data has been helpful in improving performance of classification systems on some benchmark classification tasks. While this cannot be universally true, we believe it is worth exploring further in the context of transcriptomic data due to properties of the underlying biology of unstable molecules, available only in low copy numbers across a population of cells. It is surprising indeed that the plethora of inferential tools developed in the literature, each claiming to out-perform a sample of competing methods, appear to have overlooked such fundamental properties of the source of the data.

Acknowledgements We are grateful to Anastasia Samsonova and Guido Sanguinetti for helpful comments on the manuscript.

References

1. Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., et al. (2001). Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell, 12*(2), 323–337.
2. Stolc, V., Samanta, M. P., Tongprasit, W., Sethi, H., Liang, S., Nelson, D. C., et al. (2005). Identification of transcribed sequences in *Arabidopsis thaliana* by using high-resolution genome tiling arrays. *PNAS, 102*(12), 4453–4458. doi:10.1073/pnas.0408203102.
3. Alter, O., Brown, P. O., & Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *PNAS, 97*(18), 10101–10106. doi:10.1073/pnas.97.18.10101.
4. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., et al. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS, 97*(1), 262–267. doi:10.1073/pnas.97.1.262.
5. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., et al. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell, 9*(12), 3273–3297.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon

tissues probed by oligonucleotide arrays. *PNAS, 96*(12), 6745–6750. doi:10.1073/pnas.96.12.6745.

7. Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D., & Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: Prostate cancer-associated genes. *Genome Research, 9*(12), 1198–1203. doi:10.1101/gr.9.12.1198.
8. Califano, A., Stolovitzky, G., & Tu, Y. (2000). Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the eighth international conference on intelligent systems for molecular biology* (pp. 75–85). ISBN 1-57735-115-0.
9. Levsky, J. M., Shenoy, S. M., Pezo, R. C., & Singer, R. H. (2002). Single-cell gene expression profiling. *Science, 297*(5582), 836.
10. Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science, 297*(5584), 1183–1186.
11. Levsky, J. M., & Singer, R. H. (2003). Gene expression and the myth of the average cell. *Trends in Cell Biology, 13*(1), 4–6.
12. Storch, K.-F., Lipan, O., Leykin, I., Viswanathan, N., Davis, F. C., Wong, W. H., et al. (2002). Extensive and divergent circadian gene expression in liver and heart. *Nature, 417*, 78–83.
13. Cooper, S. (2004). Rejoinder: Whole-culture synchronization cannot, and does not, synchronize cells. *Trends in Biotechnology, 22*(6). doi:10.1016/j.tibtech.2004.04.011.
14. Lockhart, D. J., & Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature, 405*(6788), 827–836.
15. Iyer, V., & Struhl, K. (1996). Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *PNAS, 93*(11), 5208—5212.
16. Hume, D. A. (2000). Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. *Blood, 96*(7), 2323.
17. Brown, T. A. (1999). *Genomes*. Oxford: Bios Scientific. ISBN 1 85996 201 7.
18. Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics, 22*, 101–109.
19. Zhou, X., Wang, X., & Dougherty, E. R. (2003). Binarization of microarray data on the basis of a mixture model. *Molecular Cancer Therapeutics, 2*(7), 679–684.
20. Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computer Biology, 7*(3–4), 601–620.
21. Pe'er, D., Regev, A., Elidan, G., & Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics, 17*(Suppl 1), S215–224. doi:10.1093/bioinformatics/17.suppl1.S215.
22. Di Camillo, B., Sanchez-Cabo, F., Toffolo, G., Nair, S., Trajanoski, Z., & Cobelli, C. (2005). A quantization method based on threshold optimization for microarray short time series. *BMC Bioinformatics, 6*(Suppl 4), S11. doi:10.1186/1471-2105-6-S4-S11.
23. Shmulevich, I., & Zhang, W. (2002). Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics, 18*(4), 555–565. doi:10.1093/bioinformatics/18.4.555.
24. Samsonova, A. A., Niranjan, M., Russell, S., & Brazma, A. (2007). Prediction of gene expression in embryonic structures of *Drosophila melanogaster*. *PloS Computational Biology, 3*(7:e144), 1360–1372. doi:10.1371/journal.pcbi.0030144.
25. Joachims, T. (1999). Making large-scale SVM learning practical. In B. Scholkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods–support vector learning*. Cambridge: MIT Press.
26. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531–537. doi:10.1126/science.286.5439.531.
27. Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS, 98*(26), 15149–15154. doi:10.1073/pnas.211566398.
28. Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., et al. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science, 283*(5398), 83–87. doi:10.1126/science.283.5398.83.
29. Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS, 95*(25), 14863–14868. doi:10.1073/pnas.95.25.14863.
30. de Lichtenberg, U., Wernersson, R., Jensen, T. S., Nielsen, H. B., Fausboll, A., Schmidt, P., et al. (2005). New weakly expressed cell cycle-regulated genes in yeast. *Yeast, 22*(15), 1191–1201.
31. Hooper, S. D., Boue, S., Krause, R., Jensen, L. J., Mason, C. E., Ghanim, M., et al. (2007). Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis. *Molecular Systems Biology, 3*. doi:10.1038/msb4100112.
32. de Lichtenberg, U., Jensen, L. J., Fausboll, A., Jensen, T. S., Bork, P., & Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics, 21*(7), 1164–1171. doi:10.1093/bioinformatics/bti093.
33. Trotter, M. W. B. (2006). *Support vector machines for drug discovery*. Ph.D. thesis, University College London, UK.
34. Rhodes, N., Willett, P., Dunber J. B., & Humblet C. (2000). Bit-string methods for selective compound acquisition. *Journal of Chemical Information and Computer Sciences, 40*, 210–214.
35. Gunn, S. R. (1997). *Support vector machines for classification and regression*. Technical Report, University of Southampton. http://www.isis.ecs.soton.ac.uk/isystems/kernel/.
36. Zilliox, M. J., & Irizarry, R. A. (2007). A gene expression bar code for microarray data. *Nature Methods, 4*(11), 911–913. doi:10.1038/NMETH1102.
37. Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS, 99*(10), 6567–6572. doi:10.1073/pnas.082099299.
38. Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics, 20*(18), 3583–3593. doi:10.1093/bioinformatics/bth447.
39. Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *International conference on machine learning* (pp. 194–202).
40. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., Mclaughlin, M. E., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature, 415*(6870), 436–442. doi:10.1038/415436a.
41. Kendziorski, C., Irizarry, R. A., Chen, K. S., Haag, J. D., & Gould, M. N. (2005). On the utility of pooling biological samples in microarray experiments. *PNAS, 102*(12), 4252.
42. Shi, L., Jones, W. D., Jensen, R. V., Wolfinger, R. D., Kawasaki, E. S., Herman, D., et al. (2007). Reply to MAQC papers over the cracks. *Nature Biotechnology, 25*, 28–29.

43. MAQC consortium (2006). The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology, 24*, 1151–1161.
44. Tomayko, M. M., Anderson, S. M., Brayton, C. E., Sadanand, S., Steinel, N. C., Behrens, T. W., et al. (2008). Systematic comparison of gene expression between murine memory and naive B cells demonstrates that memory B cells have unique signaling capabilities. *Journal of Immunology, 181*(1), 27.

Salih Tuna is a PhD student in the Information: Signals, Images and Systems (ISIS) research group, School of Electronics and Computer Science, University of Southampton. He obtained his MSc in statistics/econometrics in 2005 and BSc in statistics in 2003 at Dokuz Eylul University/Turkey. His current research interests are machine learning applications for bioinformatics.

Mahesan Niranjan is Professor of Electronics and Computer Science at the University of Southampton, where he is head of the Information: Signals, Images and Systems (ISIS) research group. Prior to this appointment in February 2008, he has held a professorship in the University of Sheffield (1999–2008) and a lectureship in the University of Cambridge (1990–1998). At Sheffield he has served as Head of Computer Science (2002–2004) and Dean of the Faculty of Engineering (2006–2008). He received his BSc from the University of Peradeniya, Sri Lanka (1982), MEE from Eindhoven, The Netherlands (1985), both in Electronics Engineering, and his PhD from the University of Cambridge (1990). His research interests are in the algorithmic and applied aspects of Machine Learning, and he has authored or co-authored about 100 papers in peer reviewed journals and conferences. He has been Program Chair of several international workshops and has acted as a co-organizer of a six month program on Neural Networks and Machine Learning at the Isaac Newton Institute for Mathematical Sciences, Cambridge.