BIOINFORMATICS

Vol. 20 no. 17 2004, pages 2943–2953
doi:10.1093/bioinformatics/bth338

Bayesian integrated functional analysis of microarray data

Madhuchhanda Bhattacharjee[1,], Colin C. Pritchard[2], Peter S. Nelson[2] and Elja Arjas[1]*

[1]Rolf Nevanlinna Institute, Department of Mathematics and Statistics, University of Helsinki, P.O. Box 68, FIN 00014, Helsinki, Finland and [2]Division of Human Biology, Fred Hutchinson Cancer Research Centre, Seattle, WA 98109-1024, USA

Received on December 10, 2003; revised on April 30, 2004; accepted on May 16, 2004
Advance Access publication June 4, 2004

ABSTRACT

Motivation: The statistical analysis of microarray data usually proceeds in a sequential manner, with the output of the previous step always serving as the input of the next one. However, the methods currently used in such analyses do not properly account for the fact that the intermediate results may not always be correct, then leading to cumulating error in the inferences drawn based on such steps.

Results: Here we show that, by an application of hierarchical Bayesian methodology, this sequential procedure can be replaced by a single joint analysis, while systematically accounting for the uncertainties in this process. Moreover, we can also integrate relevant functional information available from databases into such an analysis, thereby increasing the reliability of the biological conclusions that are drawn. We illustrate these points by analysing real data and by showing that the genes can be divided into categories of interest, with the defining characteristic depending on the biological question that is considered. We contend that the proposed method has advantages at two levels. First, there are gains in the statistical and biological results from the analysis of this particular dataset. Second, it opens up new possibilities in analysing microarray data in general.

Contact: mab@rni.helsinki.fi

Supplementary information: http://www.rni.helsinki.fi/∼mab/

INTRODUCTION

In the literature, microarray data analysis is usually performed in a stepwise manner, starting with (1) normalization of the intensity measurements, to adjust or account for the effects of experimental conditions (dye, glass plate, replications, etc.), followed by (2) classification, which is performed on the standardized (or normalized) data to identify genes that are expressed differentially, and finally, (3) functional analysis of the identified genes. Even though all these steps are essential for a biologically meaningful analysis, they are generally viewed as distinct and independent. This separation of the different steps needed is reflected also in all the literature on the corresponding statistical methodology, where, for example, classification (clustering) is typically considered without regard to uncertainties in the results of a preceding normalization.

Some drawbacks of existing methods

Normalization and classification Measurement data from microarray experiments tend to be very noisy. The current practice has been to use a single transformation of the intensity measurements, which then also involves statistical estimation of the corresponding parameters. In our experience, however, real data are mostly not informative enough about what particular transformation, and consequently, choice of parameters, would be most appropriate. Therefore all normalized data involve a certain degree of uncertainty, which is due to both the choice of an appropriate transformation and the inaccuracy of the parameter estimates that are used in order to compute the normalized values.

This uncertainty in the normalized data is commonly not accounted for in the subsequent steps of the data analysis, such as classification. Unfortunately any efficient classification technique which makes fine distinction between the expression measurements of a gene, would naturally be susceptible to basic input data, whether it accounts for uncertainty explicitly, e.g. testing of hypothesis, or not, like algorithmic clustering techniques. In presence of uncertainty in the normalization step this would result in different list of identified genes using not so different normalization methods. Also approaches like testing of hypotheses are generally targeted for individual genes and not for collections of genes. The latter would be more appropriate considering that the final goal of such studies is to arrive at a system level understanding of the genes and their functions.

Functional analysis Most analyses of genetic data, when attempting to infer about functionality, combine the available

*To whom correspondence should be addressed.

functional information with classification results obtained by analysing experimental expression data. However, this is attempted only after the decisions concerning classification have already been made, and without paying explicit attention to the fact that the classification results are uncertain.

Annotation In attempts to combine information from different existing databases with experimental data typically gene annotations provided with the arrays are used as a key index. We have encountered situations in which, although cDNA arrays had been prepared using sequence-verified clones, re-sequencing showed that as many as 10% of the genes were still incorrectly annotated. Potentially, if such annotations are used as keys linking different sources of information, there is a very real possibility of being led to incorrect biological conclusions.

Bayesian integrated analysis

Bayesian techniques have been found useful for analysing microarray data by many authors, see e.g. Keller *et al*. (2000); Baldi and Long (2001); Dror *et al*. (2002) and Parmigiani *et al*. (2002). In fact, Bayesian methods have already been used to overcome some of the difficulties mentioned above, as in Ramoni and Sebastiani (2003), who proposed using the Bayesian approach to resolve the data transformation issues (e.g. to log or not to log) for oligo arrays.

An important reason behind the popularity of the Bayesian paradigm is that, once a model is built, inference follows automatically and the posterior distributions give direct answers, in terms of the conditional probabilities conditioned on the observed data, of statements of interest being true. In a traditional frequentist analysis for non-standard statistical models, finding optimal inference requires an enormous effort. Bayesian modelling saves this effort, which can then be directed towards building more realistic models.

We feel that the advantages of the Bayesian approach in analysing microarray data have not been explored sufficiently. This approach can help to integrate, as we demonstrate in this work, the steps of a sequential procedure into a single joint analysis, and then to quantify the remaining uncertainties in terms of joint distributions. In addition, one can take advantage of the fact that, if there are multiple sources of information which happen to point in the same direction, they will generally strengthen the substantive conclusions that can be made.

Here we follow, and extend, the approach of Bhattacharjee *et al*. (2002, 2003a), using a latent variable based Bayesian classification method where first the joint distribution of the parameters involved in the normalization step is obtained. The data are then transformed into a set of latent random variables using this joint distribution, and classification of genes is carried out for these random variables. Thus the classification is done by fully accounting for the uncertainties in the normalization step as measured under the normalization model. A comparison of results based on this approach with those obtained using ANOVA can be found in Bhattacharjee *et al*. (2002, http://www.camda.duke.edu/camda02/papers/days/papers/bhattacharjee/presentation).

Below we extend this approach further, by proposing a more general normalization technique which accounts for several more sources of errors in the measurement of data. We also illustrate how genes can be simultaneously classified according to several characteristics of interest, e.g. considering genes with a high expression (ratio), genes with a high between tissue variability, etc. Using this extended model, we can simultaneously assess the relative presence of different functionalities in a set of genes of interest.

With regard to annotation error, Bhattacharjee and Arjas (2003b) demonstrated, through a comprehensive model, that meaningful and robust biological conclusions can still be obtained when experimental expression data are augmented with existing databases in the presence of possible annotation errors and database quality uncertainty. However, this complicates the model further and should be used only in case serious annotation errors are suspected. The data used for illustration here uses re-sequence verified annotation and hence are expected to contain no or maximally very few annotation errors.

DATA

Expression data

The mouse has become an indispensable model organism for the study of development, genetics, behaviour and disease. Here, we consider a dataset based on cDNA microarray experiments carried out at the Fred Hutchinson Cancer Research Centre by Pritchard *et al*. (2001).

Normal tissue samples were collected from three organs of six male mice of C57BL6 strain. A pooled sample of mRNA from these 18 tissues was used as reference sample and samples from each of these 18 tissues were used as experimental samples, each hybridized on four arrays, with a dye swap. A detailed description of the arrays used for the experiment can be found from Pritchard *et al*. (2001). The experimental procedure and basic data extraction procedures are also described there.

By allowing maximally one flagged observation for each mouse and each organ, 3822 genes of the 5406 genes were selected for the analysis. The background corrected data produced several negative intensities, which in absence of further explanation were treated as missing. Log transformations were made on all (non-missing) reference and experimental sample intensities for all 3822 genes.

In the following (Fig. 1) the raw data from an array of these data are plotted. The parameter estimates and their SD are reported in Table 1, where we have applied a lowess type piecewise linear transformation for normalization.

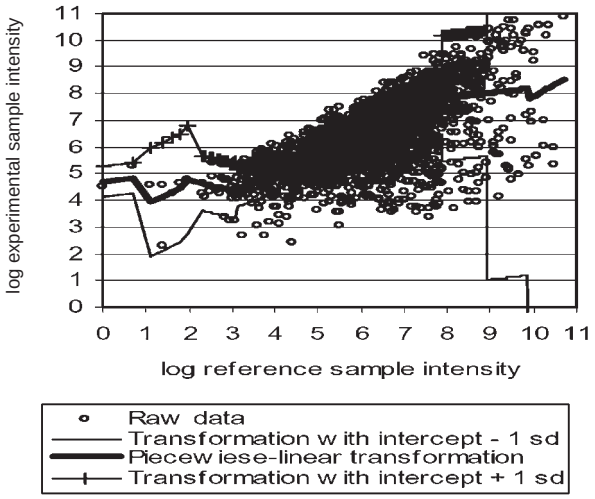

Fig. 1. Plots of raw data and piecewise linear regression transformation.

Table 1. Parameter estimates under piecewise linear regression normalization of data

Log-intensity-range		Estimates	
Lower limit	Upper limit	Intercept	Co-efficient
0.00	1.10	4.72 ± 0.57	0.15 ± 0.75
0.90	2.10	3.11 ± 2.03	0.75 ± 1.32
1.90	3.10	5.57 ± 1.03	-0.04 ± 0.37
2.90	4.10	2.49 ± 0.70	0.63 ± 0.19
3.90	5.10	2.60 ± 0.33	0.59 ± 0.07
4.90	6.10	1.78 ± 0.24	0.75 ± 0.04
5.90	7.10	1.53 ± 0.36	0.78 ± 0.05
6.90	8.10	4.05 ± 0.90	0.42 ± 0.12
7.90	9.10	6.38 ± 2.35	0.18 ± 0.28
8.90	10.10	6.28 ± 6.99	0.19 ± 0.74
9.90	11.10	-0.99 ± 25.01	0.89 ± 2.45

Table 1 clearly shows that the parameter estimates contain various degrees of uncertainties in them. This makes several distinctly different transformations of the data (under same normalization method) equally possible. Note however that the biological conclusions drawn based on the subsequent analysis of such normalized data then clearly depend on the parameters chosen for normalization.

A commonly made assumption is that the normalized log-ratios from an array should roughly resemble a normal distribution. However, for these data a close inspection of the design indicates that, irrespective of the actual level of expression of a gene in a tissue, the log ratios are bounded from above [by approximately log(18)] whereas not so from below (which can also be seen in Fig. 1). Also, depending on the proportions of tissue-specific genes, the right tail of the distribution of log-ratios from an array could be thinner than the left tail of the same. Therefore normality, if any, could only be assumed for the normalized log-ratios of a particular gene across the replicates, and not for the ratios from different genes on an array.

Table 2. Biological processes selected and number of genes having corresponding annotation

Biological process	No. of genes with the functionality	
	In 5401 genes on the array	In 3822 selected genes
Metabolism	1181	925
Cell growth and/or maintenance	599	474
Cell communication	390	257
Morphogenesis	149	95
Response to external stimulus	105	71
Response to stress	86	60
Death	53	43
Cell death	53	43
Cell differentiation	44	26
Cell motility	38	21
Response to endogenous stimulus	34	25
Reproduction	20	16
Pattern specification	18	14
Embryonic development	18	13
Regulation of gene expression, epigenetic	8	5

Functional data

Using the gene-level expression profiles and additional biological information available on the sampled genes, our aim is to infer possible distinguishing features between the selected genes from the point of view of their functionality. Resources like Gene Ontology (GO) (Ashburner *et al*., 2000) can provide useful information on the biological relevance of the observed expression-ratio profiles. Although the information available from the GO database has certain limitations, like incomplete information on a gene, or on the functions, in the absence of reasonable alternatives the information from GO can still be taken as indicative of the true biological processes involved under the given conditions.

The biological processes considered here were chosen from the level three annotations of the GO database, based on their relevance to the experimental conditions. Apart from a few relevant GO terms, care was taken to see that the number of genes present on the array with the considered functionality (as per GO) was not too few. The functionalities chosen are presented in Table 2.

BAYESIAN LATENT VARIABLE MODEL

In the following we specify the Bayesian model used for the analysis, in several steps. Note however that the implementation is in practice simultaneous. Wherever applicable, the hyper-parameters have been chosen to give rise to reasonably vague priors.

Integrated normalization and classification model

The need for a careful consideration of design aspects of microarray experiments has been emphasized before (e.g. Yang and Speed, 2002; Lee *et al*., 2000), and as a result the design frequently provides repeated measurements for the same sample, e.g. the same reference sample was used on all 72 arrays for the present data.

As mentioned before, the background corrected data contained missing observations. For the missing reference sample observations we use the following model based data augmentation, where the model parameters are to be estimated from the available observations. Starting from the log-intensities of reference (LIR) samples,

$\text{LIR}_{lij} \sim$ Normal $(\mu_i, 0.1)$, where

$\mu_i \sim$ Normal (0, 0.1), with

$l = 1$, 2 and 3 (organs),

$i = 1, \ldots, 3822$ (genes) and

$j = 1, \ldots, 24$ (arrays for an organ).

The two parameters of the above Normal distributions represent mean and precision (inverse variance), respectively. The same parameterization will be followed throughout the paper.

The log-transformed intensities from the experimental samples are modelled, by employing a piecewise linear model as follows:

$$\text{LIE}_{lij} \sim \text{Normal}[(\alpha_{ljb(i)w(lij)} + \beta_{ljb(i)w(lij)} \times (\text{LIR}_{lij} + \theta^1_{li})), \tau^1(C^1_{li})],$$

where

$l = 1$, 2 and 3 (organs),

$i = 1, \ldots, 3822$ (genes),

$j = 1, \ldots, 24$ (arrays for an organ),

$b(i)$ print tip/block number of i-th gene and

$w(lij)$ window number for LIR_{lij}.

For the missing experimental sample observations, data augmentation is done automatically using the above model and multiple array information. In the following, we describe the motivation behind the choice of the parameters involved in the above model.

For each array, block level piecewise linear regression normalization

The differences between the two channels were evident in these data, as it usually is for most cDNA experiments. In the present data, some degree of print tip variation was also observed (Pritchard *et al*., 2001). We propose to model the log-intensities from the experimental sample (LIE) by a piecewise linear regression of the LIR described above.

For this analysis, we assumed that the knots/break points were known and therefore fixed in advance at 5 and 7 by visual inspection. This was done because the large size of the experiments considered here would otherwise lead to a very heavy computational burden. We have successfully implemented models with unknown break points for smaller datasets.

The proposed normalization might appear simpler than some of the existing normalization methods. But note that, because of its Bayesian nature, the proposed method in reality explores through a large space of regression models, with suitable probabilities attached to each. This allows, not only the normalization uncertainty to be carried into the classification step, but also gives rise to a quite general normalization.

For example, in the above proposed model approximately 3500 regressions (72 arrays $\times$ 16 blocks/pins $\times$ 3 ranges of log-intensities) are being carried out as a part of normalization. The joint posterior distribution of all these parameters is used to transform the raw data. The transformed data, as a reflection of our uncertainty in the normalization step, are a collection of random variables and not deterministic numbers as produced by existing techniques.

The prior for the parameters involved in the normalization part of the model was specified by:

$\beta_{ljkm} \sim$ Normal (1, 0.1) and $\alpha_{ljk2} \sim$ Normal (0 , 0.1),

$\alpha_{ljk1} = \alpha_{ljk2} + (\beta_{ljk2} - \beta_{ljk1}) \times 5$,

$\alpha_{ljk3} = \alpha_{ljk2} + (\beta_{ljk2} - \beta_{ljk3}) \times 7$, where

$l = 1$, 2 and 3 (organs),

$j = 1, \ldots, 24$ (arrays for an organ),

$k = 1, \ldots, 16$ (blocks on each array) and

$m = 1$, 2 and 3 (number of windows).

Characterizing genes with respect to expression ratios

In our design, both the experimental and reference samples were normal tissues and moreover were made from the same tissue samples. Therefore, unlike in the commonly considered microarray experiments, the present dataset may not contain any 'differentially' expressed genes. However, the reference sample is not precisely the same as the experimental one, being rather a pool of all such samples, and some genes can have differences in expression across the experimental samples, e.g. across organs.

In that case, the expression levels of these genes could be higher or lower in the experimental sample extracted from one organ than in the reference sample. This would result in the corresponding log-ratio of intensities being away from the expected zero value. Then by necessity, the log-ratio of intensities for the same genes would be expected to behave in the opposite way in at least one of the remaining experimental samples.

Therefore, the normalized log-ratio of the intensity for a gene could deviate from zero, and we model this unknown

expression level in an organ through a latent variable. This latent variable can then be used to identify genes that have variable expression across the three organs.

First, assume that, a priori, every gene i has its own mean ratio of expression, say θ_i^0, where

$$\theta_i^0 \sim \text{Normal}(0, 0.1), \quad \text{with } i = 1, \ldots, 3822 \text{(genes)}.$$

However, as described above the actual expression within each organ could vary around this mean. Let the expression ratio for the l-th organ be θ_{li}^1, where θ_{li}^1, $l = 1$, 2 and 3, are assumed to be drawn from a Normal distribution with mean θ_i^0.

As mentioned earlier, for these particular datasets, a gene may not have dramatically different expression levels between the experimental and reference samples. Still it would be useful to identify genes with a relatively higher expression log-ratio in an organ compared with other genes. Accordingly, genes could be characterized according to whether or not they belong to, say, the top 10% in terms of expression log-ratio for that organ.

Characterizing genes with respect to between organ variability

Genes with widely different θ_{li}^1's across the organs indicate that there could be between organ variability of expressions for such genes. Modelling the variance of the distribution from which these θ_{li}^1's are drawn captures this aspect. We assume that this variation could be in one of three (a priori unknown) categories. Each gene i is characterized using a latent variable C_i^0 according to its variability amongst the θ_{li}^1's. Instead of variances, modelling is actually to be carried out in terms of the precision parameters ($\underline{\tau}^0$). The proportion of genes in the different precision classes in the population is considered as an unknown vector ($\underline{p}^0$). This is written as,

$C_i^0 \sim \text{Multinomial}(1, \underline{p}^0)$,

$\theta_{li}^1 \sim \text{Normal}[\theta_i^0, \tau^0(C_i^0)]$, where

$\underline{p}^0 \sim \text{Dirichlet}(\underline{1})$,

$\tau_k^0 \sim \text{Gamma}(1, 1)$,

$k = 1, 2, 3$ (the three variation/precision classes) with $\tau_1^0 < \tau_2^0 < \tau_3^0$,

l : 1, 2 and 3 (organs),

$i = 1, \ldots, 3822$ (genes).

Characterizing genes with respect to within organ variability

Even after accounting for between organ variability by including organ specific means, some genes might still exhibit variability within organs. This could be due to various reasons, e.g. due to heterogeneity between mice as reported by Pritchard *et al*. (2001). To assess and characterize genes with respect to such within organ (residual) variation, we introduce into the model the following gene- and organ-specific latent variables C_{li}^1, which indicate the level of such variability. As before, modelling is carried out in terms of the precision parameters ($\underline{\tau}^1$). For each organ, the proportion of genes in the different precision classes in the population is assumed to be unknown ($\underline{p}_l^1$). The following hierarchical model specifies a corresponding prior:

$C_{li}^1 \sim \text{Multinomial}(1, \underline{p}_l^1)$, where

$\underline{p}_l^1 \sim \text{Dirichlet}(\underline{1})$,

$\tau_k^1 \sim \text{Gamma}(1, 1)$,

$k = 1, 2, 3$ (the three variation/precision classes) with $\tau_1^1 < \tau_2^1 < \tau_3^1$,

l : 1, 2 and 3 (organs),

$i = 1, \ldots, 3822$ (genes).

When modelling both between and within organ variation one could introduce more classes of variation, or even a continuous spectrum of such classes. However, the effective sample size for such parameter estimation could be very small. For example, if the between organ variation is modelled through gene-specific precision parameters τ_i^0 instead of the proposed $\tau^0(C_i^0)$, then there are only three θ_{li}^1's for estimating the precision τ_i^0 for each gene i, and also these θ_{li}^1's are unknown. It is quite likely that such a model would not yield high-quality estimates. In spite of this fact such practices are being followed in well-known classification techniques, e.g. ANOVA, where for every gene a large number of parameters are being estimated based on only few observations. This has been one of the main reasons why we chose a predetermined and much smaller number (here 3) of possible variation classes. The values for the corresponding variation/precision parameters are estimated from the data.

Probabilistic assessment of biological processes enrichment

For the genes selected for the analysis, additional information on 15 biological processes was obtained from the GO database. Our goal is now to try to combine the functional information in the GO database with the results obtained from the statistical analysis of expression data and thereby try to assess whether some particular functionality seemed to be 'enriched' within gene classes of interest that were suggested by the data analysis. The classes of interest could consist of genes with a higher expression ratio, high between organ variability, high within organ variability, etc.

Let G_j denote the set of genes present on the array having the j-th functionality, $j = 1, \ldots, 15$. Let S denote a set of selected genes and S^c its complement. Let GS_j denote the subset of genes from S that also have the functionality j, i.e. $GS_j = S \cap G_j$, and let $GS_j^c = S^c \cap G_j$. Consider the

random variable

$$T_{Sj} = I\{\|GS_j\|/\|S\| > \|GS_j^c\|/\|S^c\|\},$$

where $I\{\text{event}\}$ is 1 if event is true, zero otherwise.

The behaviour of T_{Sj} could be taken as an indicator of enrichment of functionality j in set S. This is so, since, if functionality j was independent of the characteristics of interest that led to the construction of S, then the genes having this functionality would have been distributed randomly between S and S^c, only depending on their cardinalities. Departures from this independence could be judged by, say, checking whether the proportion of genes with functionality j is consistently higher in S than the proportion of such genes in S^c. In that case the posterior probability of the event $T_{Sj} = 1$ would be high. This posterior probability will later be called 'estimated probability of enrichment'.

Note that the current practice of determining presence or absence of any functionality is to treat S as a completely deterministic set containing genes selected by some separate decision-making mechanism such as ANOVA, where decisions are taken gene by gene. This set S is created without regard to either the uncertainty involved in the classification step or any possible dependence between the genes.

In the method proposed here the membership of a gene in the set S (and S^c) is a random variable, which in the current analysis would be a function of (unknown) variables such as $\theta_{li}^1, C_i^0, C_{li}^1$, etc. Therefore, we not only account for the uncertainty in the actual decision concerning possible membership of S but also, as the joint posterior of these latent variables is used to obtain the distribution of T_{Sj}, it automatically accounts for possible dependences amongst the genes.

The above statistic was formed, as we are primarily interested in relative enrichment of functionality in a selected set of genes. Modifications could be made to make the test more stringent, for example by adding a minimum cut-off (ε) on the proportion difference, $(\|GS_j\|/\|S\|) > (\|GS_j^c\|/\|S^c\| + \varepsilon)$. Other tests may be formed if more sensitive hypotheses are to be considered.

Implementation

We have formulated the model in the Bugs language and performed parameter estimation using WinBUGS (Spiegelhalter *et al*., 1999).

CHARACTERIZATION OF MOUSE GENES

The dataset was analysed using our model to identify genes with possibly different expression levels in normal tissues from the three organs, namely kidney, liver and testis. The estimated variation within the organs was far less than that between, indicating clearly the inherent expression difference in these organs.

The highest across organ precision (i.e. smallest variance) class had posterior mean 22.4, much less than the corresponding highest mean precision class 63.1 that was observed within the organs. The posterior mean of the precision parameter, for the class of genes highly varying within an organ, was 4.6, which is higher than both the lowest (0.4) and moderate (2.5) precision classes when variation across organs is examined. The precisions and proportions of genes in each class are presented in Table 3.

Table 3. Posterior population parameter estimates

Class	Posterior mean estimates for					
	Precision		Proportion of genes			
	Between organs (τ^0)	Within organs (τ^1)	Between organs (p^0)	Within organs (p^1) Kidney	Liver	Testis
1	0.4	4.6	0.19	0.11	0.17	0.06
2	2.5	20.5	0.45	0.34	0.54	0.37
3	22.4	63.1	0.36	0.55	0.29	0.57

In an additional analysis, we actually characterized the genes for variability within an organ when the precision/variation parameters from between organ variability are used to characterize such variation classes. This additional analysis estimated no gene to be highly varying within an organ as per this new definition of variability.

A closer inspection of the genes with variable expression across organs revealed that some of them are known to be relevant for these organs, either for mouse or for some other species, e.g. rat or human. Although at population level 19% genes were estimated as varying between organs, about ∼7% (276) of these genes were estimated with >90% probability to be so. For many of these 276 genes at present there is not enough knowledge about their organ specificity with respect to these three organs, and it would be useful to study these genes further. For some of the genes that were found variable across organs, in Figure 2a we present the estimated log-expression ratio from the organs, with brief comments (Table 4).

The estimated gene wise probabilities of having higher expression ratio in the organ showed kidney and liver had several genes in common among the genes with higher expression ratio. However those with higher expression ratio in testis did not have higher expression ratios in either kidney or liver. It could be that the presence of sperm, which are haploid cells, may cause the testis to be distinct from other organs in the body.

Estimating functional enrichment probabilities

Tests of different stringencies were carried out to assess the enrichment of functionalities amongst genes with variable expression across organs. Figure 2a clearly indicates that many genes have different expression levels in the organs. In the following, we investigate whether such behaviour could be related to the functionalities of these organs.

Table 4. Available descriptions and/or comments on some genes estimated as highly variable across organs

Gene	Comment
MGC37245	Hypothetical protein, but kidney specific protein AAD05209 in *Rattus norvegicus*
Aldrl6	Is developmentally regulated in kidney, exhibits a distinct spatiotemporal distribution, and probably plays a role in tubulogenesis in the kidney
Slc27a2	Highest levels of Vlcs activity in mouse liver and kidney tissues
Umod	The most abundant protein in normal urine
Hrsp12	Was isolated as a novel heat-responsive, tissue-specific, phosphorylated protein isolated from mouse liver
Pah	Kidney related
Pzp	Also called alpha 2 macroglobulin, made by the liver
F2	Coagulation factor F2 gene and coagulation factors are made by the liver
Spi1-3	A protease inhibitor that is part of the acute phase response and acute phase response is mediated primarily by the liver
Fga	Is the main protein in blood clots, it is made by the liver
Apoa5	Apolipoproteins are primarily made by the liver
Rbp4	Known to be expressed in liver
Gcc1	Mouse neural stem cell (differentiated) cDNA library (long) Golgi coiled-coil 1
Tle1	Tissue specificity of the gene is known to be liver for mouse
Hpxn	Haptoglobin and haemopexin together are essential for protection from splenomegaly and liver fibrosis resulting from intravascular haemolysis. Similar gene in human 'hpx', distinctly expressed in liver
Tnp1	Involved in spermatogenesis in the testis
Tcte3	T-complex-associated testis expressed 3

Assessing functional enrichment amongst genes variable between organs The functionalities that varied most among the liver, kidney and testis were reproduction, response to external stimulus, response to stress, metabolism, cell differentiation, and cell growth and maintenance. Enrichment probabilities of the functionalities were estimated for different sets of genes and a partial summary is presented in Table 5.

A summary of characterizations of genes according to between organ variability is presented in Table 6. Note that the number of genes estimated with high probability to be variable could be much less than the population estimated number/proportion of such genes.

For some functionality clear organ specific patterns in the expression can be seen. For example, 'Reproduction' was estimated to be enriched in highly varying genes among genes that were selected according to criterion S-1 (Table 5). Observe from Figure 3 that the genes that are variable across organs and also have this functionality were expressed, as could be expected, in testis only, and not in kidney or liver.

Unlike 'Reproduction' some other enriched functionality might produce multiple expression profiles. For example, the functionality 'Response to external stimulus' was found to be enriched and produced two distinct expression profiles

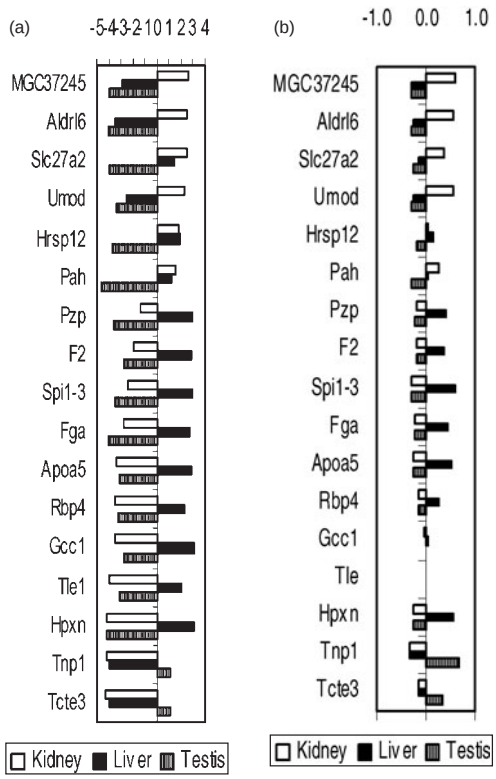

Fig. 2. (**a**) Estimated log-ratio of some between organ variable genes. (**b**) Adjusted proportions of ESTs in these three organs as per Unigene.

among variable genes (Fig. 4). In one expression profile genes expressed in testis only and in the other genes expressed in only liver. However, a further investigation on these variable genes revealed that these two profiles have been created by two different subfunctions that were organ-specific. The genes expressed in testis are for 'Response to a-biotic-stimulus', whereas the ones expressed in liver are for 'Response to biotic-stimulus'. The majority of 'Response to stress'-related genes that varied between organs were also annotated by GO as 'Response to external stimulus'. This explains why response to stress was estimated to be enriched.

'Metabolism'-related genes are expected to vary between organs because each organ performs specific metabolic functions. Liver is responsible for eliminating toxins, producing blood coagulation factors and orchestrating the acute-phase immune response. Neither kidney nor testis performs these metabolic functions. But precise explanation of all the observed expression profiles may be hard due to the diverse and complex nature of this functionality. Some of the 78 metabolism-related genes which were highly variable across organs have been presented in Figure 5 and these are annotated to have at least a dozen different subfunctions.

'Cell differentiation'-related genes are expected to vary between organs because they are related to cell-type specificity.

Table 5. Posterior probability of enrichment for different biological processes among genes estimated to be highly variable across organs

Biological process	Posterior probability of enrichment amongst genes variable between organs		
	Genes with any precision within organs (S-1)	Genes with moderate or high precision in at least two organs (S-2)	Genes with moderate or high precision within each of the three organs (S-3)
Metabolism	1.0	1.0	1.0
Cell growth and/or maintenance	1.0	0.9	1.0
Cell communication	0.0	0.1	0.3
Morphogenesis	0.8	0.7	0.6
Response to external stimulus	1.0	0.9	0.9
Response to stress	1.0	1.0	0.9
Death	0.0	0.3	0.5
Cell death	0.0	0.3	0.5
Cell differentiation	1.0	0.5	0.4
Cell motility	0.5	0.4	0.6
Response to endogenous stimulus	0.4	0.7	0.8
Reproduction	1.0	0.6	0.2
Pattern specification	0.1	0.2	0.2
Embryonic development	0.4	0.4	0.4
Regulation of gene expression, epigenetic	0.2	0.3	0.2

'Cell growth and maintenance', like 'Metabolism', is a large class with several sub-functionalities. The genes identified as variable under this functionality were also annotated to several different sub-functionalities, e.g. cell organization and biogenesis, cell proliferation, transport, etc.

Assessing functional enrichment amongst genes variable within an organ As mentioned before that the level of within organ variation is very much smaller than that between (refer Table 3). The results of the functional analysis of within variability should be viewed in this light.

Table 7 contains a partial summary of our findings on functional enrichment amongst genes having varied expression within an organ.

According to our analysis, five biological processes were enriched in all three tissues among the variable genes within organs: 'Morphogenesis', 'Response to external stimulus', 'Response to stress', 'Cell differentiation' and 'Pattern specification'. Genes that respond to stress or to external stimulus might be expected to vary between individuals due to different environmental milieus. However, the reasons behind the variability for the other three functionalities are less clear as these are functionalities that would define specific cell and tissue types. One explanation is that subtle environmental differences during organ development result in fixing some genes related to differentiation, morphogenesis and pattern specification at different levels in different individuals.

Table 6. Characterization of genes according to variability across organs (S-1)

Type of genes	Number of genes analysed	Number (proportion) estimated as varying	Number (proportion) with estimated probability of varying ≥90%
All	3822	733 (0.19)	276 (0.07)
Metabolism	925	203 (0.22)	78 (0.08)
Cell growth and/or maintenance	474	108 (0.23)	46 (0.10)
Cell communication	257	41 (0.16)	7 (0.03)
Morphogenesis	95	20 (0.21)	6 (0.06)
Response to external stimulus	71	21 (0.30)	7 (0.10)
Response to stress	60	16 (0.27)	8 (0.13)
Death	43	5 (0.12)	1 (0.02)
Cell death	43	5 (0.12)	1 (0.02)
Cell differentiation	26	8 (0.31)	5 (0.19)
Cell motility	21	4 (0.19)	0 (0.00)
Response to endogenous stimulus	25	4 (0.16)	0 (0.00)
Reproduction	16	9 (0.56)	6 (0.38)
Pattern specification	14	1 (0.07)	0 (0.00)
Embryonic development	13	2 (0.15)	1 (0.08)
Regulation of gene expression, epigenetic	5	0 (0.00)	0 (0.00)

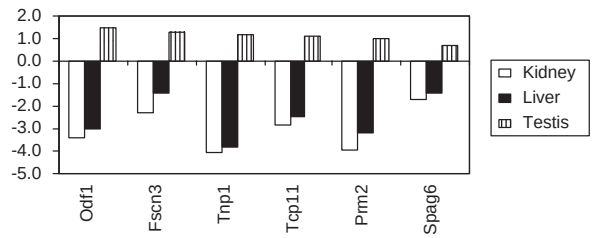

Fig. 3. Estimated log-expression ratio of reproduction related genes that are variable across organs.

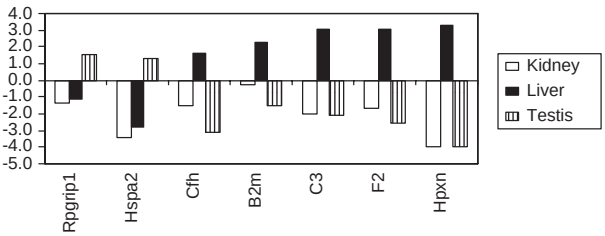

Fig. 4. Estimated log-expression ratio of response to external stimulus related genes that are variable across organs.

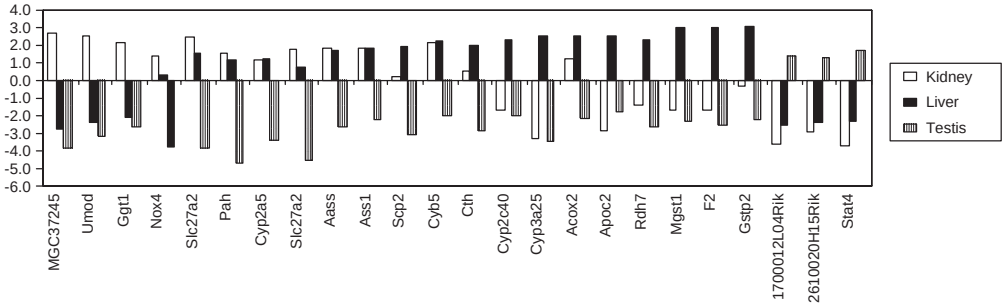

Fig. 5. Estimated log-expression ratio of some metabolism related genes that are variable across organs.

Table 7. Posterior probability of enrichment for different biological processes among genes estimated to be highly variable within an organ

Biological process	Functionalities enriched amongst genes variable within an organ		
	Kidney	Liver	Testis
Metabolism	0.0	0.0	1.0
Cell growth and/or maintenance	0.4	1.0	1.0
Cell communication	1.0	1.0	0.1
Morphogenesis	1.0	1.0	1.0
Response to external stimulus	1.0	1.0	1.0
Response to stress	1.0	0.8	1.0
Death	0.0	0.0	0.0
Cell death	0.0	0.0	0.0
Cell differentiation	1.0	1.0	1.0
Cell motility	0.7	1.0	0.3
Response to endogenous stimulus	0.0	0.0	0.0
Reproduction	1.0	1.0	0.4
Pattern specification	1.0	0.8	1.0
Embryonic development	0.0	0.8	1.0
Regulation of gene expression, epigenetic	0.0	1.0	0.0

Assessing functional enrichment amongst genes with high expression within an organ For this analysis, genes were characterized according to whether or not they belong to the top 10% in terms of expression ratio in an organ. Note however that, although here we are doing a marginal analysis of the expression ratio of a gene from an organ, we know because of the design of this particular experiment that this ratio is related to the ratios from the remaining organs too. Therefore, if a gene is characterized as having a high expression ratio in one organ, it is expected to have smaller ratio in at least one of the two other organs. This implies that such genes could show between organ variability, too, although they may not be among the most variable ones.

The enriched functionalities amongst the genes with a high expression-ratio in an organ showed both reflection of nature of the functionalities of the organ and between organ variability (Fig. 6).

'Metabolism' is enriched as liver and kidney are both organs involved heavily in metabolic activity (especially liver). These two organs are highly responsive to external stimulus also. For example, liver will up-regulate a subset of p450 metabolic enzymes in response to specific environmental toxins, and kidney is constantly adjusting its activity in response to hormones, such as cortisol and rennin. 'Reproduction' is the primary functionality of the testis. The genes related to 'Cell Differentiation' will tend to be cell type (and consequently organ) specific since they are what define a cell type. Therefore it is not surprising that each organ has this functionality enriched, since each will have a subset of cell differentiation genes with higher expression than the other two organs.

ANALYSES USING EXISTING TECHNIQUES

In addition to Pritchard *et al*. (2001), the data under consideration have been analysed previously by many. Several of these works can be found in the proceedings of the CAMDA02 conference (Johnson and Lin, 2003). In the following, we present a brief comparative summary of the existing results based on this dataset.

An interesting comparison of different normalization techniques was carried out by Warren and Liu (CAMDA02). Their findings support our observations that non-ignorable noise in the data makes it difficult to adequately normalize data by using a single transformation, as the noise in the data makes several possible normalizations equally likely and each normalization produces a different list of significant genes.

Deng *et al*. (CAMDA02) discussed the choice of a model for identifying variable genes. Bhattacharjee *et al*. (2002) demonstrated that genes identified by commonly used techniques like ANOVA might exhibit counter intuitive expression profiles, as can be seen even by a simple visual inspection. This implies that biological hypotheses such as 'variability', when translated into a statistical hypothesis testing problem, need careful consideration and appropriate modelling.

Moloshok *et al*. (CAMDA02) and Diaz-Uriarte *et al*. (CAMDA02) analysed the data in conjunction with functional information as obtained from the GO database. The former paper discusses functional enrichment among genes with a higher expression ratio within an organ. However, the relationship of such genes to genes that were variable across different

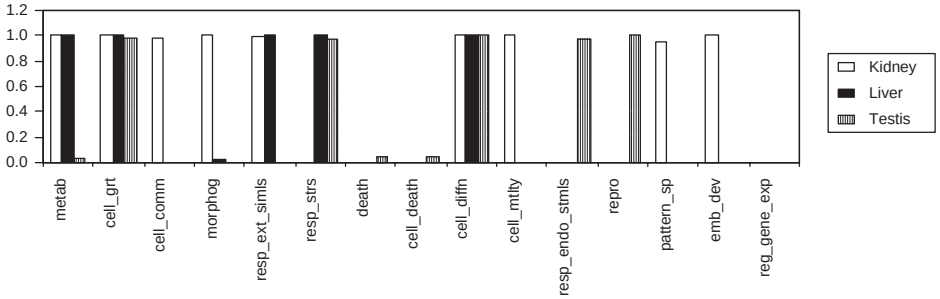

Fig. 6. Estimated enrichment probability amongst genes with higher expression ratio within an organ.

organs was not mentioned therein. The latter paper by Diaz-Uriarte *et al*. (CAMDA02) is closer to the analysis carried out here, although carried out in three distinct steps without accounting for the uncertainties in the decision-making. However, they found only three functionalities differing between organs, namely 'Metabolism', 'Reproduction' and 'Transport' (currently classified as level 4, under level 3 term 'Cell growth and/or maintenance'). Our method not only finds these three functionalities with high probabilities but additionally we identified more subtle but relevant GO terms like 'Response to external stimulus' and 'Response to stress'. Recall that in the original paper by Pritchard *et al*., these functionalities were only hypothesized to be enriched but the traditional techniques failed to identify these.

In addition to these analyses, we carried out a similar stepwise functional analysis of the data using existing popular softwares. The data were first normalized using print-tip specific lowess type normalization. This was followed by an analysis using SAM (Tusher *et al*., 2001) software. For simplicity, we carried out only an analysis to identify genes that would be variable across organs. For this analysis, the *d*-scores produced by SAM were quite comparable to the variability probability estimated under criterion S-1 in Table 5. Next, using EASE (Hosack *et al*., 2003) software, we carried out an analysis of functional enrichment for genes that were variable across organs. However, the software identified only two functionalities from level 3, namely 'Response to external stimulus' and 'Response to stress'. Although this analysis was quite similar to the one carried out by Diaz-Uriarte *et al*., CAMDA02, they produced very different results for the same data.

DISCUSSION

The proposed method has advantages at two levels. First, there are gains in the statistical and biological results from the analysis of this particular dataset. Second, it opens up new possibilities in analysing microarray data in general.

Analysis of Project Normal data

The refined normalization applied here explains the variability in the data better than that used by Bhattacharjee *et al*. (2002, 2003a). The best precision achieved by the earlier method within each organ was 14.2 which is less than even the moderate precision achieved in the present analysis (Table 3).

Identification of genes with an interesting expression profile appears to have been quite successful and in agreement with previously existing knowledge in spite of the fact that all model assumptions were made at the (population) level of all genes on the array. This establishes the inherent strength and accuracy of this model. As has been presented for genes variable across organs (in Fig. 2a and Table 4), similar details for genes with a high variability or high expression ratio within an organ can also be obtained. However, a detailed discussion covering all these is beyond the scope of this paper.

Functional data have been used by many to exploit the fundamental assumption that the genes with similar expression behaviour (i.e. co-expressed genes) are also co-regulated, directly or indirectly. Our finding for this particular dataset also suggests that genes that were selected on the basis of expression-ratio information also have distinct functional patterns across tissue types.

We noticed that different functionalities created very similar expression profiles. On the other hand, a broad functionality class might produce distinctly different expression profiles. Irrespective of this many-to-many relationship, in several cases, a one-to-one correspondence between the observed expression profile of the selected genes and functionalities could be established.

Analysis of microarray data in general

An adequate statistical analysis of microarray data often necessitates the use of complex models and involved inferential procedures. This is because such experiments are frequently targeted to answer much more complex questions than just identifying differential genes. For a particular disease or developmental problem, it is not uncommon for the (known) relevant genes to exhibit varied expression patterns (differential expression being merely one such). For example, dysfunction of a gene in the disease state could be indicated by higher variation across homogeneous samples, whereas in some other situations genes of interest could be the ones with

unaltered expression under both treatment and control state. In a real complex problem, one or more of these features could be of interest.

The method of analysis proposed here has been targeted to answer such more complex questions based on microarray experiments than just to identify differential genes, although it can be used quite efficiently to answer such questions too. Note that the proposed model is capable of assessing approximately 650 ($=2^3 \times 3^3 \times 3$) different profiles/patterns/behaviours of genes (based on expression ratios, and their within and between organ variations). Unfortunately, the dataset considered here is not suitable for illustrating when and why one would need such complex profiles.

The commonly available softwares are typically built to assess a single profile at a time; therefore, it would be a difficult task to obtain answers to the large collection of questions an integrated model can answer in a single step, from such softwares. Moreover, the underlying statistical model used in the software may not enable one to carry out assessment of some of the more subtle profiles of interest.

The statistical methods for analysing microarray data are applied, without exception, in a sequential manner, with the output of each step in the analysis serving as the input for the next. However, numerous examples can be given for obtaining conflicting results from the same data when a complete analysis is carried out in sequential manner, by a simple change of model/method in any of the steps.

Two major reasons of obtaining such contradictory results are: (1) if any of the steps involve uncertainty in its conclusions then the subsequent steps can be sensitive to the choice of method/software used for that step and (2) the technique/software used for an individual step may be based on sound statistical assumptions and appropriate for decision-making involved in that step, but the overall analysis could very well be a mixture of statistical assumptions with no guarantee of internal consistency.

For such reasons, it is generally acknowledged that integration of such stepwise procedures would be a desirable goal. The idea of an integrated model, for microarray data, was left as an open challenge in Sebastiani *et al*. (2003). Here we illustrate a possible way to meet this challenge, by presenting an integrated model in which normalization uncertainty is fully accounted for in classifying genes. Also the biological conclusions are drawn considering the uncertainty in the classification (and automatically normalization). We found the Bayesian framework to be highly convenient for implementing such an integrated model; however, we also feel that any efforts made towards such integration should be encouraged irrespective of the statistical paradigm.

ACKNOWLEDGEMENT

Authors sincerely thank Andrew Thomas for careful reading of the manuscript and useful comments.

REFERENCES

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Baldi,P. and Long,A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Bhattacharjee,M., Sillanpää,M.J. and Arjas,E. (2002) *Bayesian Characterization of Natural Variation in Gene Expression*. Critical Assessment of Microarray Data Analysis, Durham, NC.

Bhattacharjee,M., Pritchard,C.C., Sillanpää,M.J. and Arjas,E. (2003a) Bayesian characterization of natural variation in gene expression. In Johnson, K. and Lin,S. (eds), *Methods of Microarray Data Analysis III*. Kluwer Academic Publishers.

Bhattacharjee,M. and Arjas,E. (2003b) One step analysis of microarray data. *Theme Conference of the Royal Statistical Society, Statistical Genetics and Bioinformatics*, 14–17 July, Diepenbeck, Belgium.

Dror,R.O., Murnick,J.G. and Rinaldi,N.A. (2002) A Bayesian approach to transcript estimation from gene array data: the BEAM technique. In *RECOMB 2002: Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology*. ACM PRESS.

Hosack,D.A., Dennis,G., Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, 4.

Johnson,K. and Lin,S. (ed.) (2003) Methods of microarray data analysis III. *Proceedings of the CAMDA02 Conference*. Kluwer Academic Publishers.

Keller,A.D., Schummer,M., Hood,L. and Ruzzo,W.L. (2000) Bayesian classification of DNA array expression data. *Technical Report, UW-CSE-2000-08-01*, Department of Computer Science and Engineering, University of Washington, Seattle, WA.

Lee,M.T., Kuo,F.C., Whitmore,G.A. and Sklar,J. (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridisations. *Proc. Natl Acad. Sci., USA*, **18**, 9834–9839.

Parmigiani,G., Garrett,E.S., Anbazhagan,R. and Gabrielson,E. (2002) A statistical framework for expression-based molecular classification in cancer. *J. R. Stat. Soc. B*, **64**, 717–736.

Pritchard,C.C., Hsu,L., Delrow,J. and Nelson,P.S. (2001) Project normal: defining normal variance in mouse gene expression. *Proc. Natl Acad. Sci., USA*, **98**, 13266–13271.

Ramoni,M.F. and Sebastiani,P. (2003) Bayesian methods for microarray data analysis. *IMA Workshop 1: Statistical Methods for Gene Expression: Microarrays and Proteomics*, Minneapolis, USA.

Sebastiani,P., Gussoni,E., Kohane,I.S. and Ramoni,M.F. (2003) Statistical challenges in functional genomics. *Stat. Sci.*, **18**, 33–70.

Spiegelhalter,D.J., Thomas,A. and Best,N.G. (1999) WinBUGS, Version 1.2. User Manual MRC Biostatistics Unit.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

Yang,Y.H. and Speed,T.P. (2002) Design issues for cDNA microarray experiments. *Nat. Rev. Genet.*, **3**, 579–588.