Gene expression

Bayesian modelling of shared gene function

P. Sykacek1,*, R. Clarkson2, C. Print3, R. Furlong4 and G. Micklem5,6

1Department of Biotechnology, BOKU University, Vienna, Austria, 2School of Biosciences, Cardiff University, UK, 3Department of Molecular Medicine & Pathology, University of Auckland, New Zealand 4Department of Pathology, 5Department of Genetics and Cambridge Computational Biology Institute and 6Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK

Received on September 21, 2006; revised on May 8, 2007; accepted on May 18, 2007
Advance Access publication May 31, 2007
Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Biological assays are often carried out on tissues that contain many cell lineages and active pathways. Microarray data produced using such material therefore reflect superpositions of biological processes. Analysing such data for shared gene function by means of well-matched assays may help to provide a better focus on specific cell types and processes. The identification of genes that behave similarly in different biological systems also has the potential to reveal new insights into preserved biological mechanisms.

Results: In this article, we propose a hierarchical Bayesian model allowing integrated analysis of several microarray data sets for shared gene function. Each gene is associated with an indicator variable that selects whether binary class labels are predicted from expression values or by a classifier which is common to all genes. Each indicator selects the component models for all involved data sets simultaneously. A quantitative measure of shared gene function is obtained by inferring a probability measure over these indicators.

Through experiments on synthetic data, we illustrate potential advantages of this Bayesian approach over a standard method. A shared analysis of matched microarray experiments covering (a) a cycle of mouse mammary gland development and (b) the process of in vitro endothelial cell apoptosis is proposed as a biological gold standard. Several useful sanity checks are introduced during data analysis, and we confirm the prior biological belief that shared apoptosis events occur in both systems. We conclude that a Bayesian analysis for shared gene function has the potential to reveal new biological insights, unobtainable by other means.

Availability: An online supplement and MatLab code are available at http://www.sykacek.net/research.html#mcabf
Contact: peter@sykacek.net
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Robust methods for microarray data analysis are important for advancing biological research, e.g. allowing more focused drug design and better assessment of pathogenicity (Bild et al., 2006; Dave et al., 2006). Many approaches are available for the analysis of single experiments, including methods based on statistical testing (Pan, 2002; Reiner et al., 2003; Tusher et al., 2001; Wernisch et al., 2003), Bayesian automatic relevance determination (ARD) (Li et al., 2002), Bayesian variable selection (Lee et al., 2003; Bae and Mallick, 2004) and model selection (Li and Yang, 2002). While experiments on relatively homogeneous cell lines grown in tissue culture can give oversimplified results, analysis of individual experiments using multicellular tissues provides a blurred view of the molecular mechanisms operating in the tissue. This is caused by the superimposition of biological processes in, and between, multiple cell types. We propose in this article a Bayesian analytical method to identify common gene activity from multiple microarray experiments and refer to this approach as an analysis for shared gene function. Analysing different biological systems for shared genes is of interest in its own right (Hockley et al., 2006). In addition, the same kind of analysis can be applied to several well matched biological experiments that have a subset of cell types and active biological processes in common. Such analysis has the potential to provide more focused gene lists than an analysis of individual experiments. The success of such attempts, however, also depends to a great extent on a careful choice of all involved assays. Previous approaches for shared analyses have been through meta-analysis of gene lists (DeConde et al., 2006; Yang et al., 2005) and, based on expression values, through multivariate regression (Gilks et al., 2005) and by modelling Bayesian networks of pairwise correlations (Huttenhower et al., 2006).

Our approach to inferring shared gene function is a fully Bayesian assessment of whether we can establish, across experiments, a relationship between binary biological classifications (e.g. mutant versus wild type) and gene expression measurements. Using gene expression values as regressors, we model individual predictors by probit link regression (Lee et al., 2003; Spang et al., 2002). Like (Lee et al., 2003; Li et al., 2002), we use a Bayesian generalized linear model (GLM). However, we consider additional aspects: to provide an inference about shared molecular biological processes, our analysis combines information from heterogeneous sources, such as whole tissues, with well-matched cultured cells, the identification of which requires a thorough research of all involved biological systems. Such an approach also has the potential to allow data from
a new assay to be combined in a principled way with pre-existing data. This increases both the statistical power and cost-effectiveness of experiments. To allow calculation of probability measures reliably, we compare individual genes against a reference model. We suggest use of a reference which predicts biological classifications according to prior probabilities which are estimated from the class frequencies in the training data. This reference model does not use information from microarrays and must clearly be outperformed by functionally important genes. The sensitivity of the results to subjectively chosen hyperparameters is minimized by following (Bae and Mallick, 2004) and using hierarchical priors.

Our first experiment is a simulation using synthetic data which compares the Bayesian approach with a simple meta-analysis. We then apply the model to the shared analysis of two time-course experiments: (1) a cycle of growth and regression in mammary glands in vivo (Clarkson et al., 2004) together with (2) an assay of programmed endothelial cell death investigated in vitro (Johnson et al., 2004). Apoptosis of endothelial cells is known to occur during the mammary gland cycle and may play an important role in this process (Djonov et al., 2001; Matsumoto et al., 1992). Computer simulations support this prior biological belief, provide high predictive accuracy and confirm the Bayesian analysis proposed here.

2 METHODS

The following discussion assumes that biological samples are available with both microarray gene expression measurements and a known binary classification. In such situations, the importance of individual genes can be assessed by Bayesian model criticism (Bernardo and Smith, 1994). Similar approaches have previously been used by Lee et al. (2003) and Li et al. (2002), who applied Bayesian variable selection to obtain a measure of gene importance. Assuming an overall number of T genes, classical Bayesian variable selection attempts to infer a probability measure over a 2^T dimensional space. To maintain feasibility, we follow previous strategies (Pan, 2002) and consider single gene models.

2.1 Quantifying shared gene function

A probabilistic quantification of shared gene function over several microarray experiments is obtained by generalizing Bayesian model assessment for individual data sets. We assess gene function by quantifying the importance of a gene to a classifier which predicts a particular biological classification (e.g. mutant versus wild type) from its expression values. We suggest in particular comparing two generalized linear regression (McCullagh and Nelder, 1989) models (GLMs) for every gene. One GLM predicts class labels from gene expression measurements and an intercept term. The other GLM predicts class labels only from an intercept. The latter GLM predicts class prior measurements and an intercept term. The other GLM predicts class labels from gene expression vectors, with \( \xi_{\text{t}} \) denoting suitably transformed and normalized gene expression values. The prediction of \( y_{\text{t}} \) can thus be regarded as a two component mixture of GLMs. In a Bayesian context, (Lee et al., 2003; Spang et al., 2002) found that it is convenient to model binary classifications by probit link regression. The probit link GLM predicts

\[
P(Y_{\text{t}} = 0|x_{\text{t}}, \beta, I_t) = \Phi(x_{\text{t}}\beta)
\]

where \( \Phi \) is the Gaussian cumulative distribution function (cdf) with mean \( x_{\text{t}}\beta \) and unit SD at zero. If we denote all classifications by \( D_t = [y_1, \ldots, y_n] \) and all regressors by \( X_t = [x_{1t}, \ldots, x_{nt}] \), the probabilities

\[
P(Y_{\text{t}}|x_{\text{t}}, \beta, I_t) = (1-I_t)\Phi(x_{\text{t}}\beta) + I_t\Phi(-x_{\text{t}}\beta)
\]

give rise to the likelihood

\[
p(D_t|x_{\text{t}}, \beta, I_t) = \prod_{n} P(Y_{\text{t}}|x_{\text{t}}, \beta, I_t).
\]

To obtain the joint distribution, \( p(D_t, \beta, I_t) \), Bayesian inference requires multiplying the likelihood with a prior over regression coefficients, \( p(\beta|I_t) \). The functional importance of genes is quantified by the posterior probability \( P(I_t = 1|D_t, X_t) \), (Bernardo and Smith, 1994). Normalizing the product of prior probability \( P(I_t) \) and marginal likelihood \( P(D_t|X_t, I_t) = \int p(D_t|\beta, X_t, I_t)d\beta \), we obtain the posterior

\[
P(I_t|D_t, X_t) = \frac{P(I_t)\prod_{n} P(D_t|I_t, X_t)}{\sum_{I_t} P(I_t)\prod_{n} P(D_t|I_t, X_t)}
\]

This principle is extended to inferring genes that show a shared functional importance in several microarray experiments. After standardizing the naming of genes across experiments, e.g. by means of orthology mappings, index \( t \) represents the same gene in all s microarray experiments. Assuming that, given \( I_t \), all s experiments are conditionally independent,

\[
P(I_t|D_{t,1}, X_{t,1}, \ldots, D_{t,s}, X_{t,s}) = \prod_{s} P(I_t|D_{t,s}|I_t, X_{t,s})
\]

provides a measure of shared gene importance. Typically, an analysis of shared gene function selects subsets of important genes from individual data sets and searches for genes shared by all subsets (e.g. Hockley et al., 2006). Unfortunately this discards rank information and in addition is prone to censoring of important genes. The latter will happen, if a gene that is in general among the most important ones appears in one experiment just below the threshold and does therefore not appear in the final list. An analysis of shared gene function along the lines of Equation (3) calculates shared importance directly and does therefore not suffer either drawback.

There are potential dangers if the proposed approach is implemented naïvely: Bayesian inference can suffer from sensitivity problems (Bernardo and Smith, 1994). As a result, the model probabilities in Equation (3) and the implied ranking might depend crucially on the hyperparameters used to parametrize the prior \( p(\beta|I_t) \). We avoid this problem by specifying a prior over all the hyperparameters that could contribute to such an adverse effect, and then inferring the hyperparameters as well.

2.2 Robust modelling of shared gene function

A model which takes these considerations into account is presented in Figure 1. The core of the model is a latent variable implementation (Andrieu et al., 2002; Holmes and Denison, 2003; Lee et al., 2003) of the binary classifier discussed earlier. We use index \( s \) as index over microarray experiments, \( t \) as gene index and \( n \) as observation index within experiments. Variable \( z_{nt,s} \) denotes a latent variable which, conditional on its parents, has a univariate Gaussian distribution with mean \( x_{nt}^s/\beta_s \) and unit SD. For \( I_t = 1 \), we use the log expressions as regressor \( x_{nt,s} \). For \( I_t = 0 \), the regression is based on a common reference model which uses only an intercept. The indicator \( I_t \) determines the component models for all s microarray experiments.
Equation (6) in the Appendix. GLM. Mathematical details of the joint distribution can be found in Denison et al. (2005). The posterior distributions over Gamma distributions as prior over the diagonal elements of matrices. The posterior distributions over all variables depend on all genes, the influence of the hyperparameters and on all information the data provides about these variables. Since precision is specified indirectly by giving the probability that any particular set of variables is relevant to the model, this allows inference of the probability measures of shared gene function. This directed acyclic graph (DAG) illustrates a probabilistic model for the analysis of shared gene function. Large rectangles indicate a replicated conditional independence relationship. Circles represent continuous and small squares discrete random variables.

Variable $\beta_{t,s}$ denotes the regression coefficients of the GLM. The prior over $\beta_{t,s}$ is a Gaussian distribution with zero mean and diagonal precision $\Lambda_s$. The robustness of inferring $P(I_t|D_{t1},X_{t1},\ldots,D_{tS},X_{tS})$ is improved by specifying this prior hierarchically and using a product of Gamma distributions as prior over the diagonal elements of $\Lambda_s$. This Gamma prior is parametrized by the coefficients $h_t$ and $g_s$. The binomial prior over $I_t$ is specified indirectly by giving the probability $\pi$ a beta prior with counts $s$. Using these hierarchical priors minimizes the sensitivity of $P(I_t|D_{t1},X_{t1},\ldots,D_{tS},X_{tS})$ to the choice of hyperparameters. The posterior distributions over $\Lambda_s$ and $\pi$ depend on the prior and on all information the data provides about these variables. Since both variables depend on all genes, the influence of the hyperparameters is greatly reduced. This important aspect of the model is further investigated in the experiments section below. Concerning the relationship between the latent variable, $z_{n,t,s}$, and the biological classifications, $y_{n,t,s}$, if $z_{n,t,s} < 0$, the probability $P(y_{n,t,s} = 0|z_{n,t,s})$ is 1, otherwise it is 0. By integrating over $z_{n,t,s}$, (Denison et al., 2002) show that this setting corresponds to a probit link GLM. Mathematical details of the joint distribution can be found in Equation (6) in the Appendix.

### 2.3 Variational inference

Inferring shared gene function requires calculating the marginal posterior distributions over all $I_t$ from the DAG in Figure 1. A Markov Chain Monte Carlo (MCMC) technique along the lines of Green, (1995) could, at the expense of high computational cost, approximate $P(I_t|D_{t1},X_{t1},\ldots,D_{tS},X_{tS})$ with arbitrary accuracy, but careful model checking would be essential and multiple MCMC runs would be required. With conventional computer infrastructure such an approach quickly becomes infeasible. A Laplace (Chu et al., 2005; MacKay, 1992) or a variational approximation (Attias, 1999, Frey, 1998; Jordan et al., 1999) is computationally simpler and better suited for our purposes. We decided to base model inference on the variational learning framework that was recently used in Beal et al. (2005). Variational learning implies approximating the joint distribution of the model [Equation (6) in the Appendix], by a factorizing Ansatz. For brevity we use $\theta$ for all random variables, $D = \{D_{t1},\ldots,D_{tS}\}$ and $X = \{X_{t1},\ldots,X_{tS}\}$ to obtain the approximation

$$p(\theta|G,H,\delta,D,X) \approx Q(\theta) = Q(\pi) \prod_t Q(\Lambda_t)$$

$$\times \prod_t \left( \prod_{I_t} Q(I_t) \prod_{Z_{n,t,s}} Q(z_{n,t,s}) \right).$$

We can now use Jensen’s inequality and obtain a lower bound on the log marginal likelihood of the DAG.

$$\log \left( \int_\theta p(\theta, D|G,H,\delta,X) \right) \geq$$

$$\int_\theta \left( \log p(\theta, D|G,H,\delta,X) - \log Q(\theta) \right) Q(\theta) d\theta$$

Variational learning requires maximizing the lower bound [second line in Equation (5)]. This is done iteratively by integrating the negative free energy with respect to all but one $Q$-distributions from Equation (4) and maximizing the resulting functional with respect to the remaining $Q$-distribution. This provides, for every $Q$-function in Equation (4), a separate update rule. The essential difference between typical modelling approaches and the DAG in Figure 1 is the hierarchical priors that couple all individual gene models. Details of the $Q$-function updates are provided in the Appendix.

### 2.4 Computation of shared gene function

An algorithm which infers shared gene function will iterate over all maximization steps for the $Q$-distributions in Equation (4) and monitor the negative free energy. Although there is no empirical proof of optimality, it is very important to use all possible safeguards to detect potentially misleading conclusions. In this case, we have to ensure that the calculated measure of shared gene function does not overly depend on the chosen prior and so a sensitivity analysis is vital. Such an analysis examines the effect of chosen hyperparameters on the probability measure of shared gene function $Q(I_t)$ in Equation (12) and $P(G = i|D,X)$ in Equation (13). An additional sanity check is obtained by cross-validation. This provides an estimate of generalization accuracy and several gene measures, each obtained from a slightly perturbed data set. Poor generalization accuracy or large variation in the gene measures would be warning signs. If one has prior biological knowledge about the gene expression assays, one can also check whether this is in agreement with an inference of Gene Ontology (GO) categories (Dopazo, 2006). This can be accomplished by inferring active GO biological process categories by Fisher’s exact test (Al-Shahrour et al., 2004; Lewin et al., 2006). Details regarding the implementation of our approach can be found in the online Supplementary Material (Sykacek et al., 2007).
2.5 Biological experiments

A combined analysis of a microarray time course of mouse mammary gland development (Clarkson et al., 2004) and an in-vitro culture of endothelial cells under growth factor withdrawal (Johnson et al., 2004) is used to infer genes that are related to cell number control in both assays. By converting the inferred measure of shared gene function into a list of active GO categories, this data is used as gold standard. The biological motivation for a combination of these assays is that apoptosis of endothelial cells is known to occur during the mammary gland cycle and may play an important role in this process (Djonov et al., 2001; Matsumoto et al., 1992). The prior knowledge that both assays involve apoptosis can be used as a benchmark for the analysis of active GO categories and thus implicitly as benchmark for the inferred gene measure.

3 RESULTS

Using synthetic data, we demonstrate in this section two advantages of the Bayesian approach over a standard method for shared gene analysis. An analysis of two microarray assays for shared gene function allows a discussion of important aspects for securing and assessing the reliability of inferred rank lists. In this context, we highlight the importance of diagnosing the influence of hyperparameters and demonstrate benchmarking of inferred rank lists which does not rely on ‘gold standard’ gene sets. Such benchmarking is obtained indirectly by predicting the biological state of independent test data to get generalization accuracy estimates. As a second benchmark, we compare prior knowledge about shared biological processes with GO terms which are inferred from the measure of shared gene function.

3.1 Synthetic data

Synthetic data serves two purposes. We illustrate the working principle of Bayesian modelling of shared gene function and compare the proposed approach with a simple meta analysis. To obtain candidates of genes that might be of shared relevance, the latter approach thresholds the $P$-value lists (e.g. a $t$-test of differential expression) obtained separately for each individual assay. Genes that appear in all lists are then assessed to be of shared relevance. Such analysis is for example used by Hockley et al. (2006). There are two aspects we want to highlight. Unlike the simple filtering-based meta analysis in Hockley et al. (2006), which can only reveal an unordered list of shared genes, the proposed Bayesian approach ranks genes with respect to shared gene function. In addition, all meta analyses which are based on thresholded gene lists are likely to suffer from a censoring effect. Censoring is caused by removing seemingly unimportant genes before combining the experiments. Such approaches imply that genes which are not selected in all assays are not of shared relevance. This is a dangerous simplification of the actual implication of $P$-values, which just assess, whether a certain observation can be explained by chance. A gene which is top ranked in one assay and has with a small $ε>0$ in a second assay a $P$-value of $0.01 + ε$, is potentially one of the most relevant common factors. Despite that, it will not appear in the list of genes that are of shared importance, if we use a $P$-value threshold of 0.01.

| Table 1. Means of Gaussian distributions used to generate synthetic data |
|--------------------------|----------------|-----------------|-----------------|
| Experiment | Gene group | Mean assay 1 | Mean assay 2 |
| Ranking | 1 | $± 2$ | $± 2$ |
| 2 | $± 0.5$ | $± 0.5$ |
| 3 | $± 0.05$ | $± 0.05$ |
| Censoring | 1 | $± 2$ | $± 2$ |
| 2 | $± 4$ | $± 0.5$ |
| 3 | $± 0.1$ | $± 0.1$ |

Data is generated by drawing ‘synthetic genes’ from Gaussian distributions with unit SD and means as are specified in Table 1. The sign before the mean indicates the dependency on the class label. Simulations assume two separate assays, each containing three groups of identically distributed variables. We used four synthetic genes per group, each replicated six times. The ranking experiment investigates the aspect of ordering with respect to the combined relevance of genes. Genes in earlier groups provide better class separability and we may expect that this corresponds to higher ranks in shared gene function. This overweights a simple approach of filtering for gene symbols that appear in separately thresholded lists, since this must inevitably lose rank information.

A second drawback of filtering individually thresholded gene lists for shared gene symbols is that this might censor important genes randomly. Thresholding converts a continuous measure into a binary decision about gene function and random effects like selecting particular biological replicates can alter the continuous measures. All borderline genes will thus, to some extent by chance, appear relevant or not. If one assay assesses particular genes as most relevant and a second assay finds these genes just outside the threshold, combining the lists will censor promising candidates. Using synthetic data, we can illustrate this effect. The key in the censoring experiment in Table 1 are the data in the second group, for which the first assay shows very large and the second only moderate class separability. Analysing the second assay separately will, for all genes in this group, result in $P$-values close to the selection threshold and thus in a random selection of the genes being picked as relevant. As a result, the combined gene list is censored randomly and only one of the genes from group two appears in the combined list. Results are in detail reported in the online Supplementary Material (Sykacek et al., 2007). Bayesian modelling of shared gene function provides the results in Table 2. Column ‘Evaluation of Ranking’ shows that the Bayesian approach indeed ranks the different groups as highly relevant. By inferring shared gene function directly, the Bayesian approach overcomes the censoring problem inherent to all meta analyses which are based on thresholded gene lists.
Table 2. Gene numbers and Bayesian measure of shared gene function \(Q(I_i)\) for two experiments

<table>
<thead>
<tr>
<th>Evaluation of ranking</th>
<th>Evaluation of censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene number</td>
<td>(Q(I_i))</td>
</tr>
<tr>
<td>Group 1</td>
<td></td>
</tr>
<tr>
<td>gene 4</td>
<td>0.999</td>
</tr>
<tr>
<td>gene 1</td>
<td>0.999</td>
</tr>
<tr>
<td>gene 3</td>
<td>0.999</td>
</tr>
<tr>
<td>gene 2</td>
<td>0.999</td>
</tr>
<tr>
<td>Group 3</td>
<td></td>
</tr>
<tr>
<td>gene 10</td>
<td>0.040</td>
</tr>
<tr>
<td>gene 11</td>
<td>0.039</td>
</tr>
<tr>
<td>gene 9</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unlike searching for common genes the proposed approach combines information about gene importance which provides the expected ordering of genes (evaluation of ranking) and avoids censoring of early genes (evaluation of censoring).

### 3.2 Searching for shared gene function

The Bayesian approach of inferring shared gene function can be applied in every situation where several microarray assays covering biological state transitions need be assessed for common genetic behaviour. Here, we are interested in inferring which genes are of shared relevance during two transitions: (1) the transition from lactation to involution in a mouse mammary gland cycle in vivo, and (2) the transition from normal growth conditions to growth factor withdrawal in an in vitro culture of human endothelial cells. The biological motivation for this investigation is that apoptosis of endothelial cells is known to occur during the mammary gland cycle and may play an important role in this process (Djonov et al., 2001; Matsumoto et al., 1992). Within the limit that the in vitro assay of endothelial cell apoptosis will not capture all aspects of those cells within the mammary gland tissue, both assays involve apoptosis of endothelial cells. A shared analysis has thus the potential to provide genetic markers that are involved in endothelial cell death within the mammary gland, even though no gene expression measurements were obtained from these cells in isolation. The mouse mammary gland data was taken from (Clarkson et al., 2004) and time-points were labelled as being apoptotic on the basis that apoptosis is induced in involution. Expression values were obtained from two biological replicates measured using Affymetrix Murine U74 arrays, with six samples from lactation and 10 samples during involution. From Johnson et al. (2004), who studied apoptosis in human endothelial cells, we took five samples under growth factor withdrawal and five control samples that were measured with Affymetrix human U95 arrays. Cross annotations were taken from the Affymetrix databases such that human and mouse genes that are orthologues were labelled the same. To increase consistency, we followed (Mecham et al., 2004) and took only such probes that could be matched in sequence by a NCBI blast search. This left us with 4581 cross annotated genes. The expression values were extracted with MAS 5.0, converted to log scale and normalized by within-slide mean removal and scale adjustment. To ensure that all regressors are on a similar scale, we transformed the log expressions of each gene to zero mean and unit SD. Analysis for shared gene function compared the gene specific GLMs with an intercept only GLM that models endothelial cell apoptosis and mammary tissue involution based on the prior frequencies of labels as observed in the training data.

#### 3.2.1 Sensitivity to prior choices

To ensure that inferences about shared gene function do not depend crucially on the chosen hyperparameters, we specify the priors \(P(I_i|x)\) and \(P(p_{i,j} | A_i)\) indirectly. We interpret \(p\) as the a priori fraction of genes we believe to be involved in the biological process. This requires us to be uninformative about \(p\) by using a small prior count like \(k=1\). The situation with \(A_i\) is more subtle, since our choice will have an indirect effect on \(Q(I_i)\). Therefore, it is imperative to study the sensitivity of the model to choices of \(g_i\) and \(h_i\). A sensitivity analysis will depend mainly on the variance, \(VAR[A_i[d, d]]\), with respect to the prior \(P(A_i | g_i, h_i)\). Thus, we may fix the prior expectation \(E[A_i[d, d]]\) to 0.01 and vary \(VAR[A_i[d, d]]\) linear on a log scale from \(10^{-6}\) to 1 by using \(g_i = [10^{-3}, 10^{-4}, \ldots, 10^4]\) and \(h_i = [10^{-3}, 10^{-2}, \ldots, 10^4]\). Computer simulations revealed that the prior over \(A_i[d, d]\) has no effect on the approximate posterior \(P(A_i[d, d])\), if we choose a variance that is larger than \(10^{-3}\). See online Supplementary Material for a graph with details of this investigation. The ordered probability measures \(Q(I_i, q)\) are in Figure 2 allow the same conclusion from the gene measure. Variances smaller than \(10^{-3}\) result in \(P(A_i | g_i, h_i)\) having an undesired effect on \(Q(I_i)\). Curves for prior variances that are larger than \(10^{-3}\) are not shown as they are essentially indistinguishable from the curve obtained for that value. Therefore, we concluded that sensitivity to the chosen prior is avoided if we set \(h_i \leq 10^4\) and \(g_i \leq 10^{-1}\).

#### 3.2.2 Analysis of gene function

If we assume equal cost for both types of error in deciding about gene function, we should select all genes that have \(Q(I_i, q)\) larger than 0.5. For the chosen hyperparameters, this suggests that 2164 genes are potentially of interest: a ranked list in tab-delimited format is provided in the online Supplementary Material. This is a large though plausible number of genes. Given that we chose uninformative hyperparameters, we see that the hierarchical model allows the prior over regression coefficients to adjust to the data sets. Here, the data favours small regression coefficients. An illustration of this effect on synthetic data is provided in the online Supplementary Material. For the mammary gland data, the expected variances in the priors over intercept and regression coefficients are 0.13 and 0.93. The respective values for the apoptosis data are 0.11 and 1.77. The small variance of the effective prior over the regression parameter implies a small complexity penalty for the larger model (Jefferys and Berger, 1992). Therefore genes are favoured over the intercept-only model, even if they provide only a little information about the biological
classification. To validate the result, we used the model for
10-fold cross testing. Figure 3 illustrates the fold variation of
the 10 largest values of the gene measure, $P(G = t | D)$.

We see that fold-based rankings and the overall ranking give
similar results. However, there are some deviations which
indicate that some slides are more influential than others.
This effect should be reduced by using larger sample sizes.
Cross testing is based on averaging predictions which
are weighted according to Equation (13) (cf. Sykacek et al.,
2007 for further algorithmic details). Selecting the top-
ranked gene predictions [until the cumulative gene measure,
$\sum P(G = t | D)$, reaches 0.8], produces on average 424
genes, and we obtain for both data sets a generalization
accuracy of 100%. High generalization accuracy is reassuring
since it suggests that the probability measure did favour
informative genes.

To assess the biological plausibility of our shared gene
measure, we followed (Lewin et al., 2006), who inferred active
GO categories by Fisher’s exact test. To do so, we regarded the
top 30% of the genes from the rank list as active and the 30%
genes at the lower end as inactive and inferred, for every
GO category, the significance level of abundance of active over
inactive genes. To increase the robustness of this approach,
we used the fold-based gene measures as they arose from
estimating the generalization accuracies. A gene is counted as
active, if its indicator probability $Q(I_t | fold)$ is larger than 0.5 and
the regression coefficients $\beta_t$ have the same sign for all
experiments $s$. To adjust for multiple testing, we map the
$P$-values to false discovery rates (Fdr) (Benjamini and
Hochberg, 1995) and obtain at the 0.05 threshold 588 GO
categories with a significant abundance of active over inactive
genes. Such analysis finds many GO categories that are
indicative of shared metabolic changes. Our prior expecta-
tion that both assays have certain events related to cell
death is confirmed since we find ‘cell death’ (GO:0008219),
‘programmed cell death’ (GO:0012501), ‘regulation of
programmed cell death’ (GO:0043067), ‘negative regulation of
programmed cell death’ (GO:0043069), ‘regulation of
apoptosis’ (GO:0042981), ‘negative regulation of apoptosis’
(GO:0043066), ‘anti-apoptosis’ (GO:0006916), ‘caspase activa-
tion via cytochrome c’ (GO:0008635), ‘induction of apoptosis
via death domain receptors’ (GO:0006825) ‘apoptotic mito-
chondrial changes’ (GO:0008637), ‘DNA damage response,
signal transduction resulting in induction of apoptosis’
(GO:0008630) and ‘apoptosis’ (GO:0006915) are, with high
significance, enriched by active genes. An XML file with all
active GO categories is provided as an online Supplementary
Material. A list of active apoptosis genes is presented in
Table 3. This list was obtained by selecting all co-regulated
genes that were annotated to the apoptosis subgraph of the GO
DAG and by ranking according to the measure of shared gene
function $P(I_t | D)$. For a complete gene list, we refer to the online
Supplementary Material.

As a result of the hierarchical prior chosen in this work,
we find in an equal cost scenario many genes with func-
tional importance. It is evident from Figure 2 that we will
obtain fewer genes with probabilities larger than 0.5 by
forcing small precisions in the prior over regression coefficients.
This, however, implies constructing a convenient probability
measure that has little support from the data. Instead,
we recommend a pragmatic approach of taking as many
genes as one can afford in subsequent steps according to the
ranking of shared functional importance, and possibly using
additional criteria such as unknown GO or pathway
annotation.

4 DISCUSSION

In this article, we propose a probabilistic model for a principled
integrated analysis of several microarray experiments. The
proposed model is a result of careful considerations of
sensitivity issues. By specifying priors hierarchically, we
reduce the effect of all hyperparameters of the algorithm
and provide conclusions that are justified by the data. The
proposed approach shares with meta analyses (DeConde et al.,
2006; Hockley et al., 2006; Yang et al., 2005) the advantage
of combining data sets where the actual expression levels
of different experiments need not be matched. A consider-
able advantage of a Bayesian analysis is that it provides
rank information and does not suffer from the censoring
effects of simple approaches that combine thresholded
gene lists.

An application to shared analysis of gene function in mouse
mammary gland tissue and an endothelial cell line illustrates
how to diagnose and avoid potential sensitivity problems.
Assessments of predictive accuracy and a confirmation of
biological expectations reinforce the plausibility of the
proposed approach. The results suggest that avoiding sensitiv-
ity is imperative in analysing microarray data, even if one
cannot follow up a large number of positive genes. The
proposed approach has two desirable properties. First, it allows
us to increase statistical power and cost efficiency by combining
new assays with existing data. Even more important is the
prospect of a successful search for molecular biological

---

**Fig. 2.** Ordered probabilities of gene function plotted over all genes for
different priors over $\lambda_t | d_t$. The expectation of $\lambda_t | d_t$ is fixed at 0.01
and the prior variances change from $10^{-3}$ to $10^{-6}$.
mechanisms that are shared by developmental processes or state transitions in different tissues or species. A fully Bayesian analysis for shared gene function therefore has the potential to lead to biological insights that might be unobtainable by other means.

ACKNOWLEDGEMENTS

The authors thank David MacKay for his advice and the editor and the reviewers for their helpful comments. This work was supported by the BBSRC’s Exploiting Genomics initiative under ref. 8/EGH16106, ‘Shared Genetic Pathways in Cell Number Control’. P. S. is currently funded by the WWTF, ACBT, ARC Seibersdorf and Baxter AG and grateful for their support.

Conflict of Interest: none declared.

REFERENCES


Clarkson, R. W. E. et al. (2004) Gene expression profiling of mammary gland development reveals putative roles for death receptors and...


APPENDIX

Joint distribution

If we abbreviate the regression coefficients of all \(T\) genes and \(S\) experiments as \(B = \{b_{1,1}, \ldots, b_{T,S}\}\), all indicators as \(J = \{I_1, \ldots, I_T\}\), the precision matrices as \(L = \{L_1, \ldots, L_S\}\), the hyperparameters as \(G = \{g_1, \ldots, g_S\}\) and \(H = \{h_1, \ldots, h_S\}\), we may express the joint distribution as

\[
p(\pi, L, B, J, Z, D|G, H, \delta) = p(\pi|\delta) \times \prod_{s=1}^{S} p(A_s|g_s, h_s) \prod_{t} p(I_t|\pi) \times \prod_{t=1}^{T} \prod_{s=1}^{S} p(b_{t,s}|A_s, I_t) \times \prod_{n=1}^{N} \left( p(z_{n,t,s}|b_{t,s}, I_t) p(y_{n,t,s}|z_{n,t,s}) \right).
\]

The conditional probability of \(y_{n,t,s}\) given \(z_{n,t,s}\) is

\[
P(y_{n,t,s} = 1|z_{n,t,s} = 1) = \begin{cases} 1, & \text{if } z_{n,t,s} > 0 \\ 0, & \text{if } z_{n,t,s} \leq 0 \end{cases}
\]

The conditional probability of \(z_{n,t,s}\) given \(b_{t,s}\) and \(I_t\) is a univariate Gaussian

\[
p(z_{n,t,s}|b_{t,s}, I_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-0.5\left( z_{n,t,s} - b_{t,s} \right)^2 \right)
\]

The prior over \(b_{t,s}\) is a multivariate Gaussian

\[
p(b_{t,s}|A_s, I_t) = \frac{1}{\sqrt{2\pi}\sigma} |A_s|^{0.5} \exp\left(-0.5b_{t,s}'A_s^{-1}b_{t,s} \right)
\]

The prior over \(I_t\) is Bernoulli distributed

\[
p(I_t|\pi) = \pi^I_t (1-\pi)^{1-I_t}
\]

The prior over \(A_s\) is a product of Gamma distributions

\[
p(A_s|g_s, h_s) = \prod_{d} \left( \frac{h_s}{g_s} \right)^{\frac{1}{2}} \Gamma(\frac{1}{2}) \Gamma(\frac{h_s}{g_s}) \exp(-A_s[d,d]g_s^{-1})
\]

and the prior over \(\pi\) a Beta distribution

\[
p(\pi|\delta) = \frac{\Gamma(\delta + d_2)}{\Gamma(\delta)} \frac{1}{(\delta + d_2)!} \exp(-\delta_2^{-1} (1-\pi)(\delta_2^{-1} - 1)).
\]

To indicate a conditional dependency on \(I_t\), we use the latter as index. The equations link to the graph in Figure 1 by \(\pi = [\Theta, (1-\Theta)]\) (i.e. a two-state probability) and \(\delta = [\delta_1, \delta_2]\) specifying the prior counts in the distribution over \(\pi\). The \(d\)-th diagonal element of the matrix \(A_s\) is denoted as \(A_s[d,d]\).
Variational maximization

Variational learning follows the generic approach sketched in Section 2.3. We iterate over integrating the negative free energy from Equation (5) with respect to all but one Q-distributions and maximizing the resulting functional. For the Q-distributions in Equation (4) we get the following update equations.

Maximizing with respect to \( Q(z_{i,n,t,s}) \) results in a truncated normal distribution

\[
Q(z_{i,n,t,s}) = (2\pi)^{-0.5} \frac{1}{\Phi(b_{n,t,s}) - \Phi(a_{n,t,s})} \exp\left(-0.5(z_{i,n,t,s} - \hat{z}_{i,n,t,s})^2\right) 
\]

where \( \hat{z}_{i,n,t,s} = x_{i,n,t,s}^T \tilde{\beta}_{i,t,s} \).

We use \( \tilde{\beta}_{i,t,s} \) for the mode of the Q-distribution over \( \beta_{i,t,s} \). The expressions \( \Phi(a_{n,t,s}) \) and \( \Phi(b_{n,t,s}) \) denote Gaussian cdfs with mean \( \hat{z}_{i,n,t,s} \) and unit SD at \( a_{n,t,s} \) and \( b_{n,t,s} \). The latter are implied by the definition of \( P(y_{n,t,s}|z_{i,n,t,s}) \); for \( y_{n,t,s} = 1 \), we get \( a_{n,t,s} = 0 \) and \( b_{n,t,s} = \infty \); for \( y_{n,t,s} = 0 \), we get \( a_{n,t,s} = -\infty \) and \( b_{n,t,s} = 0 \).

Maximizing with respect to \( Q(\beta_{i,t,s}) \) results in a Gaussian distribution

\[
Q(\beta_{i,t,s}) = (2\pi)^{-0.5} \delta_{\beta_{i,t,s} \beta_{i,t,s}} \exp\left(-\frac{1}{2}(\beta_{i,t,s} - \beta_{i,t,s})^T \hat{\Lambda}_{i,t,s}^{-1} (\beta_{i,t,s} - \beta_{i,t,s})\right) 
\]

where \( \hat{\Lambda}_{i,t,s} = \langle \Lambda_{i,t,s} \rangle + \sum_{n=1}^{N_t} x_{i,n,t,s} x_{i,n,t,s}^T \).

\[
\beta_{i,t,s} = \langle \beta_{i,t,s} \rangle + \sum_{n=1}^{N_t} x_{i,n,t,s} z_{i,n,t,s} \]

\[
\langle z_{i,n,t,s} \rangle = \hat{z}_{i,n,t,s} - f(b_{n,t,s}) - f(a_{n,t,s}) 
\]

\[
\hat{\beta}_{i,t,s} = \langle \beta_{i,t,s} \rangle \quad \hat{\Lambda}_{i,t,s} = \langle \Lambda_{i,t,s} \rangle 
\]

with \( \langle \Lambda_{i,t,s} \rangle = \text{diag}(\hat{\lambda}_{i,1}, \ldots, \hat{\lambda}_{i,N_t}) \) denoting the expectation under the Q-distribution. In addition to previously defined symbols, \( f(b_{n,t,s}) \) and \( f(a_{n,t,s}) \) denote Gaussian density functions with mean \( \hat{z}_{i,n,t,s} \) and unit SD. We use \( \Lambda_{i,t,s} \) to indicate that \( I_t \) will select a sub matrix of \( \Lambda_i \).

Maximizing with respect to \( Q(\Lambda_i) \) results in a product of Gamma distributions over the diagonal terms of the prior precision matrix \( \Lambda_i \).

\[
Q(\Lambda_i|d, d) = \frac{\hat{\gamma}_{i,d}}{\Gamma(\hat{\gamma}_{i,d})} \Lambda_i |d, d|^\hat{\gamma}_{i,d}^{-1} \exp(-\Phi(\lambda_{i,d}^T|d, d)^{\hat{\gamma}_{i,d}}) 
\]

where \( \hat{\gamma}_{i,d} = \gamma_i + \frac{1}{2} \sum_{l=1}^{T} \sum_{|d_i|+\sum_2 \sum_2 Q(I_l)} 
\]

\[
\hat{\lambda}_{i,d} = \lambda_i + \frac{1}{2} \sum_{l=1}^{T} \sum_{|d_i|+\sum_2 \sum_2 Q(I_l)} 
\]

Maximizing with respect to \( Q(\tau) \) results in a Beta distribution over the binary probability \( \tau \)

\[
Q(\tau) = \frac{\Gamma(\hat{\delta}_1 + \hat{\delta}_2)}{\Gamma(\hat{\delta}_1)\Gamma(\hat{\delta}_2)} \theta^{(\hat{\delta}_1-1)}(1 - \theta)^{(\hat{\delta}_2-1)} 
\]

\[
\hat{\delta}_1 = \delta_1 + \sum I_l = 1 \quad \hat{\delta}_2 = \delta_2 + \sum I_l = 0 
\]

Inferring probabilities of shared gene function requires maximizing the lower bound of the log marginal likelihood in Equation (5) with respect to \( Q(I_t) \) which results in Bernoulli distributions over \( I_t \).

\[
Q(I_t) = \prod_i P_i^{\hat{I}_t(i|y_i)} \quad \text{where} \quad P_i = \frac{\exp(f_i)}{\sum_j \exp(f_j)} 
\]

\[
f_i = \psi(\hat{\delta}_i) - \psi(\hat{\delta}) + \sum_j \left( \frac{1}{2} \log |\hat{\Lambda}_{i,t,s}| \right) 
\]

\[
+ \frac{1}{2} d_{i,t,s} + \frac{1}{2} \sum_{b=1}^{d_{i,t,s}} (\psi(\hat{\gamma}_{i,d}) - \log(\hat{\delta}_{b,d})) 
\]

\[
- \frac{1}{2} (\hat{\beta}_{i,t,s}^T \hat{\Lambda}_{i,t,s}^{-1} \hat{\beta}_{i,t,s} + \text{tr} \hat{\Lambda}_{i,t,s}^{-1} \hat{\Lambda}_{i,t,s}) 
\]

\[
+ \sum_{n=1}^{N_t} \left( \log(\Phi(\beta_{n,t,s}) - \Phi(\alpha_{n,t,s})) \right) 
\]

\[
- \frac{1}{2} x_{i,n,t,s}^T \hat{\Lambda}_{i,t,s}^{-1} x_{i,n,t,s} \right) 
\]

We use \( I_t \) as index to indicate the conditional dependency of the variables on the model indicator. \( \beta_{i,t,s} \) for the mode of the Q-distribution over \( \beta_{i,t,s} \) and \( \Lambda_{i,t,s} \) as the expectation of \( \Lambda_{i,t,s} \) under the Q-distribution. The expressions \( \Phi(a_{n,t,s}) \) and \( \Phi(b_{n,t,s}) \) denote Gaussian cdfs with mean \( \hat{z}_{i,n,t,s} \) and unit SD at \( a_{n,t,s} \) and \( b_{n,t,s} \). The latter are implied by the definition of \( P(y_{n,t,s}|z_{i,n,t,s}) \); for \( y_{n,t,s} = 1 \), we get \( a_{n,t,s} = 0 \) and \( b_{n,t,s} = \infty \); for \( y_{n,t,s} = 0 \), we get \( a_{n,t,s} = -\infty \) and \( b_{n,t,s} = 0 \). In addition we have \( \psi(x) \) as the digamma function, \( \hat{\delta} = \sum I_t \hat{\delta}_i \) and \( \text{tr} \) to denote the matrix trace operation. Shared gene function of the \( t \)-th gene is captured by \( Q(I_t = 1) \). We provide with \( Q(I_t) \) an approximate measure of shared gene function, which compares a gene specific model against a common alternative. We can thus assess genes relative to each other by transforming all \( Q(I_t) \) into one measure over a \( T \)-dimensional ordinal variable \( G \).

\[
P(G \equiv I(D, X)) \approx \frac{Q(I_t = 1)}{Q(I_t = 0)} / \left( \sum_{k=1}^{T} Q(I_k = 1) / Q(I_k = 0) \right) 
\]