**ABSTRACT**

Motivation: A fundamental task in systems biology is the identification of groups of genes that are involved in the cellular response to particular signals. At its simplest level, this often reduces to identifying biological quantities (mRNA abundance, enzyme concentrations, etc.) which are differentially expressed in two different conditions. Popular approaches involve using \(t\)-test statistics, based on modelling the data as arising from a mixture distribution. A common assumption of these approaches is that the data are independent and identically distributed; however, biological quantities are usually related through a complex (weighted) network of interactions, and often the more pertinent question is which subnetworks are differentially expressed, rather than which genes. Furthermore, in many interesting cases (such as high-throughput proteomics and metabolomics), only very partial observations are available, resulting in the need for efficient imputation techniques.

Results: We introduce Mixture Model on Graphs (MMG), a novel probabilistic model to identify differentially expressed submodules of biological networks and pathways. The method can easily incorporate information about weights in the network, is robust against missing data and can be easily generalized to directed networks. We propose an efficient sampling strategy to infer posterior probabilities of differential expression, as well as posterior probabilities over the model parameters. We assess our method on artificial data demonstrating significant improvements over standard mixture model clustering. Analysis of our model results on quantitative high-throughput proteomic data leads to the identification of biologically significant subnetworks, as well as the prediction of the expression level of a number of enzymes, some of which are then verified experimentally.

Availability: MATLAB code is available from http://www.dcs.shef.ac.uk/~guido/software.html

Contact: guido@dcs.shef.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Research in molecular biology has led to the uncovering of an exceptional wealth of information about the interactions of biological quantities in living cells. This information is conveniently represented using large graphs or networks, often with hundreds of nodes in each connected component, and is stored in highly accessed databases such as KEGG (Kanehisa and Goto, 2000), TRANSPATH (Krull et al., 2006), etc. According to the type of interaction they describe, these networks can be further classified into regulatory networks (protein-DNA interactions), protein–protein interaction networks, signalling pathways and metabolic networks.

These networks provide a large-scale picture of the interactions within a cell; however, it is largely a static picture, and does not account well for the dynamics of cellular processes. For example, biological intuition suggests that cells will react to external signals by switching on in a coordinated fashion small/middle scale subnetworks of these large networks; this is supported by the observation that in general external signals result in an alteration in the expression levels of relatively small sets of proteins. How to determine, experimentally or computationally, exactly which subnetworks correspond to a particular signal is an important question in systems biology.

Traditionally, identifying differential expression and subnetworks have been treated as two separate tasks. A wide variety of statistical and probabilistic methods have been used to determine differential gene expression from microarray data: these involve \(t\)-testing (Tusher et al., 2001), empirical Bayes (Newton et al., 2001, Efron et al., 2001) and mixture models (Ghosh, 2004), to cite but a few. The results of the statistical analysis are then mapped onto a network structure (derived for example from one of the sources cited above) to determine that certain submodules are consistently differentially expressed. Crucially, however, the network structure is not used in the statistical analysis.

Intuitively, a priori usage of the network structure could lead to significant advantages in the statistical analysis: subtler changes in expression could be detected if consistent patterns appear in the neighbouring nodes, inferences could be made more robust against noise, and missing data could be imputed using the network structure. While microarrays now provide reasonably clean and complete data for gene expression, issues of robustness to noise and missing data are vital in the analysis of proteomic and metabolomic data, where high-throughput mass spectrometry (MS) is particularly vulnerable to noise,
and can at best interrogate a small fraction of an organism’s proteome.

Two recent studies have been exploiting these ideas for detecting differential expression in microarray data. Rapaport et al. (2007) use the network structure to obtain a graph Laplacian which can then be used to Fourier-transform expression profiles. The idea here is that different conditions (e.g. differentially versus non-differentially expressed genes) will have widely differing spectral profiles. While this method is elegant and produced biologically meaningful results, it is not a probabilistic method, so it is difficult to assess the level of confidence in the results, or how the method could be modified to cope with missing data.

Almost simultaneously, Wei and Li (2007) proposed a probabilistic model for using a priori information on network structure in the determination of differentially expressed genes. The model borrowed techniques from image analysis using Markov random fields (MRFs) to model the interdependencies between the latent variables representing differential expression. The authors could then draw upon efficient approximate estimation techniques to maximize the likelihood of the model to obtain estimates of the parameters and of the latent variables. However, the model’s complexity rules out a proper Bayesian analysis, so that this approach can only supply point estimates of the quantities of interest. In particular, there is no obvious way to estimate posterior probabilities of differential expression, and it is not clear how a maximum likelihood strategy could cope with the severe multimodality missing data would introduce.

In this contribution, we focus primarily on the analysis of high-throughput quantitative proteomic data for the study of cell metabolism. This type of data is becoming increasingly available through the development of technologies such as isobaric tagging for relative and absolute quantification (iTRAQ) [Ross et al., 2004], which allow the use of LC-tandem Mass Spectrometry (LC-MS/MS) for quantitative proteomics. The characteristics of this type of data, however, make its analysis much more challenging than conventional gene-expression data analysis. For a start, only a small fraction (generally around 10–20%) of the proteins in a metabolic network are usually measured. Secondly, there are potentially several sources of noise involved in the data generation, meaning that the signal to noise ratio can often be relatively small, making a probabilistic analysis of the data mandatory. On the other hand, identification of significant submodules of the metabolic network is an extremely important task: it allows not only the prediction of which enzymes are differentially expressed, but also which metabolites are required in the process. This carries profound consequences for the understanding of the cellular process, with immediate implications for engineering and synthetic biology applications.

We introduce Mixture Model on Graphs (MMG), a fully Bayesian probabilistic model to integrate a priori network structure information into the statistical analysis of high-throughput data. The model defines a conditional prior on class membership which allows us to define a simple and efficient Gibbs-sampling strategy. The model can easily accommodate missing data, weighted networks and multi-class classification. Extensive testing on simulated data reveals a marked improvement in performance over standard mixture modelling estimation. We then apply the model to a recently published proteomic data set (Stensjö et al., 2007) investigating nitrogen fixation in cyanobacteria. The results of the analysis provide biologically meaningful subnetworks, which well elucidate the cellular processes under way. The dataset contains a large proportion of missing data, and, as a by-product, our model provided posterior estimates of the expression level of these enzymes. Some of these predictions have been validated independently by a recent experimental assay (Ow et al., 2008).

2 MODEL AND METHODS

Let us assume, we have (possibly incomplete) observations \( y_j \) \( j=1, \ldots, N \) of some biological quantity of interest (e.g. mRNA expression levels from microarray experiments or log-ratios of protein concentrations from MS measurements). These quantities are linked together in a weighted network, whose structure is encoded in a connectivity matrix \( X \). \( X_{ij} \) is zero if there is no edge in the network connecting nodes \( i \) and \( j \), and \( X_{ij} = w_{ij} \) a positive constant (weight) encoding how strongly linked nodes \( i \) and \( j \) are. In many applications, for example for gene networks, these weights can be all set to 1, if there is no reason to believe some interactions are more important than others. There are however important examples where weighted graphs are essential: this is true in particular in metabolic networks where edges represent metabolites whose frequency in the network gives a natural weighting (Croes et al., 2006).

We assume that the data can be partitioned in \( d \) different classes, represented by latent variables \( c \) which are \( d \)-dimensional binary vectors: \( c = 1 \) indicates that \( y \) belongs to class \( p \). In the simplest case, \( d \) could be just equal to two and the two classes would then be differentially or non-differentially expressed (this is the situation considered by Wei and Li, 2007). In the analysis of metabolic networks, we are interested in both up-regulated and down-regulated submodules, so we will select \( d = 3 \) (allowing for an unchanged class). The probability of observing a certain value \( y_j \) will clearly depend on the latent class variable, as well as on a number of parameters. We will denote this probability by \( p(c_j | c, \theta) \); the explicit form of these class-conditional distributions will depend on the specific problem at hand.

2.1 Conditional prior model

We will incorporate network structure in our inference of the latent class variables through the use of conditional priors. In practice, instead of assuming the class variables to be independent a priori of each other, we will assume that the class of the \( j \)-th node is a priori influenced by the class of its neighbours. This is conveniently encapsulated in a choice of conditional prior distribution for the class variables, Wei and Li (2007) chose a MRF as a conditional prior; our choice will be much simpler and motivated essentially by robustness considerations.

We define the class probability for node \( j \) given its immediate neighbours \( j \) as

\[
p(c_j | e) = \text{Discrete}(ae + m_j)
\]

where \( e \propto 1 \) is a vector with identical entries and

\[
m_j = \sum_{i=1}^N w_{ij} e_i.
\]

Here we use the Discrete (also called Categorical) distribution over a binary vector defined as \( v \sim \text{Discrete} (\theta) \) if \( v_i = 1 \) with probability \( \theta_i \) (in particular \( \theta = (1) \)). \( m \) is given by the sum of the class membership
vectors for neighbouring nodes, weighted by the weight assigned to the connecting edges. So, if a node is connected to an up-regulated node by a very strong link, the prior probability of it being up-regulated will be increased accordingly. On the other hand, we do not want to rule out the possibility that a node will belong to a class that is different from the ones represented among its neighbours, hence the addition of a regularizing constant vector $\alpha e$. Different choices for the parameter $\alpha$ allow to interpolate smoothly from a standard mixture model (with $\alpha$ very large) and a (possibly degenerate) model where the network structure contains all the prior information ($\alpha = 0$).

### 2.2 Class conditional densities and posterior distributions

The conditional distribution for the expression of a node given its latent class is a modelling issue that needs to be tailored to the problem at hand. For example, for the detection of differential expression in mRNA levels, a Gamma-Gamma model has been proposed (Newton et al., 2001) where the two classes (differentially and non-differentially expressed genes) were modelled as two Gamma distributions with different parameters.

There is no currently accepted model for differential expression in high-throughput metabolomic and proteomic data. Drawing from experimental observations, a sensible model of MS data would need to consider long-tailed distributions, but at the same time allow for a large amount of noise in the system, so that up-regulated proteins or enzymes can often correspond to a log ratio that is barely positive. We therefore propose the following model for log-ratios of protein concentrations

$$p(y|c_i) = \begin{cases} \lambda_+ \exp(-\lambda_+y_i) & \text{if } c_i = 1 \\ \lambda_- \exp(-\lambda_-y_i) & \text{if } c_i = 0 \end{cases}$$

The model is a mixture of three components: two exponential distributions representing the up-regulated and down-regulated classes (with parameters $\lambda_+$ and $\lambda_-$, respectively), and a normal distribution representing the unchanged class. Notice that this model does not admit the possibility that an up-regulated protein has an observed expression amount of noise in the system, so that up-regulated proteins or enzymes can often correspond to a log ratio that is barely positive. We therefore propose the following model for log-ratios of protein concentrations

$$p(y|c_i) = \begin{cases} \lambda_+ \exp(-\lambda_+y_i) & \text{if } c_i = 1 \\ \lambda_- \exp(-\lambda_-y_i) & \text{if } c_i = 0 \end{cases}$$

The result is a sample-based approximation to the true posterior distribution. An advantage of this sampling strategy is that all steps are efficiently, Gibbs sampling has become one of the most popular

One of the main benefits in being able to obtain an explicit expression for the conditional posterior distributions is the ease with which a Gibbs sampling algorithm can be set up. Gibbs sampling (see e.g. Gelman et al., 2004) is a type of Metropolis-Hasting algorithm where the proposal distribution is given by the conditional-posterior distribution. An advantage of this sampling strategy is that all steps are accepted in the Markov chain, leading to much faster convergence to the target distribution (in this case, the full posterior distribution over latent variables and parameters). Because of its relative simplicity and efficiency, Gibbs sampling has become one of the most popular sampling approaches, with applications in a wide variety of statistical problems.

Essentially, our algorithm proceeds as follows:

1. Obtain initial estimates of class assignments (for example, by partitioning the data according to their expression levels)
2. Given the class assignments, obtain initial estimates of the model parameters
3. Iteratively sample class assignments from the conditional posteriors
4. Iteratively sample parameter values from the conditional posteriors over the parameters
5. Repeat until a sufficiently large sample has been obtained.

The result is a sample-based approximation to the true posterior distribution over the class memberships and the parameters.

### 2.4 Monitoring convergence

One of the difficult issues in sampling methods is assessing the convergence of the Markov chain to the target distribution. While there are no unequivocal rules about this, there are a few heuristic recipes that can help (Gelman et al., 2004).

- Run parallel chains from widely distributed starting points. Monitoring a few diagnostic quantities of interest during the Markov chains, it is possible to determine the degree of mixing of the chains, and assess when a reasonable convergence to the target distribution is achieved.
- Allow for a burn in period, discarding the initial part of a Markov chain.
- Once a state close to convergence is reached, subsample from the Markov chain to reduce the autocorrelation between samples. While in theory for large enough samples the autocorrelation...
should not give problems, it is still advisable to thin the Markov chain.

In our experiments, we adopted all of these approaches. It is to be born in mind though that Gibbs sampling is an efficient strategy and difficulties in convergence arise mainly out of correlation between variables. Since biological networks typically have low connectivity (i.e. they are sparse), we expect our Markov chains to reach convergence reasonably quickly. This is confirmed by the results reported below.

2.5 Identifying submodules
Having obtained posterior probabilities of class membership, one can then produce class assignments based on suitable criteria (e.g. highest posterior probability, or setting thresholds on the posterior probabilities). The identification of coherent submodules can then be performed using a simple percolation algorithm, whereby starting at a given node, the submodule is grown by iteratively adding all the neighbouring nodes in the same class. This is only one possible method (Sanguinetti et al., 2006), but provides a simple and efficient strategy.

3 RESULTS
3.1 Synthetic data
To validate and benchmark our model, we ran an extensive study of its behaviour on simulated data. It has been observed that many metabolic networks appear to have a scale free structure (Jeong et al., 2000). Therefore, we grew scale free networks using the algorithm of Barabasi and Albert (1999). The networks we used had 100 nodes and an average connectivity (average number of neighbours per node) of ~2. We then generated class memberships by using a sample from a Markov chain generated using the conditional priors of equation (1) from a random initialization, and simulated expression levels obtained from the class conditional model of equation (2). The values of the parameters were set to $\sigma = 0.3$, $\lambda_+ = 1.7$ and $\lambda_- = 1.3$, which gave a fairly large overlap among classes in terms of expression level. The parameter $\alpha$ in equation (2) was fixed to 1 in the inference. It is important to bear in mind that, since our model is defined in terms of conditional priors, it is not a generative model, so it is not straightforward to simulate data from the model. The synthetic class memberships we use are not generated by the model, but have been sampled so that it exhibits some spatial patterning intuitively analogous to the situation we wish to model.

Sampling was carried out using the procedure described in the previous section. In each case, a burn-in period of 1000 iterations was kept, and 1000 samples were taken from the following 5000 iterations with a thinning of 5.

Figure 1a shows a comparison between our model and a standard mixture model using the same conditional distributions and with parameters $\lambda_+$ estimated using the EM algorithm (Dempster et al., 1977) (both MMG and the standard mixture model used the true value of $\sigma = 0.3$; see the Supplementary Material for details of how the EM algorithm was implemented). The comparison was repeated across ten different network structures of the same size, and for each structure 20 different sets of class assignments and expression data were generated as described above. In all cases but one our model clearly outperforms the model were network structure is not included. In one case the standard mixture model does slightly better (although not with any statistical significance); curiously, the parameter values inferred by the EM procedure are quite different from the true values (which our model estimates well) in this case. A mixture model with the correct values actually underperforms our model in this specific case.

As we pointed out in section 2.4, a crucial question in sampling-based methods is the speed of convergence to the target distribution. Figure 1b shows the first 100 samples of the parameter values $\lambda_+$ and $-\lambda_-$ from two separate Markov chains with dispersed starting points. As can be observed,
the Markov chains mix very quickly (within 10 iterations) and converge fast to the true value (shown as a solid line). While these are only two possible diagnostics, we feel confident that a burn-in of 1000 cycles is more than appropriate to reach convergence on any reasonable network structure.

We also questioned the robustness of our model to missing data. To do so, we sampled 20 different class assignments and data for a fixed network structure with 100 nodes. We then removed at random data with probability 20% and compared the results of the inference performed with the complete data and the incomplete data. The results showed that our model is remarkably robust to missing data: on average, 93% of the class assignments made with the complete data coincided with the assignments made with the incomplete data (SD of 3%). At random, one would only expect ~86% of inferences to coincide, so the result shows the predictive power that the inclusion of network structure can bring.

Finally, we investigated the dependence of the results of our analysis on the parameter $\alpha$ which sets the scale of the regularizing coefficient in equation (2). To do so, we generated synthetic data with the procedure described above with $\alpha = 1$, and then ran MMG with 19 different values of $\alpha$ ranging from 0.1 to 10. To assess the degree of similarity of the results, we used the probability of being up-regulated as a diagnostic. We first transformed the probabilities using the logit function

$$f(z) = \log \frac{z}{1 - z}$$

to map them onto the real space. We then analysed the covariance between these 20 results. Strikingly, 92.4% of the variance in the transformed probability was in the direction spanned by the diagonal vector $e \propto (1, \ldots, 1)$ (this percentage rose to 96.2% if we restricted to nodes that were assigned to the up-regulated class with a probability $>0.5$). This indicates a very high degree of consistency for the model results across two orders of magnitudes in variation for $\alpha$. While there are some differences, particularly where the assignments are not clear cut, these results indicate a weak dependence on the choice of the parameter $\alpha$ (within a reasonable range). A similar analysis on the sensitivity to the parameter $\sigma$ also revealed weak dependencies, with very consistent predictive performance over a range between 0.1 and 1. Further details are contained in the supplementary material.

### 3.2 Nitrogen metabolism of Nostoc

We now consider a recently published proteomic dataset (Stojnouch et al., 2007). In this experiment, an iTRAQ-based study of the proteome of the cyanobacterium Nostoc sp. PCC7120 was carried out comparing N$_2$ fixing conditions with non-N$_2$ fixing conditions. Nostoc is an important genus of cyanobacteria due to its ability to metabolize nitrogen to produce ammonium, a reaction catalysed by the enzyme nitrogenase that in turn results in the production of hydrogen. These properties make Nostoc, and more generally cyanobacteria, some of the most promising organisms for the bacterial production of renewable fuels (for a recent review see e.g. Rupprecht et al., 2006). Hydrogen production, however, happens only under specific environmental conditions (N$_2$-fixing conditions), so it is of great relevance to identify the particular pathways which are activated (or repressed) under fixing conditions.

The map of the metabolic network we will use is obtained from the KEGG database¹ (Kanehisa and Goto, 2000). Metabolic networks can be represented in several ways: perhaps the most informative representation is as a directed bipartite graph, where there are two types of nodes, metabolites and enzymes. However, it is also common to represent metabolic networks as networks of metabolites (connected by edges if there is at least a reaction involving them), or networks of enzymes, where the edges represent metabolites involved in reactions catalysed by the enzymes. Since our data consists of relative concentrations of enzymes, it is this last representation that we will choose. As explained in the introduction, the edges in the network are naturally weighted: common metabolites (such as water) receive a lower weight than more biologically relevant ones.

The maximal connected component of the KEGG pathway for which we had a reasonable amount of data consisted of 422 enzymes, and had a total number of edges equal to 2024. As expected, the distribution of the number of edges per node follows the approximate power law decay characteristic of scale-free networks. The data available covered only 73 nodes out of 422 (17.3%); the distribution of the data appears to be reasonably heavy tailed (despite the size of the sample), motivating the use of exponential distributions to model the protein concentrations. We ran MMG on this data with parameter $\alpha = 1$ and $\sigma = 0.3$; the sampling procedure was the same as described above with a burn-in of 1000 and 1000 samples collected from the subsequent 5000 iterations.

The large amount of missing data clearly leads the model to produce highly uncertain predictions, i.e. in many cases the posterior probability of belonging to any of the three classes is almost identical. We therefore filtered the results of the model by considering as significant assignments only those where the probability of being in a class exceeded the probability of being in any other class by a certain specified amount (in this case fixed to be 0.1). In this way, we found a single, large up-regulated subpathway with 33 nodes, as well as several, uninteresting singletons or groups of two-three enzymes. The up-regulated subpathway is shown in Figure 2. This contains 13 highly expressed enzymes (red discs, expression ratio $>1.5$) 12 weakly up-regulated enzymes (yellow discs, expression ratio between 1 and 1.5) and eight enzymes whose expression ratio was not measured.

Some of the enzymes in this subpathway are obvious suspects: in particular, the subpathway contains the nitrogenase enzyme itself (the large red disc in the right hand of Fig. 2, labelled (a)), together with a number of other enzymes related to nitrogen metabolism at the right end of Fig. 2. On the other hand, the left end of the pathway in Fig. 2 comprises many of the enzymes involved in the pentose phosphate pathway (PPP) (labelled (b)), a component of the glycolysis machinery. This again is hardly surprising: glycolysis is the main energy producing machinery of the bacterium, and nitrogen fixation

¹http://www.genome.jp/kegg/
is a very energy consuming process for the organism. What is interesting, however, is that the model suggests a possible direct link between these two cellular functions; in other words, these two distinct processes seem to be up-regulated as a whole. The connection between the two sides of the graph goes through some enzymes that are either unobserved or only weakly up-regulated; it is therefore unlikely that it would have been recognized by an approach which does not use network structure in the determination of differential expression.

Interestingly, a recent experiment by our collaborators, which interrogated a different subset of enzymes in the Nostoc proteome in the same experimental conditions, confirms some of the predictions of our model (Ow et al., 2008).

Fig. 2. Experimental results on Nostoc proteomic data: up-regulated subpathway. Red discs indicate enzymes with an expression ratio >1.5 (size of the disc is proportional to the expression ratio); yellow discs are enzymes with expression ratio between 1 and 1.5 and black discs are enzymes with unobserved expression ratio. The left hand end of the pathway contains enzymes involved in energy production (glycolysis), the right hand end contains enzymes involved in nitrogen metabolism (in particular, the large red disc represents nitrogenase). (a) Nitrogenase; (b) Pentose Phosphate Pathway (PPP) and (c) 6-phosphogluconolactonase, determined as up-regulated in the recent experiment of Ow et al. (2008).

Fig. 3. Experimental results on Nostoc proteomic data: down-regulated subpathway. Yellow discs represent weakly down-regulated enzymes (expression ratio between 0.6 and 1), black discs are unobserved enzymes. (a) polymerase enzymes and (b) cytosine deaminase, which has been validated as a down-regulated enzyme in Ow et al. (2008).

While the biological insights provided by the up-regulated subpathway are important, the fact that a large number of enzymes are significantly up-regulated made the inferential task somewhat easier. In fact, a careful implementation of a heuristic method (Noirel et al., 2008) was able to obtain a similar subpathway which includes many of the same key enzymes.

An analysis of the down-regulated enzymes proves much harder. For a start, the range of expression ratios is much narrower (the lowest expression ratio identified by Stensjö et al. [2007] is 0.7). This means that the model’s predictions will carry a much higher uncertainty, unless several weakly down-regulated enzymes neighbour each other.

MMG returned only one down-regulated subpathway with more than three nodes. This is shown in Figure 3 and contains 12 enzymes, six of which are weakly down-regulated and six of which are unobserved. The biological significance of this subpathway is clear: the enzymes in the large cluster on the left in Figure 3 (labelled (a)) are related to polymerase, and are
Therefore involved in cell growth. This is another energy intensive cellular process, and it makes sense that the cell will scale it down in the nitrogen fixating conditions in order to give more resources to the nitrogen metabolic pathway.

Again, the new experiments (Ow et al., 2008) allowed the measurement of two of the undefined enzymes shown in Fig. 3. Notably, cytosine deaminase (the last enzyme on the right in Fig. 3, labelled (b)) turned out to be significantly down-regulated (fold change 0.45), while pyruvate kinase (part of the large cluster on the left in Fig. 3) appears to be weakly down-regulated.

4 DISCUSSION

In this article we introduced a novel Bayesian method to identify submodules of biological networks that show coherent behaviour under different conditions. This builds on the idea that using a priori knowledge of the network structure should result in an increased predictive power in the statistical analysis of ‘omic datasets. This idea has recently been applied to microarray expression data with interesting results (Rapaport et al., 2007; Wei and Li, 2007). However, the analysis of high-throughput proteomic data presents completely different challenges from the analysis of microarray data: in particular, the large proportion of missing data, as well as high levels of noise, calls for models that are capable of giving measurements of the uncertainty in their predictions, not only point estimates. Most recently, Wei and Pan (2008) have proposed a fully Bayesian model to incorporate network effects into the determination of differential expression. Their model shares many aspects of the MRF model of Wei and Li (2007) while differing subtly in the way prior distributions are modelled. This follows the same broad philosophical approach of our contribution of exploiting whenever possible Bayesian methods to account for uncertainty in data; however, the significant differences between MMG and the MRF model of Wei and Li (2007) (different priors and likelihoods) are still all present, making our contribution quite different from Wei and Pan (2008).

We develop a fully Bayesian method that allows an efficient sampling of the posterior distribution of the differential expression state of network nodes given noisy observations. An extensive validation on synthetic data leads to two conclusions: using the network structure does result in an improvement in performance, and the fully Bayesian approach we adopt is remarkably robust in the face of missing data.

Experimental results on a real high-throughput proteomic data set give results with a clear biological meaning in terms of cellular metabolic functions. It also produces a number of predictions for the expression ratio of unobserved enzymes. Some of these have been validated experimentally in a new high-throughput proteomic experiment.

While the probabilistic nature of our model ensures a principled handling of experimental noise, we did not consider the fact that measurements of different enzyme concentrations can have different associated noise levels. Indeed, the uncertainty associated with different spectra can differ greatly, and it would be desirable to keep this information during the inference stage. Propagating uncertainty in probabilistic models has been successfully done for the analysis of microarray data (Sanguinetti et al., 2005); it would be interesting to apply similar ideas in this setting.

ACKNOWLEDGEMENTS

The authors wish to thank Saw Yen Ow for many useful discussions about Nostoc metabolism, as well as granting access to unpublished data for validating our model’s predictions. J.N. and P.C.W. gratefully acknowledge support from the EPSRC under grants EP/E036252/1 and GR/s84347/01, as well as the BiomodularH2 EU-FP6 project NEST (contract number 043340).

Conflict of Interest: none declared.

REFERENCES

Stensjö, K. et al. (2007) An (TRAQ)-based quantitative analysis to elaborate the proteomic response of Nostoc sp. pcc7120 under N2 fixing conditions. J. Proteome Res., 621, 621–635.