BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–3

Probabilistic assignment of formulas to mass peaks in metabolomics experiments - Supplementary material

Simon Rogers[1]*, Richard A. Scheltema [2], Mark Girolami [1], and Rainer Breitling [2]

[1] Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, United Kingdom
[2] Groningen Bioinformatics Centre, University of Groningen, 9751 NN Haren, The Netherlands.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 INTRODUCTION

This document includes supplementary material to the bioinformatics submission. More information and code and data downloads can be found at http://www.dcs.gla.ac.uk/inference/metsamp.

2 MODEL DESCRIPTION

We observe a set of M mass peaks, with masses given by $\mathbf{x} = [x_1, \ldots, x_M]^T$. In addition, we are provided with a set of C potential formulas with theoretical mass given by $\mathbf{y} = [y_1, \ldots, y_C]^T$. For each of the M masses, we would like a posterior probability distribution over the C potential formulas. Let the indicator variable z_{cm} equal 1 if we assign mass m to forumla c and zero otherwise. We start by defining the following likelihood (or noise model) to link the observed and true masses

$$p(z_{cm} = 1|x_m, y_c, \gamma) \propto \mathcal{N}\left(\frac{x_m}{y_c}\middle| 1, \gamma^{-1}\right)$$

i.e., the ratio of the two values is distributed as a Gaussian with mean 1 and precision γ. This reflects the fact that measurement error in mass spectrometry tends to be proportional to the magnitude of the mass being measured. One of the benefits of the proposed framework is that any noise model could be used here. For example, if one wished to use a distribution with heavier tails, a laplace may be more suitable. In this work we restrict ourselves to Gaussians, determining the most appropriate distribution for this particular data is an area for future investigation.

We now introduce the the $C \times C$ matrix $\mathbf{W}$ where $w_{cc'}$ is 1 if a relationship exists between potential formulas c and c', and zero otherwise. Typically this matrix will be rather sparse.

The novelty in this work is in the definition of a prior distribution over the set of potential compounds based on this connectivity. Sets of connected compounds are considered more likely than their unconnected counterparts. Such a prior is defined over a complete set of assignments from masses to formulas - i.e. if we collect the $C \times M$ values of z_{cm} into a single matrix $\mathbf{Z}$, we are defining a prior on $p(\mathbf{Z})$ that cannot be factorised as $\prod_c \prod_m p(z_{cm})$. Enumeration of this becomes infeasible for realistic numbers of masses and formulas. Fortunately, a common technique in the area of Bayesian inference allows us to overcome this computational bottleneck. Gibbs sampling (for example, Gelman *et al.* (2004)) allows one to generate samples from a posterior distribution by repeatedly sampling values for each individual parameter (in our case each z_{cm}) conditioned on the other parameters . In this example, this means sampling an assignment for the mth mass, conditioned on the samples of the other $M-1$ masses. Defining $\mathbf{Z}^{-m}$ as the set of assignments with the mth removed, we can factorise our prior as $p(\mathbf{Z}) = p(\mathbf{Z}_{\cdot m}|\mathbf{Z}^{-m})p(\mathbf{Z}^{-m})$ where $\mathbf{Z}_{\cdot m}$ is the column corresponding to the assignment of the mth mass. When sampling $\mathbf{Z}_{\cdot m}$ we are just choosing one of the C entries to set to one and as such, the second term in this expansion $p(\mathbf{Z}^{-m})$ is just a normalising constant and can be ignored.

In our prior, we choose to make the probability of selecting formula c for mass m proportional to the connectivity of c to the other currently assigned compounds, *without* the current assignment for mass m. Mathematically, this connectivity can be computed as

$$\begin{aligned}\beta_{cm} &= \mathbf{W}_{c\cdot}\mathbf{Z}^{-m}\mathbf{1}^{M-1} \\ &= \mathbf{W}_{c\cdot}\mathbf{Z}\mathbf{1}^{M} - \mathbf{W}_{c\cdot}\mathbf{Z}_{\cdot m}\end{aligned}$$

where $\mathbf{1}^j$ is a $j \times 1$ vector of 1s and the second expression is an alternative way of viewing it, with the second term removing the effect of the current assignment of mass m.

Our conditional prior is thus defined as

$$p(z_{cm} = 1|\mathbf{Z}, \delta) = \frac{\delta + \beta_{cm}}{C\delta + \sum_{c'} \beta_{c'm}}$$

with the smoothing parameter δ ensuring that we don't get zero probability for unconnected formulas and controlling the trade-off between connectivity and mass.

Combining this prior with the likelihood, we obtain the following distribution used to generate our Gibbs samples

$$p(z_{cm} = 1|x_m, \mathbf{y}, \delta, \gamma) = \frac{\mathcal{N}\left(\frac{x_m}{y_c}\middle| 1, \gamma^{-1}\right)}{\sum_{c'} \mathcal{N}\left(\frac{x_m}{y_{c'}}\middle| 1, \gamma^{-1}\right)} \frac{\delta + \beta_{cm}}{C\delta + \sum_d \beta_{dm}}.$$

Each iteration of the sampler involved resampling the assignments for each mass using the above equation. The samples can then be averaged over a complete set of samples to obtain posterior values for $p(z_{cm} = 1)$ as required.

*to whom correspondence should be addressed

3 HYPER-PARAMETERS DETAILS

Our model includes two hyperparameters, γ (mass accuracy) and δ (connectivity smoothing). Often, the approximate mass accuracy of the mass spectrometer will be known. From this it is possible to derive a sensible prior value for γ. For example, if the mass measurements are expected to be accurate to 1ppm, and noting that the noise is defined on the ratio between the two masses, we may wish to choose a prior such that the expected value of variance, $\gamma^{-1} = (\frac{1}{3} \times 10^{-6})^2$, i.e., the standard deviation ($\sigma^{-1/2}$) is such that 99.9% of the probability mass is within $\pm$ 1ppm. A suitable prior would therefore be a Gamma distribution (the variance must be positive)

$$\mathcal{G}(\gamma|a,b) = \frac{b^a}{\Gamma(a)}\gamma^{a-1}e^{-b\gamma}$$

with parameters a and b such that the expected value $a/b = (\frac{1}{3} \times 10^{-6})^{-2}$, and a/b^2 gives some suitable variance that reflects the uncertainty in the quoted noise value. With a Gamma prior, the Gibbs sampling distribution for γ is readily available and is also Gamma distributed, as given in equation 1. This can be easily incorporated into the sampling scheme described above, by re-sampling a γ^s at each sample, using the current sample of assignments, $\mathbf{Z}^s$.

The second hyper-parameter, δ acts as a smoothing parameter for the connectivity term. In fact, we can think of it as the parameter for a Dirichlet prior on a multinomial distribution over the components of $\boldsymbol{\beta}$. Imagine the values in $\boldsymbol{\beta}$ coming from $\boldsymbol{\beta}^T\mathbf{1}$ draws from a multinomial distribution parameterized by some vector $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_c, \ldots, \theta_C]^T$. Now, we can marginalize the parameter $\boldsymbol{\theta}$ to obtain $p(\boldsymbol{\beta}|\boldsymbol{\delta})$ thus

$$\begin{aligned}
p(\boldsymbol{\beta}_m|\boldsymbol{\delta}) &= \int p(\boldsymbol{\beta}_m|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\delta})d\boldsymbol{\theta} \\
\text{where } p(\boldsymbol{\theta}|\boldsymbol{\delta}) &= \frac{\Gamma(\sum_c \delta_c)}{\prod_c \Gamma(\delta_c)} \prod_c \theta_c^{\delta_c - 1} d\boldsymbol{\theta} \\
p(\boldsymbol{\beta}_m|\boldsymbol{\delta}) &= \frac{\Gamma(\sum_c \delta_c)}{\prod_c \Gamma(\delta_c)} \int \prod_c \theta_c^{\beta_{cm}+\delta_c - 1} d\boldsymbol{\theta} \\
&= \frac{\Gamma(\sum_c \delta_c)\prod_c \Gamma(\beta_{cm}+\delta_c)}{\Gamma(\sum_c \beta_{cm}+\delta_c)\prod_c \Gamma(\delta_c)}.
\end{aligned}$$

In practice, we want the probability of the next sample from this distribution – we have already observed the samples in $\boldsymbol{\beta}_m$, so, conditioned on this and the Dirichlet prior, what are the new multinomial probabilities? Assuming that we are interested in the dth compound, and that adding one observation to the dth compound gives us $\boldsymbol{\beta}$, we have the form given in equations 2 to 5, where we have used the identity $\Gamma(x+1) = x\Gamma(x)$. If we make the assumption that $\delta_c = \delta \ \forall c$, then this prior is identical to the conditional prior described in equation (1) of the main paper. The use of a Dirichlet distribution to smooth a multinomial is common in text processing (see, for example Zhai and Lafferty (2004)), and can be thought of as adding δ pseudo-observations to each possible multinomial outcome. This overcomes the problem of no probability mass being assigned to compounds that cannot be created with the current assignment. It is possible to add an extra layer of abstraction and place a prior on δ. However, in this work we will assume it is a fixed parameter defined by the user.

4 TRYPANOSOME EXAMPLE

4.1 Data generation

The measured masses from the Trypanasoma dataset reported by Breitling *et al.* (2006) were matched to KEGG metabolites, using a mass window of 10 ppm. Matching entries were retrieved with the SOAP interface provided at the KEGG website. All unique molecular formulas were selected and stored with the mass. The automated matching was performed by the MetabolomeExplorer software (Scheltema et al., in prep.) (available from http://gbic.biol.rug.nl//supplementary/2008/MetabolomeExplorer/), using a Java implementation based on the KEGG library (`keggapi.jar`, http://www.genome.jp/kegg/soap/) utilizing the functions `search_compounds_by_mass` (retrieves all compound KEGG ids within the provided mass range) and `bget` (retrieves all information in the KEGG database for a given id). Dedicated software was written to interpret the results from the bget function. This process took several hours, mainly due to delays accessing KEGG. The full annotation file is available for download at http://www.dcs.gla.ac.uk/inference/metsamp/downloads/tryp10ppm.xls.

4.2 Transformations

The full list of chemical transformations can be downloaded from http://www.dcs.gla.ac.uk/inference/metsamp.

4.3 Connectivity matrix

The connectivity matrix can be seen in figure 1. Of the $\sim$ 80000 possible connections (bearing in mind that connections are not directional), only 343 are present (0.4%) suggesting an efficient sparse implementation would be possible.

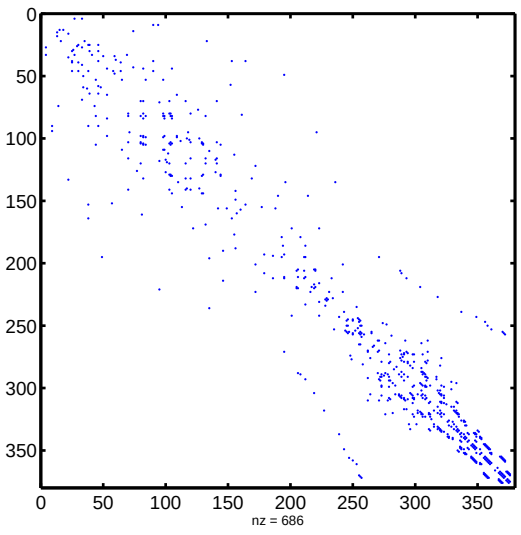

Fig. 1. Connectivity matrix for the Trypanosome example.

4.4 Additional results

in Figure 2 we plot the probability of the most likely assignment from the sampler against that from just using the likelihood (i.e.,

$$p(\gamma|a,b,\mathbf{x},\mathbf{y},\mathbf{Z}) = \mathcal{G}\left(\gamma \left| a + \frac{M}{2}\right., b + \frac{1}{2}\sum_{m=1}^{M}\sum_{c=1}^{C} z_{cm}\left(\frac{x_m}{y_c} - 1\right)^2\right) \tag{1}$$

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\beta}_m,\boldsymbol{\delta}) &= \frac{p(\boldsymbol{\beta},\boldsymbol{\beta}_m|\boldsymbol{\delta})}{p(\boldsymbol{\beta}_m|\boldsymbol{\delta})} && (2)\\
&= \left[\frac{\Gamma(\sum_c \delta_c)\prod_c \Gamma(\beta_{cm}+\delta_c)}{\Gamma(\sum_c \beta_{cm}+\delta_c)\prod_c \Gamma(\delta_c)}\right]^{-1} \frac{\Gamma(\sum_c \delta_c)\Gamma(\beta_{dm}+1+\delta_d)\prod_{c\neq d}\Gamma(\beta_{cm}+\delta_c)}{\Gamma(1+\sum_c \beta_{cm}+\delta_c)\prod_c \Gamma(\delta_c)} && (3)\\
&= \frac{\Gamma(\sum_c \beta_{cm}+\delta_c)}{\Gamma(1+\sum_c \beta_{cm}+\delta_c)} \times \frac{\Gamma(\beta_{dm}+1+\delta_d)\prod_{c\neq d}\Gamma(\beta_{cm}+\delta_c)}{\prod_c \Gamma(\beta_{cm}+\delta_c)} && (4)\\
&= \frac{\beta_{dm}+\delta_d}{\sum_c \beta_{cm}+\delta_c} && (5)
\end{aligned}$$

mass alone). We can see that of the 63 peaks where the probability changes (the remaining 383 peaks had equal probability under both models), the majority (42) show increased probability under the full model, corresponding to an increased confidence in the assignments.

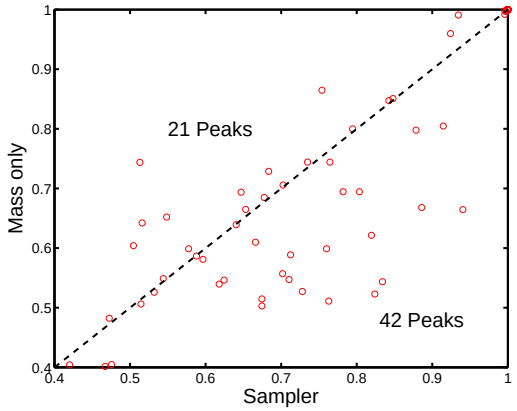

Fig. 2. Assignment probabilities based on the maximum of the likelihood (mass alone) and the sampler (mass and connectivity). Of the 63 where the probability changes, in 42 cases (67%) it is increased with the sampler.

4.5 Timing information

The Trypanosome data-set consists of 446 measured masses and 379 potential compounds. The transformation matrix was created by exhaustively comparing each pair of potential compounds to the list of allowable transformations. This step was implemented in Matlab and took 19.8 seconds. The sampler was then run for 3000 burn in samples and then 2000 further samples from which probabilities were computed. This took 225.3 seconds (just under 4 minutes). It is worth noting that the implementation is not particularly efficient (unnecessarily compares each mass to all compounds, most of which will have a likelihood of effectively 0 and does not use the sparse properties of the connectivity matrix). The time taken to generate the samples compares very favorably with the time taken to generate the original list of potential compounds (several hours). It is therefore practical to quickly compare several settings of the hyper-parameters.

REFERENCES

Breitling, R., Ritchie, S., Goodenowe, D., Stewart, M. L., and Barrett, M. P. (2006). Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics*, **2**, 155–164.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*.