MIXTURE MODEL CLUSTERING FOR PEAK FILTERING IN METABOLOMICS

Simon Rogers[1], *Rónán Daly*[2] *and Rainer Breitling*[2,3]

[1]School of Computing Science, University of Glasgow, UK
[2]Institute of Molecular, Cell and Systems Biology, University of Glasgow, UK
[3]Groningen Bioinformatics Centre, University of Groningen, Netherlands
simon.rogers@glasgow.ac.uk, ronan.daly@glasgow.ac.uk, rainer.breitling@glasgow.ac.uk

ABSTRACT

In recent years, the use of liquid chromatography coupled to mass spectrometry has enabled the high-throughput profiling of the metabolic composition of biological samples. However, the large amount of data obtained is often difficult to analyse. This paper focuses on a particular problem, that of detecting and potentially removing derivative peaks of a substance of interest. A mixture model for clustering peaks based on chromatographic peak shape correlation is presented, and comparison of this model to the behaviour of a leading mass spectrometry analysis tool is presented. Based on the results, the mixture model is shown to have better overall performance characteristics.

1. INTRODUCTION

Recent studies have shown that changes in an organism's metabolome composition are more closely correlated with phenotypic variation than changes in either the transcriptome or the proteome [1], motivating the use of high-throughput metabolomic assays for a wide variety of biomedical applications. The most popular method for analysing the metabolome is mass spectrometry (MS) coupled to a separation phase resulting in data consisting of a set of chromatographic peaks characterised by their mass and retention time. Identifying the peaks in these spectra makes metabolomic analysis more challenging than, for example, microarray analysis, where the structure of the mRNA makes it possible to build a probe to which the molecule of interest will bind with high specificity. Whilst, in some cases, matching a measured mass to a database of known masses will result in a correct identification, sometimes several matches might be found within a window defined by the mass accuracy of the equipment [2].

The problem is made more complicated by the fact that an experimental sample containing a few dozen metabolites will result in the production of several hundred peaks. The large number of additional peaks can come from various sources, including isotopes, adducts, molecular fragments, and multiply charged ions [3]. As some of these derivative peaks will have masses very similar to known metabolites, filtering them out is crucial to avoid the overwhelming number of false identifications that would be produced if they were all compared to a database of known masses.

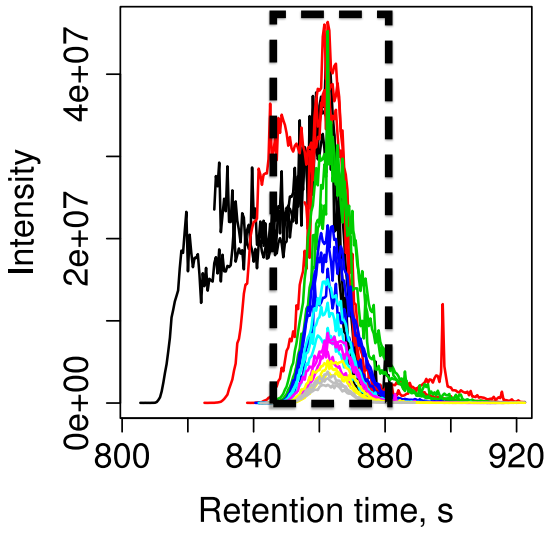


Figure 1. A chromatogram, showing peak intensity as a function of retention time in the separation phase. Co-eluting peaks often have a similar shape; the peaks tightly clustered around 863 s (boxed) are derived from the same metabolite.

Fortunately, peaks that are derived from the same metabolite share characteristics that make it possible to group them together. In particular, they will elute from the separation phase at the same time (their retention times will be very similar) and their peaks will have similar shapes [3], as can be seen in Figure 1. Here, we introduce a mixture model for clustering peaks based on the correlation between their peak shapes. The remainder is organised as follows: In the next section we briefly introduce mzMatch, one of the most popular open-source metabolomics analysis pipelines, as this is the system against which we compare our proposed approach. In Section 3 we introduce our model and in Section 4 demonstrate its performance on some standard metabolomic datasets. Finally, in Section 5 we present a brief discussion and conclusions.

2. MZMATCH

mzMatch is a popular open-source metabolomic data analysis pipeline that enables researchers to perform end-to-end analysis of diverse biological datasets [4]. Whilst the complete toolkit performs many functions from peak extraction

from raw data files, to metabolite identification, for the purposes of this paper we are focusing on the peak-derivative detection phase, also known as peak clustering.

Currently, peak clustering in mzMatch is done via a simple greedy algorithm. The following algorithm describes mzMatch's behaviour.

Algorithm 1 mzMatch clustering algorithm

while there are unclustered peaks **do**

- Find the most intense unclustered peak.
- Create a new cluster based on this peak.
- Find other unclustered peaks whose Pearson correlation over retention time with this peak is greater than a pre-defined threshold (0.75) and add these to the cluster.

end while

There are several potential issues with this approach. Firstly, the algorithm is greedy – whilst it is likely that the true metabolite peaks will have high intensity [3], repeatedly picking the most intense remaining peak and building a cluster around it results in an algorithm that could be highly sensitive to small changes in intensity values. Secondly, correlation is only computed between the initial peak and others. Most inter-peak correlation values are not used to produce the clustering – information is being thrown away.

3. MIXTURE MODEL FOR CORRELATIONS

Our observations consist of a symmetric $N \times N$ matrix of peak shape correlations (Pearson) between the N observed peaks, $\mathbf{Q}$, the n, mth element of which (the correlation between peaks n and m) is denoted by q_{nm}. Note that it would be possible to start from the actual peaks themselves rather than correlation values, and this is an interesting avenue for future work. Our aim is to partition the peaks into K clusters, each of which will potentially correspond to one metabolite and its derivatives. We use a set of binary indicator variables, z_{nk} to indicate cluster membership i.e., $z_{nk} = 1$ if peak n is assigned to cluster k. Collectively, these indicators are denoted as $\mathbf{Z}$ and our task is therefore to infer $\mathbf{Z}$ from $\mathbf{Q}$. It will be convenient to define a second set of indicators: $\epsilon_{nm} = \sum_{k=1}^{K} z_{nk} z_{mk}$, i.e. $\epsilon_{nm} = 1$ if peaks n and m are in the same cluster, and zero otherwise.

We assume that the correlation values were generated by a mixture model with two components, one describing correlations between peaks within the same cluster, and one describing correlations between peaks in different clusters. Inspection of typical outputs from mzMatch [4] led us to choose an exponential-type distribution for the former, and a Gaussian for the latter. In addition, to allow the correlation matrices to be produced at low computational cost, peak correlations are only calculated for peaks that co-elute within a particular retention time window. Hence, we additionally observe a binary $N \times N$ matrix $\mathbf{R}$, the elements of which, r_{nm}, equal 1 if the correlation between peaks n and m is observed and 0 otherwise. We therefore assume the following

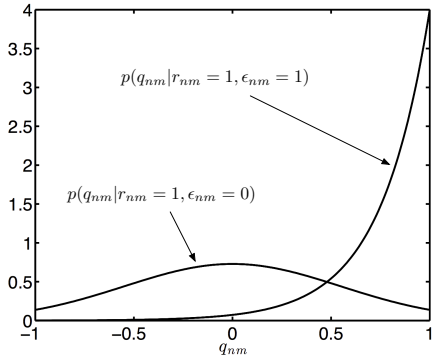


Figure 2. Example of the generative densities (assuming the value is observed, $r_{nm} = 1$) where $\lambda = 4, \mu = 0, \sigma^2 = 0.3$.

generative distributions:

$$p(q_{nm}, r_{nm} = 1|\epsilon_{nm} = 1) = (1 - p_1)\, \lambda e^{-\lambda(1-q_{nm})} \quad (1)$$

$$p(q_{nm}, r_{nm} = 0|\epsilon_{nm} = 1) = p_1 \delta\left(q_{nm}\right) \quad (2)$$

$$p(q_{nm}, r_{nm} = 1|\epsilon_{nm} = 0) = (1 - p_0)\, \mathcal{N}\left(q_{nm}|\mu, \sigma^2\right) \quad (3)$$

$$p(q_{nm}, r_{nm} = 0|\epsilon_{nm} = 0) = p_0 \delta\left(q_{nm}\right), \quad (4)$$

where conditioning on $p_0, p_1, \lambda, \mu, \sigma^2$ is omitted. An example is given in Figure 2. The density of q_{nm} conditioned on a particular value of r_{nm}, e.g.

$$p(q_{nm}|r_{nm}, \epsilon_{nm} = 1) = r_{nm} \lambda e^{-\lambda(1-q_{nm})} + (1 - r_{nm})\, \delta\left(q_{nm}\right),$$

is similar to the 'spike-and-slab' distributions used in DNA microarray analysis (see, e.g., [5]). The likelihood of the complete set of observed correlations is given by:

$$\mathcal{L}(\mathbf{Q}, \mathbf{R}|\mathbf{Z}) = \prod_{n=1}^{N-1} \prod_{m=n+1}^{N} p(q_{nm}, r_{nm}|\epsilon_{nm} = 1)^{\epsilon_{nm}} \times p(q_{nm}, r_{nm}|\epsilon_{nm} = 0)^{1-\epsilon_{nm}},$$

where the various components in the product are given by Equations 1 to 4. To complete our model specification, we require a prior density over the membership of cluster k, i.e. $p(z_{nk} = 1)$. To avoid having to *a priori* specify the number of components, we use a Dirichlet Process (DP) prior, with concentration parameter α (see, e.g., [6] for more details).

3.1. Inference

The distribution of interest here is the posterior density over cluster assignments, $p(\mathbf{Z}|\mathbf{Q}, \mathbf{R})$. Analytical inference is not tractable, but it is possible to generate samples using a Gibbs sampling scheme. At each stage, we re-sample the assignment for one peak conditioned on the current assignments of all other peaks. Using $\mathbf{Z}^{-n}$ to denote all of the assignments except that for peak n, the conditional distributions required are given by:

$$P(z_{nk} = 1|\ldots) \propto s_k \mathcal{L}(\mathbf{Q}, \mathbf{R}|\mathbf{Z}^{-n}, z_{nk} = 1),$$

Table 1. Comparison of mixture model to mzMatch

Sample	Ionisation Mode	# Peaks	mzMatch				Mixture Model			
			TPR	FPR	BA	# Clusters	TPR	FPR	BA	# Clusters
Standard 1	Negative	3664	0.548	0.211	0.668	799	0.466	0.060	0.703	250
Standard 1	Positive	6291	0.583	0.220	0.682	1403	0.444	0.061	0.692	409
Standard 2	Negative	2386	0.571	0.247	0.662	621	0.446	0.071	0.688	206
Standard 2	Positive	6883	0.597	0.340	0.628	2370	0.5	0.076	0.711	565
Standard 3	Negative	999	0.632	0.418	0.607	421	0.474	0.212	0.631	216
Standard 3	Positive	2316	0.632	0.497	0.567	1151	0.474	0.187	0.643	440

for the current cluster k that has s_k members (not including peak n), and:

$$P(z_{nk*} = 1|\ldots) \propto \alpha\mathcal{L}(\mathbf{Q}, \mathbf{R}|\mathbf{Z}^{-n}, \epsilon_{nm} = 0\,\forall\, m),$$

for a new cluster, k^*, where $\epsilon_{nm} = 0\,\forall\, m$ describes the fact that n is in its own cluster and hence not in the same cluster as any other peak. As we are interested in a single set of assignments to compare against the clustering produced by mzMatch, we typically run the sampler for a fixed number of iterations and keep the sample with the highest posterior value.

4. TESTING AND RESULTS

In order to compare the clusterings produced by the mixture model against those produced by mzMatch, both algorithms were ran against the data files produced from a set of LC-MS experiments. In this case, the data are based on standard samples used to calibrate chromatographic columns. Each sample consisted of a mixture of known compounds, with known mass and expected retention time. Three samples were used, comprising 104, 96 and 40 compounds respectively, with runs in both positive and negative mode, for a total of six data sets. The mixture model parameters were set at $\mu = 0, \sigma^2 = 0.4, \lambda = 8, p_1 = 0.001, p_0 = 0.97, \alpha = 1$. This is our initial choice, and it is unlikely that these will be optimal. Note also that the Gaussian density will provide some probability mass for values outside the feasible correlation range. Optimising the form and parameters of these densities based on the known constituents of standard samples is ongoing work.

After the clusterings were produced, each of the peaks in each of the clusters was labelled as a base peak or a derivative of a base peak, using the algorithm given by mzMatch (the base peak is defined as the most intense peak in a mass spectrogram or in a cluster). An identification was also attempted on each peak, using the known masses of the compounds in the sample and the mzMatch identification algorithm. To assess the performance characteristics of the clusterings, each peak was compared against the known masses of the compounds and assigned a label. The labels were:

True Positive (TP) If the peak is a base peak and identified as a compound that is known to be present in the sample.

False Positive (FP) If the peak is a base peak and does not correspond to an expected compound.

True Negative (TN) If the peak is not a base peak and not identified as an expected compound.

False Negative (FN) If the peak is not a base peak, but matches an expected compound.

The following statistics were calculated:

True Positive Rate: $TPR = \frac{TP}{TP+FN}$.

False Positive Rate: $FPR = \frac{FP}{FP+TN}$.

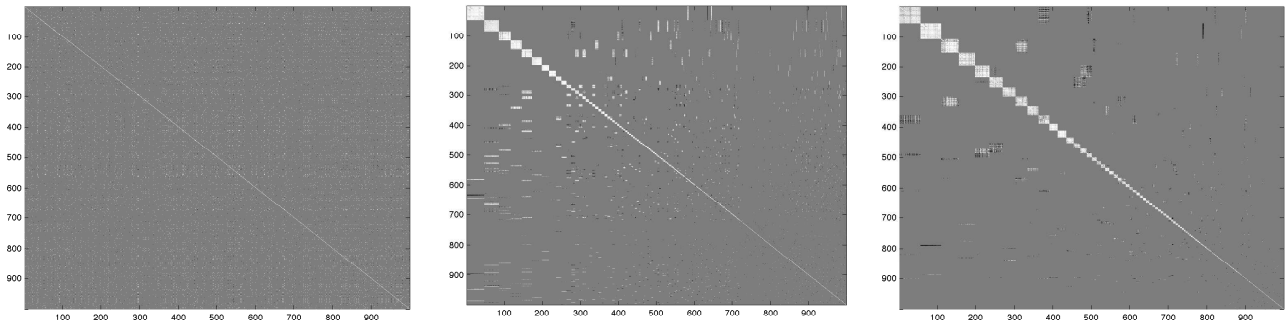
Balanced Accuracy: $BA = \frac{0.5*TP}{TP+FN} + \frac{0.5*TN}{TN+FP}$.

The results of the experiments are summarised in Table 1. The first thing to notice is that mzMatch produces many more clusters than the mixture model. This gives rise to higher TPR (if every peak were assigned to its own cluster, we would observe a TPR of 1). The smaller number of clusters given by the mixture model causes a considerable improvement in the FPR. This improvement outweighs the lower TPR, as can be seen by the mixture model having a higher BA in all samples. These results suggest that the mixture model represents a very promising alternative to the clustering currently used by mzMatch.

In addition to these results, Figure 3 shows the correlation matrix $\mathbf{Q}$ for Standard 3 in Negative mode unordered (a) and then ordered according to the mzMatch clustering (b) and the mixture model clustering (c). The larger clusters produced by the mixture model are clearly visible, as is the marked reduction in high correlation values between peaks in different clusters (white colouring off the diagonal).

5. DISCUSSION AND CONCLUSION

The results presented in the previous section suggest that the mixture model approach can cluster, and subsequently filter, peaks more effectively than the algorithm used in mzMatch. Whilst mzMatch finds more compounds, it does so as the result of degenerate behaviour when clustering peaks



(a) Unordered **Q** matrix

(b) **Q** ordered by mzMatch assignments (clusters ordered by size)

(c) **Q** ordered by mixture model assignments (clusters ordered by size)

Figure 3. Various orderings of the correlation matrix for Standard 3 Negative data. White corresponds to high positive correlation, black to high negative correlation. Unobserved values are given the value zero and are shown as grey.

together, and hence produces many spurious identifications. This behaviour can be understood as a consequence of the greedy behaviour and fixed threshold of mzMatch; by committing early to a particular clustering, and by not being able to update clusters in the light of new information, peaks that might naturally be clustered with others are left out if they do not correlate to the base peak. With no other peaks to correlate to (they are all in the original cluster) they are left by themselves in singleton clusters.

The clustering produced by the mixture model is still far from perfect. One way of further improving the filtering process is through the incorporation of additional information into the clustering process. For example, it is straightforward to extend the model to handle technical and/or biological replicates by assuming a single clustering across all replicates and taking the product of the likelihoods of the observed data in each clustering conditioned on the assignments. Alternatively, mass is currently not used in the clustering at all. Many peaks deriving from the same metabolite should have explainable mass differences – development of a likelihood function that can take this information into account is a promising avenue for future work [7].

Taking a longer view, peak filtering is just one step in the complete analysis pipeline. Often, one will be interested in identifying the metabolites and performing differential analysis across time, or different experimental conditions. Rather than extracting a single clustering from the Gibbs sampler, one could combine this filtering model with other probabilistic models that assign peaks to metabolites [7] or perform differential analysis [8] and thus propagate any uncertainty in the clustering stage through the analysis in a manner similar to that done for microarray data in [9].

6. ACKNOWLEDGEMENTS

Many thanks to Andris Jankevics and Stefan Weidt for very useful discussions and for supplying the data.

7. REFERENCES

[1] J. Fu, J. J. B. Keurentjes, H. Bouwmeester, T. America, F. W. A. Verstappen, J. L. Ward, M. H. Beale, R. C. H. de Vos, M. Dijkstra, R. A. Scheltema, F. Johannes, M. Koornneef, D. Vreugdenhil, R. Breitling, and R. C. Jansen, "System-wide molecular evidence for phenotypic buffering in *Arabidopsis*," *Nature Genetics*, vol. 41, pp. 166–167, 2009.

[2] T. Kind and O. Fiehn, "Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm," *BMC Bioinformatics*, vol. 7, pp. 234, 2006.

[3] R. A. Scheltema, S. Decuypere, J.-C. Dujardin, D. G. Watson, R. C. Jansen, and R. Breitling, "A simple data reduction method for high resolution LC-MS data in metabolomics," *Bioanalysis*, vol. 1, no. 9, pp. 1551–1557, 2009.

[4] R. A. Scheltema, A. Jankevics, R. C. Jansen, M. A. Swertz, and R. Breitling, "PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis," *Analytical Chemistry*, vol. 83, no. 7, pp. 2786–2793, 2011.

[5] C. Carvalho, J. Chang, J. Lucas, J. Nevins, W. Wang, and M. West, "High-dimensional sparse factor modelling: Applications in gene expression genomics," *J Am Stat Assoc*, vol. 103, no. 484, pp. 1438–1456, 2008.

[6] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems 12*. 2000, pp. 554–560, MIT Press.

[7] S. Rogers, R. A. Scheltema, M. Girolami, and R. Breitling, "Probabilistic assignment of formulas to mass peaks in metabolomics experiments," *Bioinformatics*, vol. 25, no. 4, pp. 512–518, 2009.

[8] I. Huopaniemi, T. Suvitaival, J. Nikkila, M. Oresic, and S. Kaski, "Two-way analysis of high-dimensional collinear data," *Data Mining and Knowledge Discovery*, vol. 19, no. 2, pp. 261–276, 2009.

[9] R. Pearson, X. Liu, G. Sanguinetti, M. Milo, N. Lawrence, and N. Rattray, "PUMA: A Bioconductor package for propogating uncertainity in microarray analysis," *BMC Bioinformatics*, vol. 10, 2009.