BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 11 2006, pages 1391–1396
doi:10.1093/bioinformatics/btl085

*Data and text mining*

# Progressive peak clustering in GC-MS Metabolomic experiments applied to *Leishmania* parasites

David P. De Souza[1,2], Eleanor C. Saunders[1,2], Malcolm J. McConville[1,2] and Vladimir A. Likić[2,*]

[1]Department of Biochemistry and Molecular Biology and [2]The Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, 3010, Australia

Received on January 4, 2006; revised on February 16, 2006; accepted on March 4, 2006
Advance Access publication March 9, 2006
Associate Editor: Martin Bishop

**ABSTRACT**

**Motivation:** A common problem in the emerging field of metabolomics is the consolidation of signal lists derived from metabolic profiling of different cell/tissue/fluid states where a number of replicate experiments was collected on each state.

**Results:** We describe an approach for the consolidation of peak lists based on hierarchical clustering, first within each set of replicate experiments and then between the sets of replicate experiments. The problems of finding the dendrogram tree cutoff which gives the optimal number of peak clusters and the effect of different clustering methods were addressed. When applied to gas chromatography-mass spectrometry metabolic profiling data acquired on *Leishmania mexicana*, this approach resulted in robust data matrices which completely separated the wild-type and two mutant parasite lines based on their metabolic profile.

**Contact:** vlikic@unimelb.edu.au

## 1 INTRODUCTION

Metabolomics is an emerging tool of functional genomics that is increasingly being used to identify new protein functions and to model whole cell metabolism (Fernie *et al*., 2004; Fiehn, 2002; Goodacre *et al*., 2004; Sumner *et al*., 2003). Two approaches are commonly associated with metabolomics: metabolic fingerprinting and metabolic profiling (Fiehn, 2002). Metabolic fingerprinting refers to the analysis of patterns in the molecular response profiles (detected by NMR, MS or other spectroscopic techniques), without an attempt to resolve individual analytes. In contrast, metabolic profiling aims to resolve, identify and quantitate individual analytes. In non-targeted profiling all metabolites resolved by a particular analytical technique are quantitated, even those of unknown chemical structure.

Metabolomic studies frequently utilize gas chromatography mass spectrometry (GS-MS) (Fiehn *et al*., 2000; Roessner *et al*., 2001; Urbanczyk-Wochniak *et al*., 2003), liquid chromatography mass spectrometry (LC-MS) (Allen *et al*., 2003; Tolstikov *et al*., 2003), capillary electrophoresis mass spectrometry (CE-MS) (Guillo *et al*., 2004; Sato *et al*., 2004), direct infusion electrospray ionization mass spectrometry (ESI-MS) (Castrillo *et al*., 2003; Goodacre *et al*., 2002) and NMR (Choi *et al*., 2004; Raamsdonk *et al*., 2001). Irrespective of the technique, the analytical instrumentation produces signal which consists of informative peaks embedded in a continuum of background noise. Signal peaks are associated with individual analytes (metabolites) and provide a quantitative measure of the concentration of individual analytes.

A common problem in metabolic profiling is the correlation of signal peaks from two or more cell/tissue states (i.e. wild-type versus mutant, healthy versus diseased, etc.). For example, if two cell/tissue states $A$ and $B$ are examined, with observed signal peaks $A_1, A_2, \ldots, A_N$ and $B_1, B_2, \ldots, B_M$, it is important to establish the correspondence between these signals, i.e. which signals in $A$ and $B$ refer to the same analyte or metabolite. Furthermore, some signals in $A$ may not have the corresponding signal in $B$ because of the metabolites produced in one state but not the other. The correspondence between $A_1, A_2, \ldots, A_N$ and $B_1, B_2, \ldots, B_M$ leads directly to the data matrix, whose rows represent individual experiments and columns represent unique analytes observed in the two sets of experiments. The generalization to more than two experiments is straightforward.

Metabolic profiling experiments are relatively rapid and inexpensive, with typically multiple replicates are recorded for each cell/sample state (Allen *et al*., 2003; Fiehn *et al*., 2000; Urbanczyk-Wochniak *et al*., 2003). Multiple replicate experiments facilitate robust statistical analysis, and have also been used to explore the inherent biological variability in metabolomic studies (Fiehn *et al*., 2000; Sumner *et al*., 2003). If replicate experiments are to be explicitly included in the data matrix, the correspondence between signals observed in different replicates must be established. By taking the previous example of two cell states $A$ and $B$, and assuming that $P$ replicate experiments of $A$ and $Q$ replicate experiments of $B$ were performed, the two sets of experiments may be represented by a series of signals $A_{11}, A_{12}, \ldots, A_{1N}, A_{21}, A_{22}, \ldots, A_{2N}, A_{P1}, A_{P2}, \ldots, A_{PN}$ and $B_{11}, B_{12}, \ldots, B_{1M}, B_{21}, B_{22}, \ldots, B_{2M}, B_{Q1}, B_{Q2}, \ldots, B_{QM}$, respectively. The resulting data matrix will have $P + Q$ rows. The number of columns will depend on the correspondence between signals observed in the two sets of experiments, but cannot exceed $N+M$.

The problem of peak consolidation in multiple experiments has been addressed recently with the program MSFACTs (Duran *et al*., 2003). Here we further explore the idea of peak clustering for the consolidation of signal lists. Specifically, we propose a two-step hierarchical clustering of peak signals, first within each set of replicate experiments and then between the sets of replicate experiments. We apply this approach to gas chromatography-mass spectrometry (GC-MS) metabolic studies of *Leishmania mexicana*.

*To whom correspondence should be addressed.

 

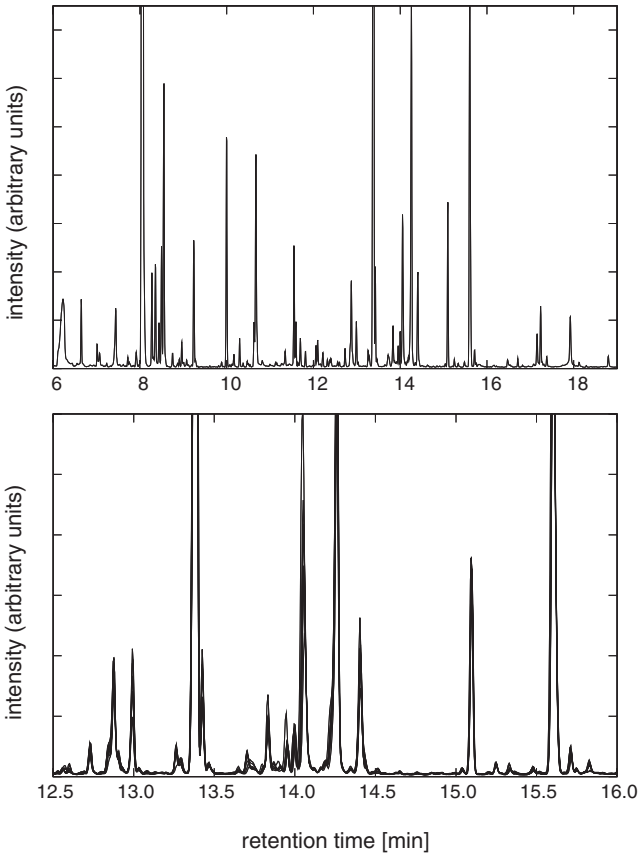


**Fig. 1.** An example of total ion chromatogram (TIC) for wild-type *L.mexicana* (upper panel). The TICs for eight replicate experiments overlaid in the region of 12.5–16.0 min are shown in the lower panel.

*Leishmania* is a sandfly-transmitted parasite endemic throughout the tropic and subtropics which infects around 12 million people worldwide (Davis *et al*., 2004). The proposed approach resulted in robust data matrices which completely separated the wild-type and two mutant parasite lines based on their metabolic phenotypes.

## 2 METHODS

Metabolic profiles of wild-type and two mutants of *L.mexicana* were analyzed. The wild type strain (*wt*), and two mutant strains deficient in the three functional glucose transporters (Δ*gt*) or the enzyme phosphomannose isomerase (Δ*pmi*) were derived from the same parental strain M379 and displayed very similar growth rates in rich medium (Burchmore *et al*., 2003; Garami and Ilg, 2001). Parasites were cultivated in RPMI medium containing 10% fetal calf serum and harvested at day 6 in stationary growth phase. Parasite metabolism was quenched by immersion of the culture flask in ethanol–dry ice bath, and chilled parasites harvested by centrifugation of 1 ml culture medium in a microfuge (15 000 rpm, 20 s, 0°C). Metabolites were extracted with chloroform:methanol:water (1:3:1 v/v) and polar and apolar metabolites were separated by phase partitioning. Following derivatization (methoximation and trimethylsilylation, TMS) the polar metabolite extracts were analyzed by GC-MS (Roessner, *et al*., 2000). For each genotype eight replicate experiments were prepared. One Δ*pmi* replicate experiment was contaminated and was not included in the analysis. The final data set consisted of eight replicate experiments for *wt* and Δ*gt* mutant each, and seven replicate experiments for the Δ*pmi* mutant.

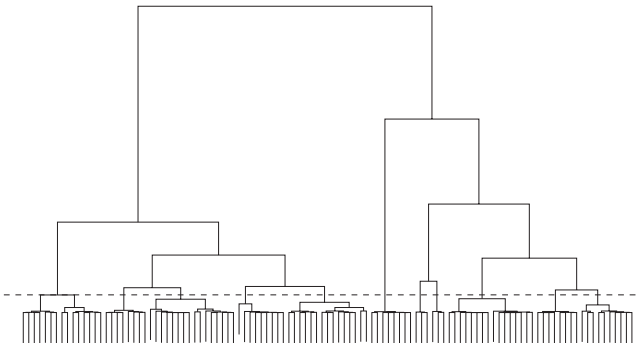

**Fig. 2.** An example of a complete dendrogram tree created after peak clustering. This figure shows a dendrogram tree created from peaks found in the region of 13.5–15.2 min of wild-type *L.mexicana* experiments. The cumulative list of peaks from eight replicate experiments contained a total of 111 peaks. Peaks were clustered by the retention times with complete linkage method. A hypothetical (non-optimal) cutoff is shown by the dashed line.

The total ion-chromatogram (TIC, Fig. 1) was integrated in ChemStation (MSD Chemstation D.01.02.16, Agilent Technologies) by using the default integrator. Resulting peak tables were exported to external files for further processing.

## 3 RESULTS

Initially the TIC (Fig. 1) was examined visually for each experiment, and peak lists were edited to mark the reference peak and uninformative peaks originating from the derivatizing agent (TMS). Subsequently, the areas of all peaks were normalized with the area of the reference peak, and the reference peak was removed from each peak list.

### 3.1 Peak clustering within each set of replicate experiments

For the purpose of clustering all peaks from a single set of replicate experiments were pooled together, and hierarchical clustering was performed to create a complete dendrogram tree. The distance between two peaks was defined as the absolute difference between the retention times recorded at peak apexes. In general, the details of the dendrogram tree depend on the clustering method used to generate the tree. Furthermore, once the method has been chosen and the dendrogram tree created, cutting the tree at any definite height would produce a certain number of peak clusters (Fig. 2). This suggests that at least two questions must be answered for successful application of this approach to peak clustering: (1) how to choose a dendrogram tree cutoff to obtain the optimal number of peak clusters and (2) how the results depend on the clustering method.

*3.1.1 The dendrogram tree cutoff* If the average number of peaks in each replicate experiment is $N_{\text{aver}}$ we expect to observe approximately $N_{\text{aver}}$ distinct chemical compounds in a set of replicate experiments. Thus a reasonable initial assumption would be that the dendrogram tree should be cut at a height to yield $N_{\text{aver}}$ peak clusters. To investigate how this would perform in practice we focused on the region between the 12.5 and 16.0 min in the *wt* set of experiments (Figure 1, lower panel). In this region ChemStation peak finding algorithm identified between 31 and 36 peaks (depending on the specific replicate experiment), with $N_{\text{aver}} = 32.5$. To create an 'ideal' table of peak correspondence, manual analysis of the region 12.5–16.0 min was carried out across all eight replicate

experiments which included verification of mass-spectra at peak apexes. Manual analysis revealed 41 unique analytes in the region 12.5–16.0 min. This result suggested that the optimal number of peak clusters is greater than $N_{aver}$. Inspection of individual peak tables revealed that this effect occurred because some peaks of low intensity were detected in some, but not all replicate experiments. Moreover, different peaks fell into this category in different replicate experiments. Although this effect was strictly verified only on the data subset analyzed here, we expect it to apply more generally.

*3.1.2 Errors in automated peak clustering* Two types of errors can occur in an automated peak clustering procedure: (1) given a true peak group across the set of replicate experiments, one or more peaks may be assigned to a different group and (2) a true peak group may be split into two or more groups. The first case is likely to result in one or more 'peak collisions', the effect whereby more than one peak from the same experiment is joined into a single cluster (Duran *et al*., 2003). The second case involves creation of one or more artificial peak groups. Provided that some reasonable clustering is performed errors of type (1) will predominate when the final number of clusters is too small relative to the optimal number of clusters, while errors of type (2) will predominate when the final number of clusters is too large.

*3.1.3 The effect of the clustering method* Several well-established clustering methods were tested, including single linkage, complete linkage and centroid methods. The full dendrogram tree was created by each clustering method, which was then cut to yield the final peak clusters. To accommodate for the effect described above (the optimal number of peak clusters being larger than $N_{aver}$) the tree was cut to yield $N_{aver} + FN_{aver}$ peak clusters, where $F$ was an 'expansion factor' set empirically. The performance of automated peak clustering was tested by comparing the results with the manually generated table of peak clusters (the region 12.5–16.0 min, Fig. 1). Given the choice of $F$ in the range 0–0.30 the total number of clusters was recorded, the number of peak collisions, and the number of true clusters artificially split. The last two parameters were obtained by comparing the clusters obtained via hierarchical clustering with the manually generated table of clusters. Figure 3 shows these three values as a function of $F$ when single linkage, complete linkage and centroid clustering methods were used.

*3.1.4 Accuracy peak clusters* Analysis of the accuracy of final clusters was carried out for the complete clustering method and expansion factors of $F = 0$, 0.1 and 0.2 by comparing the clustering output with the manually generated table of clusters. Figure 4 shows that as the expansion factor was increased from 0 to 0.2, peak collisions were reduced while the accuracy of final clusters was preserved. Specifically, automated clustering with the expansion factor of 0.2 produced both a small number of peak collisions and accurate clusters. The final table of clusters was comparable in accuracy with the manually generated table produced by inspecting mass-spectra.

## 3.2 Peak clustering between sets of replicate experiments

In order to correlate peaks between different cell states or genotypes, peak clusters from each set of replicate experiments were pooled together and clustered. In this separate clustering step

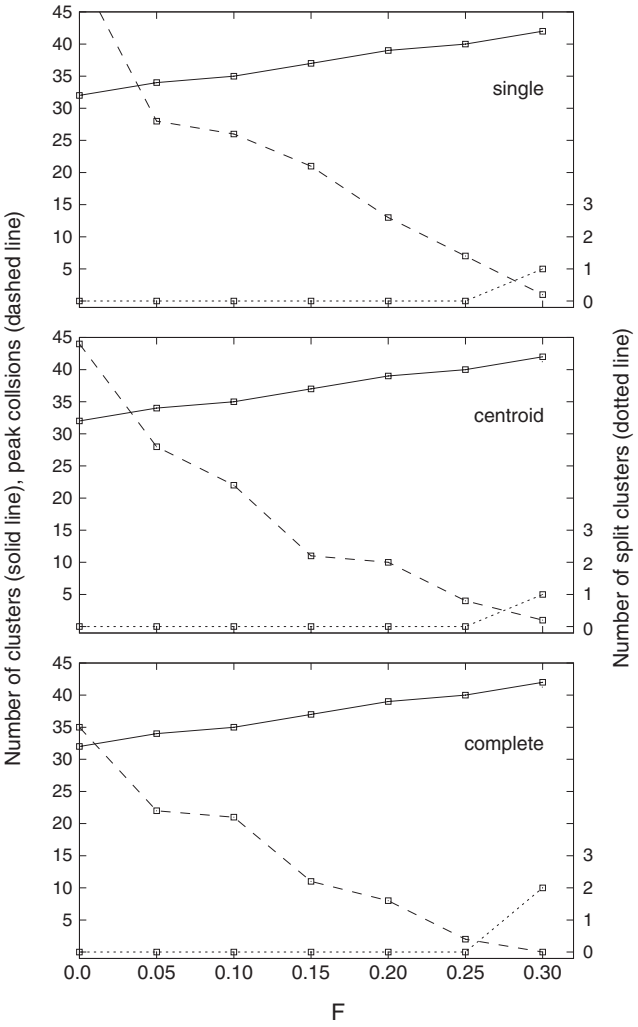


**Fig. 3.** The effect of the expansion factor ($F$) on the total number of clusters, number of peak collisions and number of split clusters when the single linkage, complete linkage, and centroid clustering methods were applied. The analysis refers to the region 12.5–16.0 min in the *wt* set of experiments, which contained eight replicate experiments (overlaid in Fig. 1, lower panel). The true number of unique compounds observed in this region was 41, as determined by manual analysis of mass spectra at peak apexes in eight experimental replicates.

the objects to cluster were peak clusters, and the outputs were clusters of peak clusters (i.e. super-clusters, Fig. 5). Super-clusters provided correlations between peak clusters, and therefore peaks observed in different sets of replicate experiments. Super-clusters correspond directly to the columns of the data matrix desired in the output (Fig. 5).

In order to perform hierarchical clustering of peak clusters one must define the distance measure. We chose the average distance between all peaks from two clusters to represent the distance between the two clusters, where the distance between the two peaks was defined above.

As in simple clustering of peaks, clustering of peak clusters produces a complete dendrogram tree. Cutting this tree at certain height would result in a set of super-clusters. If the dendrogram tree

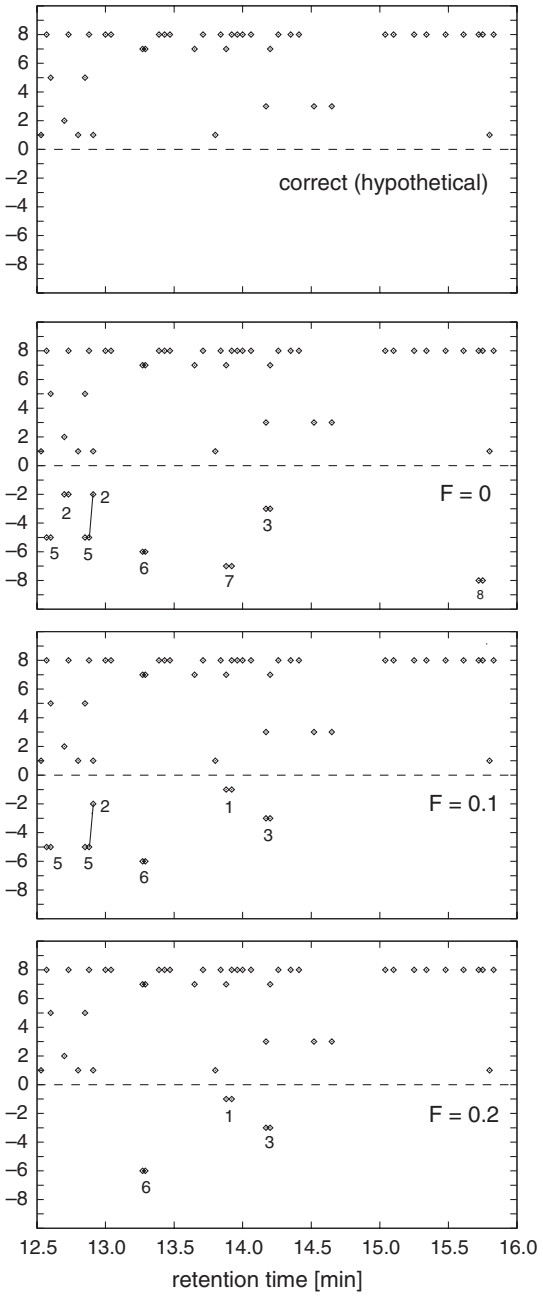


**Fig. 4.** The accuracy of peak clustering (complete linkage method). Correctly assigned peak clusters are shown above the zero line, with the number on *y*-axis corresponding to the number of peaks in that cluster. Clusters plotted below the zero line are those that involved either missing or erroneously assigned peaks. When a peak is erroneously assigned from one cluster to another, this affects two clusters from the viewpoint of an ideal answer. Therefore clusters below the zero line occur in pairs, and such pairs are joined by a line in the plot. In principle, such errors may involve more than two clusters, but this was not observed in the data shown here. For clarity, for clusters below zero the *y*-axis readout is shown next to each cluster point. The top panel shows the hypothetical situation when all peaks are correctly assigned to correct clusters. The three lower panels depict the actual results when the expansion factor $F$ was set to 0, 0.1 and 0.2 (from top to bottom). Increasing the expansion factor from 0 to 0.2 increased the accuracy of peak classification into clusters.

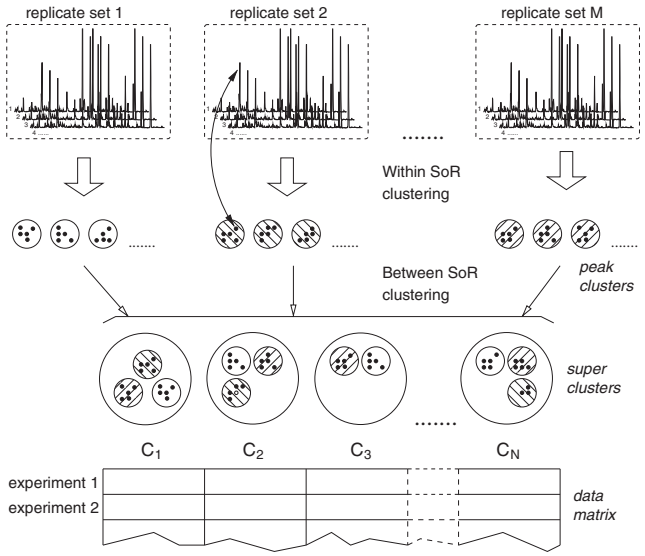


**Fig. 5.** A schematic representation of the progressive clustering procedure, showing clustering within set of replicates, clustering between sets of replicates, and the relationship to the data matrix (SoR denotes set-of-replicates).

is cut too high, two or more peak clusters belonging to the same cell state (i.e. the same set of replicate experiments) may be joined to the same super-cluster. By analogy with a peak collision, we denote this effect a 'cluster collision'. As the dendrogram cutoff is lowered, the number of super-clusters will increase and the number of cluster collisions will decline. If the dendrogram tree is cut too low relative to the optimal cutoff, some true super-clusters will be split to create two or more artificial super-clusters. When an extremely low cutoff is applied each peak cluster will become a super-cluster on its own.

In the clustering of peaks within a set of replicate experiments we cut the dendrogram tree to produce a predefined number of clusters related to the average number of peaks per replicate experiment ($N_{aver}$). A similar approach could not be applied to a dendrogram of peak clusters: when two or more cell states are analyzed it is unknown a priori how many super-clusters should be obtained. For example, in the case of two cell states with $N_1$ and $N_2$ unique analytes (metabolites) the final number of super-clusters could be anywhere between the largest of $N_1$ and $N_2$ (all metabolites found in one cell state found in the other) and $N_1 + N_2$ (none of the metabolites found in one cell state found in the other).

To find the optimal cutoff for a cluster dendrogram tree a radically different approach must be applied. We first note that individual peaks are single observations and are expected to be more prone to sampling variations relative to peak clusters, because the latter are derived from the set of replicate experiments. Furthermore, each peak cluster corresponds to a single chemical compound detected in a set of replicate experiments. Thus in any reasonable partitioning of peak clusters, cluster collisions must be rare or non-existent. This in turn suggests that the optimal dendrogram tree cutoff should be as high as possible to minimize the number of super-clusters, but not as high to produce cluster collisions.

**Table 1.** The summary of clustering results with $F = 0.15$ and $F = 0.20$ used for clustering of peaks within each set of replicate experiments

| | $N_{aver}$ | #clusters | #p.collisions | #clusters (final) |
|---|---|---|---|---|
| $F = 0.15$ | | | | |
| *wt* | 100.0 | 115 | 48 | 100 |
| Δ*gt* | 85.5 | 98 | 60 | 81 |
| Δ*pmi* | 91.6 | 105 | 43 | 94 |
| $F = 0.20$ | | | | |
| *wt* | 100.0 | 120 | 33 | 102 |
| Δ*gt* | 85.5 | 102 | 38 | 85 |
| Δ*pmi* | 91.6 | 109 | 29 | 98 |

The final number of super-clusters were 109 ($F = 0.15$) and 114 ($F = 0.20$), and the corresponding linear discriminant analysis is shown in Figure 6. $N_{aver}$ denotes the average number of peaks per replicate experiment, #clusters denotes the number of peak clusters and #p.collisions denotes the number of peak collisions. The column #clusters (final) shows the final number of peak clusters from the given set of replicate experiments (wt, Δgt, or Δ*pmi*, or Δ*pmi*) that entered the second stage of clustering, after filtering to remove spurious peak clusters (see main text).

We used this approach to analyze the GC-MS profiles of polar metabolite extracts of three strains of *L.mexicana*. After the peak clustering was performed within each set of replicate experiments, filtering was applied to remove spurious peak clusters. Because there were eight replicate experiments for *wt* and Δ*gt*, and seven replicate experiments for the Δ*pmi* genotype, any peak cluster that contained less than four peaks (*wt* and Δ*gt*) or three peaks (Δ*pmi*) was discarded. Remaining peak clusters from all three sets of replicate experiments were pooled together and were subject to hierarchical clustering. A complete dendrogram tree relating peak clusters from *wt*, Δ*gt* and Δ*pmi* experiments was generated. To obtain the final set of super-clusters the dendrogram tree was cut as high as possible, with the requirement imposed to produce no cluster collisions. This was achieved by cutting the dendrogram tree in small but finite steps, starting from an initial value close to zero. For each step the number of collisions was recalculated; when the first collision was observed, the previous cutoff was taken as optimal. Such scanning is very fast because it involves only cutting the previously created dendrogram tree (i.e. the hierarchical clustering is performed only once).

Finding the optimal dendrogram tree cutoff resulted in a unique set of super-clusters which was then transformed into a data matrix. The only empirically chosen parameter in this procedure was the expansion factor $F$ used in within-replicates peak clustering. Table 1 summarizes the results for the expansion factors $F = 0.15$ and $F = 0.20$. The final number of super-clusters was 109 ($F = 0.15$) and 114 ($F = 0.20$), and the resulting data matrices had dimensions $109 \times 23$ and $114 \times 23$. The linear discriminant analysis based on these two data matrices are shown in Figure 6.

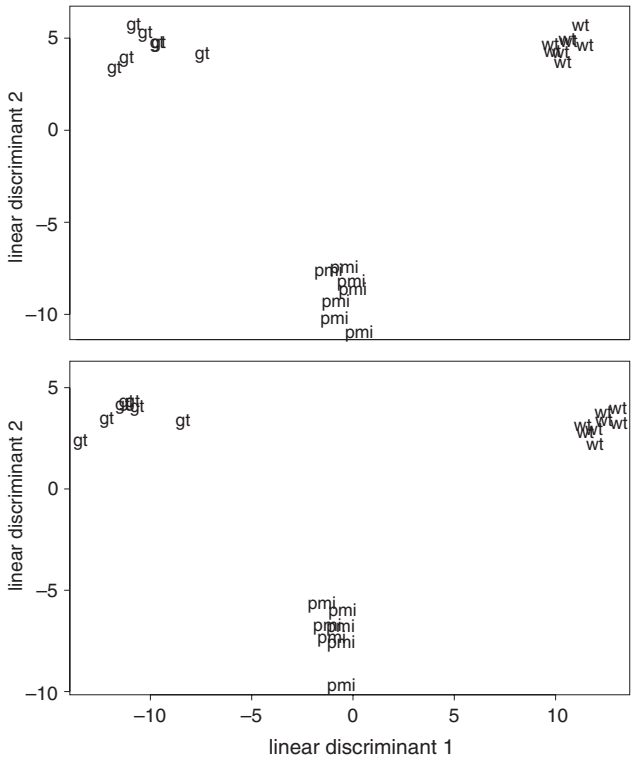


**Fig. 6.** Linear discriminant analysis of the data matrix produced by progressive clustering with $F = 0.15$ (upper panel) and $F = 0.20$ (lower panel). For simplicity, the individual experiments are labeled as follows: 'wt' for wild-type, 'gt' for glucose transporter mutant and 'pmi' for phosphomannose isomerase mutant experiments.

## 4 DISCUSSION

A common problem in the emerging field of metabolomics is the consolidation of peak lists derived from metabolic profiling of different cell/tissue/fluid states (i.e. wild-type versus mutant, diseased versus healthy, etc.), where a number of replicate experiments were collected on each cell state. This problem arises because of various experimental factors beyond the control of the experimentator; e.g. in the case of hyphenated mass spectrometry methods (GC-MS and LC-MS) the same analyte may elute at slightly different retention times in different experiments.

A recently described program MSFACTs is an attempt to address this problem (Duran *et al*., 2003). MSFACTs relies on the assumption that there is one-to-one correspondence between the peaks of any two peak lists (Duran *et al*., 2003). This assumption cannot be made when two or more different cell/tissue states are analyzed. Furthermore, the MSFACTs algorithm for peak classification depends on the minimum/maximum retention time and user-defined time interval. Here we propose to solve the overall problem sequentially: first by clustering peaks within each set of replicate experiments (within set of replicate experiments clustering) and then by clustering resulting peak clusters (between sets of replicate experiments clustering). This approach allows for signal peaks (i.e. metabolites) to be present in one not be present in others, the situation of central interest in practice. Furthermore, our approach relies on hierarchical clustering and is therefore symmetrical with respect to all peaks.

We analyzed polar metabolite extracts of three strains of *L.mexicana* parasites: a wild-type strain and two mutant strains with defects in carbohydrate metabolism owing to loss of glucose transporters or the enzyme phosphomannose isomerase (Burchmore *et al*., 2003; Garami and Ilg, 2001). Total ion chromatograms were integrated with a standard software package, and progressive clustering was applied to derive the data matrix from resulting peak lists.

The only empirical parameter which entered this analysis was the 'expansion factor' $F$ used in the within set of replicates clustering. The optimal value of $F$ was found empirically to be 0.15–0.25. The progressive clustering approach yielded a robust data matrix that completely resolved the three cell states in a discriminant analysis (Fig. 6).

The distance measure must be defined for each of the two clustering steps, i.e. (1) within a set of replicate experiments and (2) between sets of replicate experiments clustering. For (1) objects to cluster were peaks detected in individual replicate experiments, and we defined the distance between two peaks as the absolute value of the difference between their retention times. For (2) objects to cluster were peak clusters, and the average distance between all peaks from two peak clusters was used as the distance measure. Another obvious choice for the distance between two peak clusters is the distance between their peak centroids. In our preliminary investigation centroid distance measure did not produce significantly different results (data not shown).

Once the distance measure is defined, one can choose between several well-established clustering methods to produce the dendrogram tree. For clustering within replicates we investigated single linkage, complete linkage and centroid methods (Fig. 3), as well as Ward and average methods (data not shown). Given the data at hand all clustering methods performed reasonably well. Complete linkage and centroid methods performed slightly better than the single linkage method (Fig. 3). This was not surprising given that single linkage tends to produce elongated clusters, while complete linkage and centroid methods tend to produce compact, spherical clusters. Our preference is for a complete linkage method which seems to produce a minimal number of peak collisions when the dendrogram tree cutoff is smaller than optimal (Fig. 3).

An important question is how reproducible the data needs to be for the progressive peak clustering to be effective. The fundamental premise of the proposed approach was that differences in retention times of peaks originating from different analytes (metabolites) are greater compared with random variations (i.e. differences in retention times for the same peak observed in different replicate experiments). This seems to be a reasonable assumption for the data analyzed here (Figs 1 and 4). If the random variation approaches or exceeds differences in retention times between peaks originating from different analytes, the cluster analysis will be unable to assign peaks to correct clusters. This was confirmed by computer simulations based on synthetic datasets created by assuming realistic peak distributions and varying noise levels.

The analysis of manually classified peaks in the region 12.5–16.0 min (Fig. 1) showed that the noise in peak retention times is ~0.004 min (based on 30 peaks with 7 or 8 observations each). Other parts of the chromatogram showed poorer reproducibility, especially the initial parts of the chromatogram (the noise level of up to 0.014 min was observed for individual peaks in the region 6–9 min). In a typical wild-type experiment the minimum difference in peak positions was found to be ~0.020 min. This relationship between the noise level and differences in peak positions is favorable for the type of analysis proposed here.

The simple retention time distance is attractive because of its simplicity, speed of analysis and its ability to produce robust results in the analysis presented here (Fig. 6). We are currently investigating the inclusion of mass-spectra, which has the potential to increase accuracy of peak classification in multiple experiments, although at the cost of significant increase in the complexity of calculations.

## ACKNOWLEDGEMENTS

This work was supported by the Bio21 Molecular Science and Biotechnology Institute (University of Melbourne), and Australian National Health and Medical Research Council (NHMRC) Programme grant and Principal Research Fellowship (to M.J.M). D.P.D. was supported by NHMRC 'Dora Lush' Biomedical scholarship (Id. No. 359427).

*Conflict of Interest:* none declared.

## REFERENCES

Allen,J. *et al*. (2003) High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat. Biotechnol*., **21**, 692–696.

Burchmore,R.J. *et al*. (2003) Genetic characterization of glucose transporter function in *Leishmania mexicana*. *Proc. Natl Acad. Sci. USA*, **100**, 3901–3906.

Castrillo,J.I. *et al*. (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry*, **62**, 929–937.

Choi,Y.H. *et al*. (2004) Metabolomic differentiation of *Cannabis sativa* cultivars using 1H NMR spectroscopy and principal component analysis. *J. Nat. Prod*., **67**, 953–957.

Davis,A.J. *et al*. (2004) Drugs against leishmaniasis: a synergy of technology and partnerships. *Trends Parasitol*., **20**, 73–76.

Duran,A.L. *et al*. (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, **19**, 2283–2293.

Fernie,A.R. *et al*. (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol*., **5**, 1–7.

Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant. Mol. Biol*, **48**, 155–171.

Fiehn,O. *et al*. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol*., **18**, 1157–1161.

Garami,A. and Ilg,T. (2001) The role of phosphomannose isomerase in *Leishmania mexicana* glycoconjugate synthesis and virulence. *J. Biol. Chem*, **276**, 6566–6575.

Goodacre,R. *et al*. (2002) Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst*, **127**, 1457–1462.

Goodacre,R. *et al*. (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol*, **22**, 245–252.

Guillo,C. *et al*. (2004) Micellar electrokinetic capillary chromatography and data alignment analysis: a new tool in urine profiling. *J. Chromatogr. A*, **1027**, 203–212.

Raamsdonk,L.M. *et al*. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol*., **19**, 45–50.

Roessner,U. *et al*. (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J*., **23**, 131–142.

Roessner,U. *et al*. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.

Sato,S. *et al*. (2004) Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J*., **40**, 151–163.

Sumner,L.W. *et al*. (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **62**, 817–836.

Tolstikov,V.V. *et al*. (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem*., **75**, 6737–6740.

Urbanczyk-Wochniak,E. *et al*. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep*., **4**, 989–993.