BIOINFORMATICS

Vol. 1 no. 1 2002
Pages 1–8

Application of metabolomics to plant genotype discrimination using statistics and machine learning

Janet Taylor[†*], Ross D. King[†], Thomas Altmann[‡] and Oliver Fiehn[‡]

[†]Department of Computer Science, University of Wales, Aberystwyth, Penglais, Aberystwyth, SY23 3DB, UK and [‡]Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany

ABSTRACT

Motivation: Metabolomics is a post genomic technology which seeks to provide a comprehensive profile of *all* the metabolites present in a biological sample. This complements the mRNA profiles provided by microarrays, and the protein profiles provided by proteomics. To test the power of metabolome analysis we selected the problem of discrimating between related genotypes of *Arabidopsis*. Specifically, the problem tackled was to discrimate between two background genotypes (Col0 and C24) and, more significantly, the offspring produced by the cross-breeding of these two lines, *the progeny* (whose genotypes would differ only in their maternally inherited mitichondia and chloroplasts).

Overview: A gas chromotography - mass spectrometry (GCMS) profiling protocol was used to identify 433 metabolites in the samples. The metabolomic profiles were compared using descriptive statistics which indicated that key primary metabolites vary more than other metabolites. We then applied neural networks to discriminate between the genotypes. This showed clearly that the two background lines can be discrimated between each other and their progeny, and indicated that the two progeny lines can also be discriminated. We applied Euclidean hierarchical and Principal Component Analysis (PCA) to help understand the basis of genotype discrimination. PCA indicated that malic acid and citrate are the two most important metabolites for discriminating between the background lines, and glucose and fructose are two most important metabolites for discriminating between the crosses. These results are consistant with genotype differences in mitochondia and chloroplasts.

Key Words: Metabolome, Arabidopsis, Clustering

Corresponding Author: *jat@aber.ac.uk

INTRODUCTION

Post genomic molecular biology is being driven by new, and highly powerful experimental techniques which enable the large-scale, and parallel interrogation of cell states under different stages of development and defined environmental conditions. Such analyses may be carried out at the level of transcription using hybridisation-arrays – the *transcriptome* (Baldwin *et al.* (1999); Ruan *et al.* (1998)). Similar analyses may be carried out at the level of translation to define the proteome (Santoni (1998)). Most recently, the metabolome (the cells small molecule complement) has risen to prominance as an essential component in cell analysis (Raamsdonk *et al.* (2001)).

The goal of metabolome research is to be able to provide a comprehesive profile of *all* the metabolites present in a biological sample. Analysis of cells at the metabolic level has a number of advantages over the more conventional transcriptome and proteome analyses (Tretheway *et al.* (1999); Katona *et al.* (1999); Adams *et al.* (1999)):

- Changes in gene and protein expression can cause amplified changes in metabolism, making detection easier.
- Metabolome technology does not require the complete genome sequence or an large EST databases, as do many transcriptome and proteome approaches.
- There are fewer metabolite types than genes or proteins: in the order of 1000 per organism compared to several thousand genes for the smallest bacterial genomes and 10's of thousands of genes for complex multi-cellular organism.
- The technology is more generic, as a given metabolite - unlike a transcript or protein - is the same in every organism.

To test the power of metabolome analysis we selected the problem of discriminating between related genotypes of *Arabidopsis*. This problem is of clear biological interest, with applications in plant breeding and ecology; and has the data analysis advantage of providing a clear cut measure of success or failure - unlike data analysis based purely on clustering. Specifically, the problem tackled was to discriminate between to background lines of *Arabidopsis thaliana* (Col0 and C24) and their first generation (**F1**) progeny (C24 x Col0 and Col0 x C24). This problem is interesting because although previous work has indicated that it is possible to discramate between different lines of Arabidopsis using metabolomics (Fiehn *et al.* (2000)), discrimation between different types of interbred progeny is more biochemically focussed, and far more challenging. The two forms of F1 (first generation) progeny only differ by which background line was male and which female. This means that they are genetically identical (neglecting any possible imprinting) except for their maternally inherited mitichondria and chloroplast genomes. The bioinformatic questions therefore are:

- Can the background genotypes be discriminated?
- Can the F1 progeny genotypes be discriminated from the parents and themselves?
- Can the resuts of any discrimination method be biologically interpreted?

From the previous work published (Fiehn *et al.* (2000)), it is expected that the answer to the first of the above questions to be positive, but whether the two related, but differentially inherited progeny genotypes can be separated is the focus of this study.

MATERIALS AND METHODS

Plant Growth and Harvest

Two background lines of *Arabidopsis thaliana* (Col0 and C24) were crossed to produce two F1 progeny lines with differing patterns of maternal inheritance, one F1 progeny was Col0 x C24, and the second progeny was C24 x Col0. The seeds of the parent controls were derived from manual self-fertilisation of the parent lines in order to control for potential effects (seed size) caused by the crossing procedure.

Seeds were dispersed onto moist standard soil (Einheitserde GS90, Gebrüder Patzer, Sinntal-Jossa, Germany) and cultivated for 10 days (16h day, 145 μ E fluorescent light Philips TLD36W/830 + TLD36W/840, 20^{0}C, 75% rel. humidity/ 8h night, 6^{0}C, 75% rel. humidity). Seedlings (across the different populations synchronously germinated seedlings were selected) were than picked into individual pots (with standard soil: Einheitserde GS90, Gebrüder Patzer, Sinntal-Jossa, Germany) and cultivated under a day/night regime of 16h day, 120 μ E fluorescent light Philips TLD36W/830 + TLD36W/840, 20^{0}C, 60% rel. humidity / 8h night, 16^{0}C, 75% rel. humidity until flowering. Leaves were harvested from plants with a primary inflorescence of 5 cm. All plants were harvested at Boyes stage 6.0 - 6.5, weighed and frozen in liquid nitrogen.

Leaf Extract Preparation

About 100mg (fresh weight) of frozen ground tissues of Arabidopsis leaves were extracted as previously reported (Fiehn *et al.* (2000b)). 80% aqueous methanol at 70^{0}C and chloroform at 37^{0}C were used, combined, and phase separated by addition of water and subsequent centrifugation. The lipophilic phase was not investigated in this study. The polar phase was dried down in a SpeedVac concentrator. Metabolites were subsequently derivatized by methoximation and trimethylsilation prior to analysis into the Gas Chromatography Mass Spectroscopy machine (GC-MS) as described in (Fiehn *et al.* (2000b)).

Data Acquisition and Pre-processing

GC-MS chromatograms are potentially information-rich entities of some 3.24 MB, with mass spectral information and ion intensities for eash of the 0.5s long scans that are acquired over a 1200s chromatographic run time. This raw form of data is not suitable for data analysis and needs to be refined, to extract useful information from the large quantity of raw data. Our data acquisition and pre-processing procedure was designed to extract the maximum reliable information from the chromatograms. In metabolomic studies we aim to identify as many metabolites as possible (not just a pre-defined set based on background knowledge of their biological importance as has been typical of studies of metabolites). This requires a semi-automated data analysis approach.

The data analysis process begins with the selection of a reference chromatogram that is typical of the whole set of Arabidopsis analyses. This is based upon visual inspection off the number of compounds and the presence of low abundant metabolites such as organic phosphates and and certain di- and trisaccharides. Since an important objective of this study was to investigate metabolic difference between F1 and parental lines, and to inspect differences between both F1 lines in order to look for mitochondrial inheritance, an F1 sample was chosen as reference for both the lipophilic and the polar chromatograms.

The second part of the data pre-processing was based on the automated mass spectral deconvolution and identification system AMDIS developed by Steven Stein (Stein (1999)) and implemented for a variety of MS formats; including the ThermoFinnigan Quadrupole mass spectrometer used for this study. This deconvolution software enables finding of peaks in an unbiased way without prior knowledge about their mass spectral characteristics and chemical nature. Since it was developed for detecting chemical warfare agents, the software suffers from several weaknesses for automated quantitation of a multitude of metabolites in high throughput applications.

For example, depending on the threshold used, either too many false positive or false negative peak findings are generated. Further, quantitation in AMDIS uses the sum of all deconvoluted ion traces for each peak (dTIC). It is therefore prone to cause problems for all minor peaks that regularly contain noise ions from chemical background or coeluting major peaks. Finally, retention time shifts after column changes have not adequately been taken into account by the software developers. For these reasons, AMDIS deconvolution was only used for finding all peaks in the reference chromatograms.

This step was followed by use of a script written in MassLab for identifying those peaks that could be confirmed to exceed signal-to-noise ratios (S/N) of 5 and that had peak widths of at least 5s. For each of the newly defined target peaks, dedicated ions were chosen at high masses for the MassLab routine quantitation method. Using this procedure 433 peaks were positively detected in the polar phase of a single Arabidopsis leaf extract. All other chromatograms were then matched against this list of pre-defined target analytes. Peak areas were normalized to mg fresh weight and to the internal standard ribitol. Since very low thresholds were used for mass spectral quality in the MassLab routine, false negative results were minimized. A number of peaks that were detected in the reference chromatograms were apparently absent in the sample chromatograms. This mainly occured because the comarison peaks had a lower S/N than 5. Therefore, the true peak areas for these peaks were somewhere between zero and the detection limit. The value of 0.000001 was given to such peaks.

It should be noted that use of a reference chromatogram has the data analysis advantage of producing examples all with the same number of attributes. The data for analysis is therefore a simple 2D matrix, enabling the data analyst to choose from many statistical and machine learning data analysis methods (both simple and complex). Complex data structures (where different objects have different numbers of attriibutes) can often require the use of more complex data analysis approaches such as ILP (Muggleton *et al.* (1998)).

RESULTS

The data comprised eight examples of each of the four genotypes; with each sample originating from a separate plant, meaning only one replicate object in the dataset per plant. These were numbered: 1-8 = Col0 parental genotype, 9-16 = C24 parental genotype, 17-24 = Col0 x C24 F1 progeny, 25-32 = C24 x Col0 F1 progeny. In the reference GC-MS chromatogram 433 peaks corresponding to metabolites were identified; 201 of these could be identified at some detail, 92 as molecular type (e.g. “alanine”, “6-hydroxynicotinic acid”) and 109 by chemical property (e.g. “sugar alcohol”, “aromatic compound”). The remaining samples were labelled with a unique identification number. Each peak had a real valued quantity. This resulted in a processed dataset of 433 x 32 real numbers. No initial preprocessing of the data was carried out in terms of data reduction. Whilst some variable selection could have removed irrelevant variables, it was undesirable to do this when, in this case, there was no prior information concerning the relevance of each variable.

Descriptive Analysis

The first step in the analysis was to apply some standard descriptive statical methods to the data. This was done using the R statistical package (www.r-project.org).

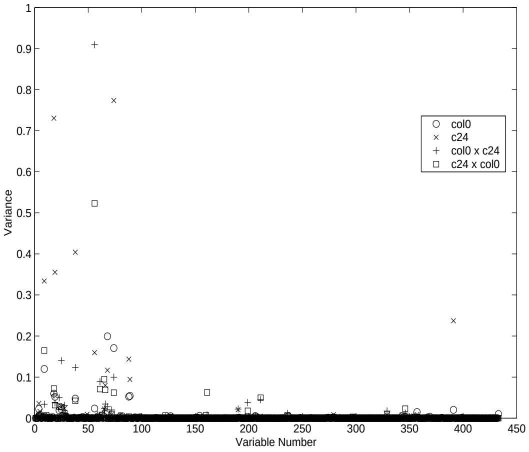

Fig. 1. The Variance of Each Metabolite for Each Genotype

Perhaps the most interesting result of this is shown in Figure 1. This displays a plot of the variance of each identified metabolite. Note that the highest variance occur within the first $\sim$ 100 peaks in the data. The peaks are listed by name in the original data, meaning that in Figure 1, the highest variances are associated with those peaks that are well defined and have a compound label associated with them. This is further shown in Table 1, where the average variance for each class of compound label is shown. The identified metabolites for each genotype have an average variance at least 10 times those of less well defined peaks.

The probable explanation of this is that these identified metabolites are generally core primary metabolites, and as such their quantities are likely to be affected by many genetic changes. This would appear to be consistant with the predictions of Metabolic Control Analysis (MCA) (Mendes (1997); Cornish-Bowden (1995); Mendes and Kell (1998)).

Table 1: The Average Variance of Each Metabolite for Each Genotype

	Label(name)	Label(property)	Label(number)
Col0	0.01	0.00022	0.000367
C24	0.037	0.0005	0.0013
Col0xC24	0.019	0.00078	0.00049
C24xCol 0	0.014	0.001	0.00052

Each metabolite peak (column) in the data was plotted as a simple bar graph so that the relative peak area for a particular metabolite may be compared between genotypes. 433 bar graphs were plotted in total (not shown) and examined visually. A number of metabolites showed marked differences in measurement across genotypes, and as described above, this was even true for key primary metabolites. 27 metabolite peaks in the reference spectrum were not measured for the Col0 parent (either due to the compound not being present or in such low quantities as to not be detected), 14 were absent from the C24 parent, and only one metabolite peak was not measured for both F1 genotypes. Of those absent from the parental strains, perhaps the most suprising was the absence of a derivative of glucose-6-phosphate, as this molecule is part of a pool of intermediates in which the reactions of carbohydrate synthesis and degredation converge and interact with other pathways (Dennis *et al.* (1990)).

Figure 2, plots a and b show the distribution of raffinose and galactose respectively. These two metabolites are further examples of some highly discrminant metabolite measurements. Raffinose is a trisaccharide commonly found in higher plants in the phloem. It is synthesized by the attachment of a D-galactose to the C6 position of the D-glucose moiety in sucrose. These sugars are important in the long distance transport of carbon from source to sink (Greenberg (1967)).

Genotype Discrimation

Following the preliminary analysis of the data using descriptive statistics we sought to discriminate between the four genotypes. Use of linear discrimination was ineffective at this, we therefore used the standard neural network approach of back-propagation. This was done using the freely available Java software package WEKA (Version 3.2, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ ml/weka/). A multilayered percepton network with sigmoid units was trained using 32-fold (leave one out) cross validation (Witten *et al.* (2000).

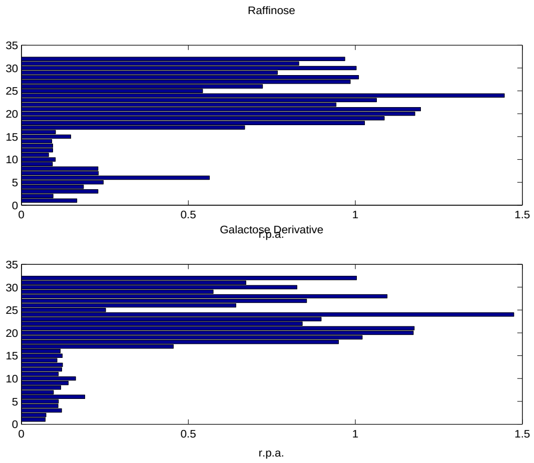

Fig. 2. Comparison of the Relative Peak Area of Two Example Metabolites between Genotypes (r.p.a = Relative peak area)

Table 2: Confusion Matrix of the Neural Network Classifier

A	B	C	D	←classified as
8	0	0	0	A = Col0
0	8	0	0	B = C24
0	0	5	3	C = Col0 x C24
0	0	3	5	D = C24 x Col0

Because the number of examples in the data is limited, to reserve some of the examples to form an independant test set could mean that the model does not learn from a truly representative training subset. Leave one out cross validation involves the ommission of each data object in turn from the training set, and using all others to train the model. The model is then judged on its ability to classify the remaining object. This is repeated for all objects in the data. Use this form of cross validation ensures that the maximum amount of data is used for the training of the model, which is particularly important when analysing a small number of samples.

The neural network trained in this way correctly classified 26 out of the 32 examples in the data. The confusion matrix for the classifier is shown in Table 2.

Of the parental strains, all of the Col0 examples were correctly classified, as were the C24 parental examples. From these results it is clear that it possible to discriminate between the two parental lines and between the parental lines and the two F1 crosses. However, it is less clear if the two F1 crosses can be discriminated. To quantitatively test this we applied a binomial test, i.e. if the discrimant was random what is the probability of getting 10 correct predictions and 6 wrong? The answer is P = 0.27, or around 1 in 3. Although this value is far higher than normally considered significant in normal tests, it must be taken into account that the sample size of 16 is very small. We therefore believe that it is fair to conclude that balance of evidence favours the hypothesis that discrimation is possible even for the case of the subtle differences between the different F1 genotypes Col0 x C24 and C24 x Col0.

Distance metrics between examples

To investigate further the differences in the metabolome between genotypes we carried out a cluster analysis of the data (this is an unsupervised approach in contrast to the supervised approach taken with the neural network (Everitt (1974)). This was done because we wished to have a better biological understanding of which metabolites were involved in the discrimination and it is notoriously difficult to directly interpret the weights in a multi-layer non-linear neural network.

The majority of the clustering algorithms undertake an initial calculation to produce a similarity or distance matrix between entities in the data. To calculate this we used the same data as for the variance calculations. Various distance (or *dissimilarity*) metrics were investigated (Duda *et al.* (1973); Jain *et al.* (1988)) and they all produced broadly similar results. We have therefore chosen to illustrate these results using the the most commonly used distance metric, that of Euclidean distance. Objects are compared in a pairwise fashion to calculate the distance. The Euclidean distance describes a circular population boundary in two dimensions, and a sphere or hyper-sphere in 3 or more dimensions. If a particular population or cluster does not have a hyper-spherical boundary (it may for example have an elliptical shape along a particular axis) then this distance measure may fail to group objects correctly.

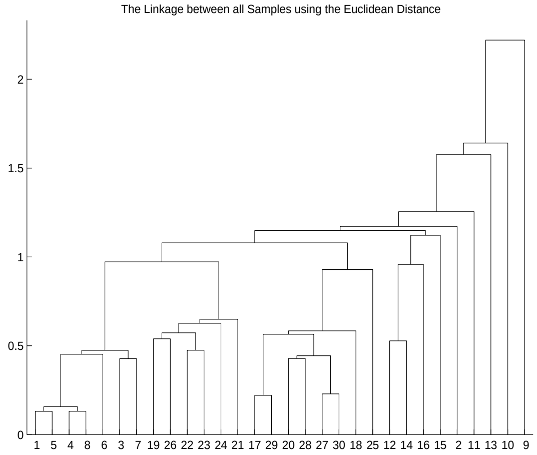

Fig. 3. Distances between each object in the dataset, according to the Euclidean distance metric. 1-8 = Col0 parental genotype, 9-16 = C24 parental genotype, 17-24 = Col0 x C24 F1 progeny, 25-32 = C24 x Col0 F1 progeny

Calculation of these distance metrics allows the plotting of a dendrogram, or hierarchical tree, which illustrates how far apart (or *dissimilar*) objects are from each other. Dendrograms may also be employed to cluster the data by terminating the tree hierarchy at a desired threshold. Figure 3 shows an example of a dendrogram produced by the calculation of the Euclidean distance. It is shown in this dendrogram that the parental Col0 and both F1 progeny group more tightly together, i.e. all members of the group are in close proximity according to the Euclidean distance. The C24 genotype, on the other hand were more distant, both from other genotypes and other members of that group. However there is a significant outlier in the parental groupings. One of the Col0 objects is found to be ’nearer’ to the C24 objects. Of the F1 progeny, five of the Col0 x C24 crosses lie next to each other, but there is no other significant grouping of the F1.

Principal Components Analysis

To investigate further the role of the different metabolites in discrimination we carried out a principal component analysis (PCA) of the data. This unsupervised multivariate data analysis approach is apppropriate when it is believed that a function of many attributes (metabolites) is involved in differences between examples. PCA is primarily concerned with the transformation of a large set of related variables into a new, smaller set of uncorrelated variables (Joliffe (1986)). The new variables are termed latent variables, or principal components (PCs). The PCs attempt to express the maximum variation of the original data. Each principal component may be thought of as an axis in multi-dimensional space, and each object can then be characterized by how far away it lies to a particular axis. This calculation gives each object a *score*. The contribution of each variable to a particular PC can also be calculated. This gives each variable a weighting value or loading for a PC. High positive or negative loading values for a variable indicate a strong contribution to that PC.

The PCA algorithm used for the following analysis was taken from the Statistics Toolbox of Matlab. The same data that were used for the calculation of the distance metrics were input to the PCA routine. From this analysis, the first three PCs were examined in detail. Figure 4 illustrates the individual and cummulative variance that is explained by each of the first few PCs. 78% of the variation in the original data is explained using the first three PCs alone.

As all the constructed PCs are orthogonal, the object scores may be plotted against each other to represent the distribution of the objects in the space. Figure 5 shows the first 3 principal components plotted against each other, giving rise to separate clusters for each of the parental genotypes. However, one of the Col0 samples lies much closer to the C24 cluster, than to other Col0 objects. This is consistent with the results gained from the distance metric described in the previous section. It is also shown that five of the Col0 x C24 samples cluster together, and of the remaining F1 progeny, seven out of the eight C24 x Col0

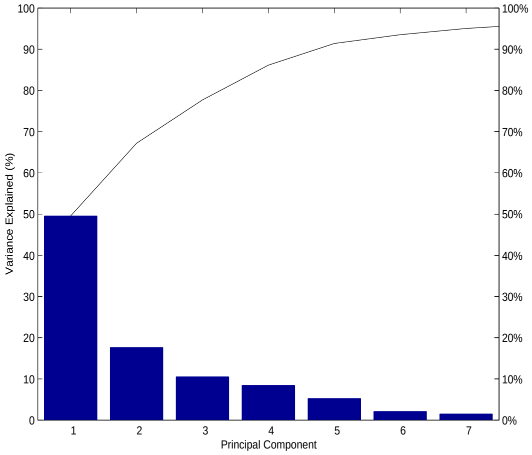

Fig. 4. The Variance of the Data Explained by the First Few Principal Components

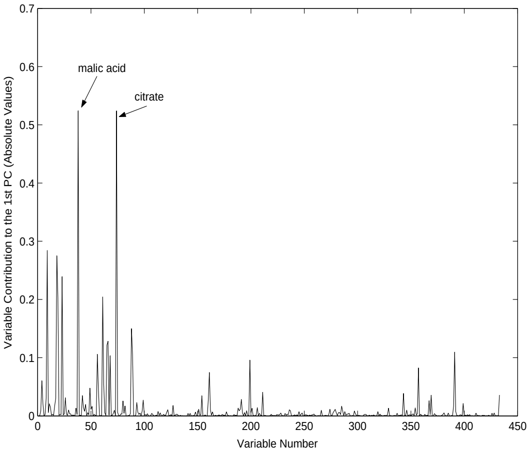

Fig. 6. The contribution of each variable (metabolite peak) to the first principal component

samples also form a loose cluster, but one which contains the remaining Col0 x C24 samples.

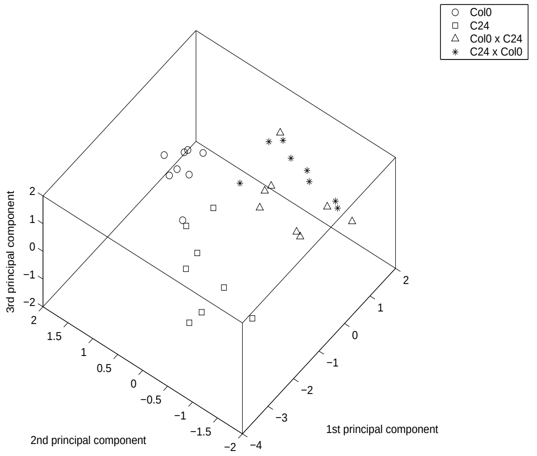

Fig. 5. Clustering of the Genotypes by 3 Principal Components

The contribution of each variable to the first PC is illustrated in Figure 6. There are two variables whose absolute loading values are much greater than any other variable. These two lie at positions 38 and 74 of the original data, and represent the relative peak areas for the metabolites malic acid and citrate.

This prompted further study of the original measurements was undertaken, firstly by returning to the descriptive statistics carried out previously. Figure 7 shows a boxplot for both metabolites for each genotype class. It is immediately apparent that the citrate peak data for the Col0 parental line contains an outlier, which lies far outside the remaining Col0 sample range. The value for this outlier lies within the range of the citrate peak values for the C24 genotype. The presence of this one outlying value may explain the misgrouping of one of the Col0 samples by the Euclidean distance measure. It may also be observed from this boxplot that the range of values for both metabolites is much wider in the C24 parent than for any other genotype class.

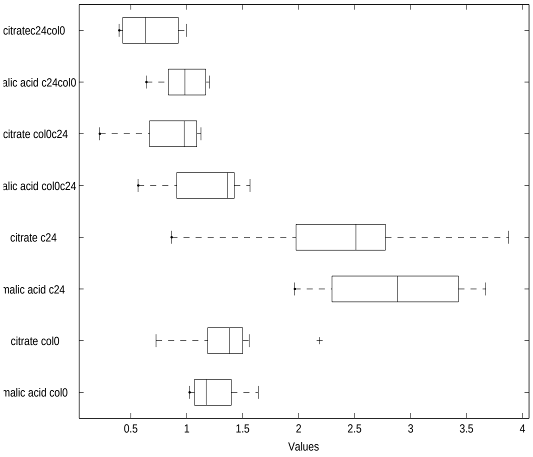

Fig. 7. Box plot of the relative peak area data for the metabolites citrate and malic acid, for both parental and F1 progeny. The plot shows the median, 25% quartile, 75% quartile and the range.

Examination of the loadings plots from the second and third PCs (plots not shown) reveal other highly

Table 3: The Variables with the Highest Absolute Loadings in the First 3 PCs

PC	1	2	3
	Malic acid	Glucose meox 2	Glucose meox 2
	Citrate	Fructose meox 1	Threonine
	Threonine	Malic acid	Galactose meox 1
	Serine TMS 3X	Fructose meox 2	Citrate
	Raffinose	Galactose meox 1	Malic acid
	Galactose meox 1	Pyroglutamic acid	Serine TMS 3X
	Serine TMS 2X	Raffinose	Fructose meox 1

contributing single metabolite peaks. In the second PC, two peaks labelled Glucose Meox 2 (a glucose derivative) and Fructose Meox 1 (a fructose derivative) had the highest contribution, whilst the same glucose derivative in the third PC contributed over 50% more than any other peak.

Further to the PCA carried out as previously described, a similar analysis was undertaken using the F1 progeny alone, to remove the variation due to the parental genotypes. In the first PC of this analysis, the glucose derivative had the highest loading value, contibuting 29% of the total loadings values for all 433 variables. If a further PCA is performed on just the parental data, malic acid and citrate again contribute most highly to the clustering of the data in the first PC.

It may be therefore inferred from the original PC analysis of all data, that the first PC partions the parental genotypes, and separates them from the F1, and the second and third PCs discriminate between the F1 progeny, using a glucose derivative as the major discriminating metabolite.

Table 3 lists the metabolites with the highest loading values for each principal component. It is observed that many metabolites appear significant in more than one PC, leading to nine individual metabolite peaks that are identified as significant by the PCA analyis.

Malic acid and citrate are two key metabolites in the TCA cycle, located within the mitochondrial matrix. Serine is also produced in the mitochondria, then is transported to the peroxisome for further processing. Raffinose and galactose, as stated previously, are important sugars in the transport of carbon in the phloem. Threonine is used either for protein synthesis, or converted to isoleucine. The enzymes involved in the synthesis and processing of threonine, and related compounds are located in the chloroplast. Pyroglutamic acid is a component in the pathway of glutathione metabolism (www.genome.ad.jp/kegg/, Anderson *et al.* (1998)). Oxidised glutathione is suggested by (Cohen (1993); Buchanan (1994)) to be associated with the oxidation of key chloroplastic enzymes to their disulphide form in the dark. This is a form of reversible covalent modification which is extremely important in the in light regulation of these enzymes, and photosynthetic electron flow (Dennis *et al.* (1990)).

Glucose and fructose are two of the most commonly occuring sugars in plants, having multiple roles within the plant cell. However, chloroplasts have the key role in producing glucose and fructose through photsynthetic driven production of glyceraldehyde 3-phosphate, and mitochondia have a key role in in removing glucose through catabolism of pyruvate in the TCA cycle.

Therefore the results are consistant with the knowledge that the only genetic differences are the maternally inherited chloroplasts and mitichondria.

DISCUSSION

The comprehensive study of the metabolome of organisms is just beginning. However, already it is clear that the general approach has great potential to increase our knowledge of the internal state of the biochemistry of cells. Technological advances are still needed, both to improve the sensitivity of the detection of metabolites, and to characterise the chemical nature of the detected metabolites. There is also a great need to be able to improve the temporal resolution of detection methods. This will be essential in moving to the goal of realistic *in silico* models of cellular metabolism. In bioinformatic terms the data presents numerous challenges. Our knowledge about metabolites needs to be formalised into an ontological structure - as as been done for other areas of biology (Schulze-Kremer (1997)). This ontology must formalise our knowledge both about the chemical structures of the metabolites (hydrophobic, hydrophilic, acid, sugar, etc), as well as their role in biochemistry (glycolysis, TCA cycle, etc.). Once this knowledge is formalised it then needs to be incorporated into automated data analysis methods. Metabolomic data can then be integrated with other forms of bioinformatic data: genomic, transcriptomic, and proteomic to provide a comprehensive description of cells and organisms.

ACKNOWLEDGEMENTS

This study is funded by the UK Food Standards Agency (Grant number G02006) and the authors gratefully acknowledge the contribution of Professor John Draper, Professor Douglas Kell, Mr Nigel Hardy, Dr Helen Fuell and the other co-workers of the consortium both in Aberystwyth and Potsdam.

REFERENCES

Joliffe,A. (1986) Principal Components Analysis. *Springer-Verlag, New York.*

Everitt,Brian (1974) Cluster Analysis. *Heinmann International Books Ltd., London.*

Baldwin,D., Crane,V. and Rice,D. (1999) A comparison of gel-based, nylon filter and microarray techniques to detect differential RNA expression in plants. *Current Opinion in Plant Biology*, **2**, 96–103.

Ruan,Y., Gilmore,J. and Conner,T. (1998) Towards *Arabidopsis* genome analysis: Monitoring expression profiles of 1400 genes using cDNA microarrays. *The Plant Journal*, **15**, 821–833.

Santoni,V. (1998) Use of proteome strategy for tagging genes present at the plasma membrane. *The Plant Journal*, **16**, 633–641.

Tretheway,R., Krotzky,A.J. and Willmitzer,L. (1999) Metabolic Profiling: A rosetta stone for genomics?. *Current Opinions in Plant Biology*, **2**, 83–85.

Katona,Z.F., Sass,P. and Molnár-Perl,I. (1999) Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography-mass spectrometry. *Journal of Chromatography A*, **847**, 91–102.

Adams,M.A., Chen,Z.L., Landman,P. and Colmer,T.D. (1999) Simultaneous determination by capillary gas chromatography of organic acids, sugars and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *annals of Biochemistry*, **266**, 77–84.

Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from GCMS data. *Journal of the American Society of Mass Spectrometry*, **10**, 770–781.

Fiehn,Oliver, Kopka,Joachim, Dörmann,Peter, Altmann,Thomas, Tretheway,Richard and Willmitzer,Lothar (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology*, **18**, 1157–1161.

Fiehn,Oliver, Kopka,Joachim and Tretheway,Richard (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry*, **72**, 3573–3580.

Muggleton,S., Srinivasan,A., King,R.D. and Sternberg,M.J.E. (1998) Biochemical knowledge discovery using inductive logic programming. *Discovery Science. Lecture Notes in Artificial Intelligence, Springer Verlag, Berlin*, 326–341.

Witten,IanH. and Frank,Eibe (2000) Data mining. Practical machine learning tools and techniques with Java implementations. *Morgan Kaufann, San Francisco*.

Mendes,PedroM. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. *Trends in Biochemical Sciences*, **22**, 361–363.

Cornish-Bowden,A. (1995) Metabolic control analysis in theory and practice. *Advances in Molecular Cell Biology*, **11**, 21–64.

Mendes,Pedro and Kell,DouglasB. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, **14**, 869–883.

Schulze-Kremer,Steffen (1997) Adding semantics to genome databases: Towards an ontology for molecular biology. *ISMB-97, Proceedings of the fifth international conference on intelligent systems for molecular biology*, 272–275.

Duda,R.O. and Hart,P.E. (1973) Pattern Classification and scene analysis. *John Wiley, London*.

Jain,A.K. and Dubes,R.C. (1988) Algorithms for clustering data. *Prentice Hall, Englewood Cliffs, NJ*.

Greenberg,D.M. (1967) Metabolic Pathways Volume 1: Energetics, Tricarboxylic Acid Cycle and Carbohydrates. *Academic Press, New York*.

Dennis,D.T. and Turpin,D.H. (1990) Plant Physiology, Biochemistry and Molecular Biology. *Longman Scientific and Technical, Harlow, UK*.

Anderson,M. and Roberts,J. (1998) Arabidopsis, Annual Plant Reviews Volume 1. *Sheffield Academic Press, UK*.

Buchanan,B. (1994) The ferredoxin/thioredoxin system: A key element in the regulatory function of light in photosynthesis. *BioScience*, **34**, 378–383.

Cohen,P. (1993) Control of Enzyme Activity. *Chapman and Hall, London*.

Raamsdonk,M.L., Teusink,B. and others, (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, **19**, 45–50.