Metabolomics (2007) 3:349–355
DOI 10.1007/s11306-007-0066-2

ORIGINAL ARTICLE

Statistical measures for validating plant genotype similarity assessments following multivariate analysis of metabolome fingerprint data

David P. Enot · John Draper

Received: 19 February 2007 / Accepted: 5 June 2007 / Published online: 23 August 2007

Abstract Metabolome fingerprinting offers opportunities for 'first pass' evaluation of compositional similarity between plant genotypes. Compositional "substantial equivalence" testing is a popular concept in the literature in relation to food safety; however reported studies do not provide a systematic and standard approach to quantify similarity in a high dimensional data context. We have undertaken a large scale screen of *Arabidopsis* genotypes for evidence that individual genetic modifications effect plant phenotype at the level of the metabolome. From this study we propose pragmatic alternative measures that could in the future be used to assess substantial equivalence in GM foods under realistic data paucity constraints and without prior feature selection. Evaluation of classifier accuracy in supervised data mining approaches by bootstrap error estimation provided a robust tool for model validation. Receiver operating characteristics (such as AUC) provide an alternative measure of predictive ability by displaying the relationship between sensitivity and specificity. Additional specific measures based on scatter matrices and sample margins have also been investigated. We illustrate the application of such metrics on a large metabolic profiling data set derived from analysis of 27 genetically distinct *Arabidopsis thaliana* mutants. We show that agreement exists between model margins, eigenvalue, accuracies and AUC characteristics produced by three different classifiers (Random Forest, Support Vector Machine and Linear Discriminant Analysis). Comparisons between mutants with no observable phenotypic differences to the parent ecotype provided a baseline for model significance metrics; whilst comparison of mutants with increasingly distinct phenotypic alterations generated predictable changes in these measures of similarity.

Keywords GM substantial equivalence · Multivariate model distance metrics · Metabolome fingerprinting · Random Forest · Model margins · Sensitivity and Specificity

1 Introduction

Several recent papers have highlighted the usefulness of metabolite profiling or metabolite fingerprinting to examine the compositional differences between genetically modified plants and their progenitor genotypes (Baker et al. 2006; Catchpole et al. 2005; Charlton et al. 2004; Choi et al. 2004; Fukusaki and Kobayashi 2005; Garratt et al. 2005; Le Gall et al. 2003; Manetti et al. 2006; Mattoo et al. 2006; Shepherd et al. 2006). A key issue in compositional assessments of genetically modified foods is the absence of systematic and understandable metrics of similarity based on metabolic profiling data (Cockburn 2002; Konig et al. 2004; Kuiper et al. 2003). Testing for similarity between genotypes (and substantial equivalence) usually refers to the application of statistical routines that aim to show whether there are, somehow, *significant* differences (Konig et al. 2004; Kuiper et al. 2001, 2002). The conclusions drawn from statistical results (for example deciding an appropriate confidence level) can be somewhat arbitrary and context-dependent as different experts may give different interpretations of the degree of significance. Importantly, there can also be a gap between the mathematical significance and the biological relevance of the

D. P. Enot (✉) · J. Draper
Institute of Biological Sciences, University of Wales, Aberystwyth, Aberystwyth SY23 3DA, UK
e-mail: dle@aber.ac.uk

effect detected. Although a central concept in food safety assessment, when testing for substantial equivalence, no precise definition of distance between a GM line and a progenitor cultivar or a pool of reference lines has been adopted. In previous work (Catchpole et al. 2005; Enot et al. 2006) and in the accompanying paper (Enot, Beckmann and Draper) we have described the use of several data analysis methods (Linear Discriminant Analysis, Support Vector Machine and Random Forest) to classify plant and food raw material samples. Focusing specifically on well characterized *Arabidopsis* mutants affected in metabolism the present study extends discussion on approaches to model validation and highlights statistical metrics that may have value in future compositional comparisons in the context of substantial equivalence.

Due to the nature of both biological questions and data characteristics, multivariate analysis is intimately linked to metabolomics and finds a wide range of application with objectives ranging from first pass tools to explore and summarise the information content to data diagnostics or construction of predictive models. When dealing with omics data, a rather pragmatic approach for data analysis is necessary because of several structural characteristics: (1) An optimum (or indeed even adequate) sample size commensurate with the problem complexity or number of variables measured will probably never be met; (2) Experimental variability will always be a major element; (3) Unlike transcriptomics data or targeted metabolite profiling, there is no systematic identity attribution of each signal measured in a metabolomics fingerprint so that noise or instrument artefacts might also enter the analysis; (4) Metabolite fingerprint coverage of the total metabolome reflects the extraction technique and thus will always present incomplete view of any biological system, hence, there is no *a priori* guarantee that the problem at hand can be solved. As a result, interpretation of multivariate models must be treated with caution with the help of adequate statistical techniques and advanced machine learning strategies in order to cope with the complexity of the metabolomics data. Caveats and pitfalls regarding the analysis of omics data have been fairly well covered in the literature (Berrar et al. 2006; Broadhurst and Kell 2006; Dıaz-Uriarte 2005; Somorjai et al. 2003). In the context of substantial equivalence testing it should be stressed that *a priori* it is expected that many genetic modifications will result in changes to the levels of specific metabolites if the transgenes concerned encode for enzymes or regulators of gene activity. Thus, with this in mind it is essential that data analysis methods are used in which individual variables are explicitly highlighted in multivariate models in order to judge whether such 'explanatory' features are associated with changes to predicted areas of biochemistry (Catchpole et al. 2005; Broadhurst and Kell 2006).

2 Testing for similarity by understanding dissimilarities

In the growing number of reported studies, multivariate analysis modelling has played a central role to derive conclusions regarding differences between progenitor and transgenic lines. Principal Component Analysis (PCA) is probably the most widely used technique to approach this problem (Baker et al. 2006; Le Gall et al. 2003; Manetti et al. 2004, 2006). In an unsupervised fashion, samples are typically mapped onto lower dimensions corresponding to the main axes of variation (PCs) and if the GM samples cluster with the parent type, it can be concluded that the GM variety is substantially equivalent to the parent, if not, it is not substantially equivalent. Despite the fact that the reader can appreciate that the genetic modification is apparent on one of the main vectors of variation, there are no metrics to describe how much further away the GM examples are away from their parent and consequently it is difficult to relate the significance of such results with other similar experiments using, for example, another GM line or a different analytical platform. In addition to the inherent difficulty of defining a general measure to decide on the quality of the clustering output, it is quite unusual for unsupervised techniques on their own to rediscover original groups in situations either dominated by noise, suffering from the curse of dimensionality or involving complex experimental design (Jain et al. 1999). Finally, unsupervised approaches do not exploit an important piece of information: after all, we know which are the GM plants and we may also want to discover *unexpected* effects. In contrast, supervised techniques use this essential piece of information (i.e. class label). For classification problems, supervised learning techniques (Hastie et al. 2001) are a class of machine learning algorithms that aim to connect pairs of input vector (e.g. fingerprint matrix) and class labels both constituting the so called training data. Ultimately, the objective of the supervised classifier is to be able to predict as accurately as possible, the class value of any valid input vector previously not seen (generalisability). The misclassification error (or its complement classification accuracy) can be used to derive conclusions relating to the generalization ability of the model and offers direct and general metrics to define distances between genotypes as the ability to discriminate classes is clearly linked to the underlying similarity of behavior within the classes (Braga-Neto and Dougherty 2005). However, an accurate estimation of the true error rate is almost always impractical unless large number of samples are available, which is very often not the case. In such situations, classification accuracy on an independent test set may not reflect subtleties of class difference, specifically when data show inherited variability.

3 Towards appropriate definitions of class separability measures

The characteristics of the underlying probability distribution of the data and more precisely, an examination of class complexity in the original input space must also be considered in conjunction to the overall predictive power of any model (Singh 2003). However for most supervised learning techniques applied to high dimensional problems, the decision boundary properties cannot directly be used because it usually involves an optimisation process that tends to overfit the training data (hence it is necessary to use external samples to assess robustness of the model rather than the re-substitution error). When using projection based techniques such as PLS-DA or PC-DFA (Manly 2004; Massart 1988), additional estimation of the number of components to be used in the modelling process has to be carried out, which might affect discrimination. For these reasons, we propose different approaches to test substantial equivalence in the original multivariate space (i.e. without prior feature selection), with an optimal use of the available information under realistic data paucity constraints and which can provide general and meaningful metrics for future comparisons.

3.1 Classifier error estimation

Estimation of classifier accuracy has received considerable interest in the bioinformatics community and particularly in the omics literature (Braga-Neto and Dougherty 2004, 2005; Fu et al. 2005; Lyons-Weiler et al. 2005). Strategies recommended for transcriptomics experiments can be directly applied in a metabolomics context as both data structures share similar characteristics. In problems related to small sample size, validation approaches are based on repetitive sub-sampling of the original training data to build the model and then use some averaging of the left-out examples to estimate the classifier error. Amongst different resampling techniques, bootstrap based techniques are known to offer an adequate trade off between bias and variance of the error estimation (Efron 1983; Hastie et al. 2001). The general idea is that the variance of an estimate based on the left-out samples is a good approximation of the true variance of the original population. Although bootstrap is widely used to assess statistical accuracy in data mining and machine learning experiments it will often display a bias towards upward accuracy (Efron 1983). To minimize this problem of overfitting, Efron and Tibshirani (1997) proposed a '.632+' estimator, B632+, designed to be a less-biased compromise between upward and downward accuracy estimation. B632, utilizes a resubsitution error (i.e fitting the training data) which is added to correct the bias inherent to the error computed by weighted average of the left out samples (bootstrap zero estimator). Despite its computational cost, bootstrap error provides a less variable estimate than an error counting only techniques (e.g. cross validation) where possible misclassification changes are increments corresponding to the inverse of total number of samples (Braga-Neto and Dougherty 2005).

3.2 Receiver operating characteristic

Receiver operating characteristic (ROC) curves can be used as an alternative measure of the predictive abilities of any binary classifier (Fawcett 2003; Sing et al. 2005). ROC curves display the relationship between sensitivity (true-positive rate) and specificity (false-positive rate) across all possible threshold values that define the decision boundary. The most common way to summarize the ROC curve is to compute the area under the curve (AUC). As a single value measure, the AUC specifies the probability that the decision boundary assigns a higher value to a positive sample than a negative one, both chosen randomly. AUC takes a value of 0.5 when the samples from both classes are uniformly distributed across the decision boundary and a value of 1 when the decision boundary can incontestably discriminate both groups. One advantage of both accuracy and AUC is that they can be calculated regardless of the algorithm used for data analysis.

3.3 Eigenvalues

Linear Discriminant Analysis (LDA; also known as Discriminant Function Analysis in the metabolomics literature) is a supervised method that computes new directions (canonical variates or discriminant functions) in which the groups are best separated (Manly 2004). The aim of LDA is to find discriminant functions that maximise between-class separability (S_B) and minimise within-class variability (S_W). The eigenvalue of the eigensystem of ($S^{-1}{}_W\, S_B$) can be used to measure similarity of both sample replicates, and, more importantly, different classes: the greater the distance between classes (large S_B) and the more compact the classes are (small S_W), the better the classes are separable. However, due to sample size problems (as are common in metabolomics), S_W is always singular and/or unstable and the inversion of S_W cannot be possible without initial data dimensionality reduction using for example PCA (Martinez and Kak 2001; Yang and Yang 2003). To avoid bias in the eigenvalue estimation introduced by the selection of the number of components, we chose the two step procedure proposed by (Thomaz et al. 2004).

3.4 Margin concept in ensemble methods

As opposed to a single model that aims to find the best hypothesis, ensemble methods are techniques that generate multiple models by running a base algorithm many times

(Dietterich 2000; Windeatt 2003). One example is Random Forest (RF) that uses the standard decision tree algorithms as the base learner (Breiman 2001). Prediction of new samples is done commonly by determining the winner class from the votes on the overall ensemble of models. Therefore, confidence in attributing a sample to a designated class can be deduced from the difference between the score (averaged number of votes) for the true class and the largest score of the rest of the classes. This is defined as the sample margin and measures the extent to which the average number of votes for the right class exceeds the average vote for any other class (i.e. the most probable misclassification). The larger the margin, the higher is the confidence that an example belongs to the actual class. In the present study, the margin of a classifier is the mean of all of margins calculated using training data in the ensemble of RF models comparing classes.

4 Model validation in a large scale comparison of *Arabidopsis* mutants

To illustrate the behavior of the statistical measures outlined above (i.e. 0.632 bootstrap accuracy, eigenvalue, AUC and margin), an initial validation was conducted on a heterogeneous set of 27 *Arabidopsis* genotypes providing wide coverage of modelling situations ranging from comparisons with no/little metabolic differences, single gene mutation/insertion effects on isolated biochemical pathways to huge ecotype divergences (Enot et al. 2006). To simulate a situation with a realistic sample size, each measure was calculated on a training set comprising 18 plant replicates per genotype whereas 12 different plant replicates were left out of the training data to provide a validation set. Features of different statistical similarity measures are illustrated in Fig. 1 using all possible 351 pairwise comparisons between the 27 *Arabidopsis* genotypes. Both RF margins and LDA eigenvalues calculated on the training data provide sufficient information regarding the generalisability and the sensitivity/specificity relationships of the models (Fig. 1a, b). In the light of these results, several remarks can be made. Although LDA and RF algorithms differ by the principles by which they operate, voting confidence in RF is higher when the between-class/within-class ratio is maximised (Fig. 1c). One of the claims regarding the fact that RF does not overfit (see Enot et al. 2006) is apparent in Fig. 1d where there is a direct correlation between the confidence in the prediction votes calculated on the training and those of the unseen data. It is thus concluded that RF margin calculations and eigenvalues are able to distinguish between models for which both bootstrap accuracy and AUC are reaching their maximal value of 1. This property is of particular interest in studies where meaningful relative distances between genotypes are sought (Enot et al. 2006).

5 Baselines for model and biological significance

It is difficult to make meaningful statements concerning the extent of compositional similarity between genotypes as it will always be possible to discriminate between any GM

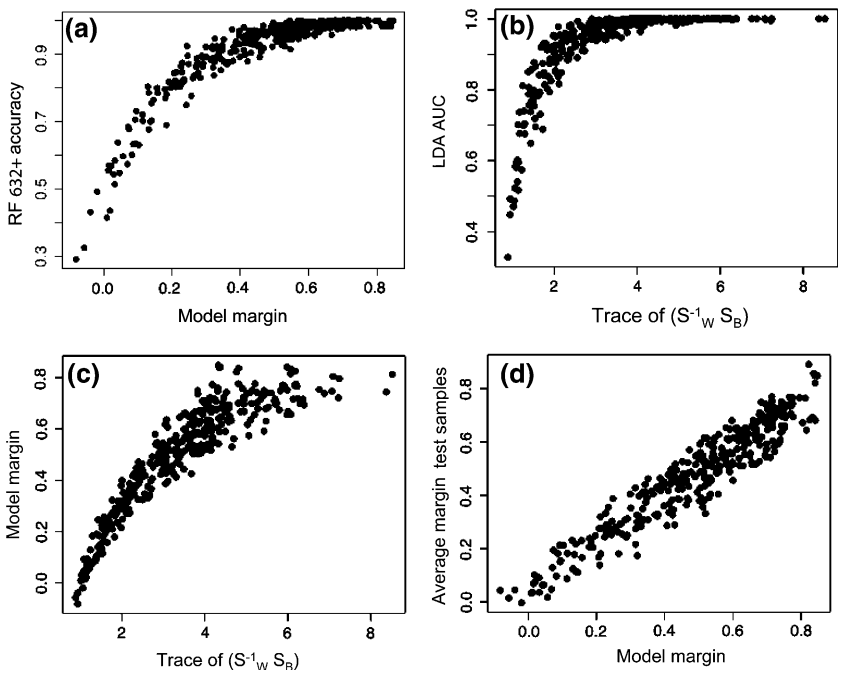

Fig. 1 Relationships between various statistical measures computed from 351 binary classifiers: (**a**) average margin of the training samples versus the *0.632+* bootstrap accuracy in RF models; (**b**) eigenvalue versus area under curve of the LDA model; (**c**) eigenvalue of the LDA classifier versus average RF margin of the training samples; (**d**) average RF margin of the training samples versus average RF margin of the unseen samples

and its progenitor in situations where the GM variety has been intentionally biochemically modified by expression of a transgene. It is therefore preferable to define the presentation of the results in a way that suits the context of the modeling experiment and attempt to match any significance measures against any predicted expectations from a biological perspective. Despite its central role in any modeling experiment, a threshold for determining model acceptance based on its robustness and generalisibility is hardly ever discussed in reports describing 'omics' data analysis. To define a baseline for model significance, one must formulate the behavior of the model properties using a form of null hypothesis stating that the distance measure is not relevant to the biological problem. This approach is quite straightforward when the quantity under study is known to satisfy a particular statistical distribution. An alternative solution is to conduct permutation tests to determine how far from chance is the actual quantity obtained (Good 2000; Lyons-Weiler et al. 2005). This is a type of significance testing in which the distribution of the statistics under study is obtained by computing all its possible values by rearrangements of the sample labels.

To determine a practical baseline for classifier significance one appropriate solution consists of investigating the properties of models that are known to carry few or no relevant biological differences and compare to classifiers discriminating genotypes with distinctive metabolic alterations (Bickel 2004; Tan et al. 2006). To illustrate these points comparisons have been made (using RF and SVM modeling) between four *Arabidopsis thaliana* mutants: *pgm1* (deficient in the isozyme phosphoglucomutase required for starch synthesis), *fah1-2* (mutation of ferulate-5-hydroxylase, an enzyme functioning in cell wall metabolism), *vtc1* (mutation of GDP-mannose pyrophosphorylase in the ascorbate synthesis pathway) and *amt14* (ammonium transporter defective mutant) and their progenitor line Columbia (*Col2*) are extracted from the overall data. In a previous study (Enot et al. 2006) we have already demonstrated that *amt14* did not have any appreciable phenotypic differences when compared to the wild types line, presumably as this particular mutation is compensated for by other functional ammonium transporters. Thus any statistical measures derived from classifiers attempting to discriminate the two lines will represent a threshold for significance. In contrast *pgm-1* had large, pleotrophic phenotypic alterations, whilst *fah1-2* and *vtc-1* displayed distinctive, but more contained, metabolic differences from the progenitor ecotype.

Receiver operating characteristic curves derived from resampling the SVM models of the four comparisons are gathered in Fig. 2. Whilst, the RF model margins of each comparison are given alongside the distribution of the null hypothesis obtained by a permutation test in Fig. 3. It can

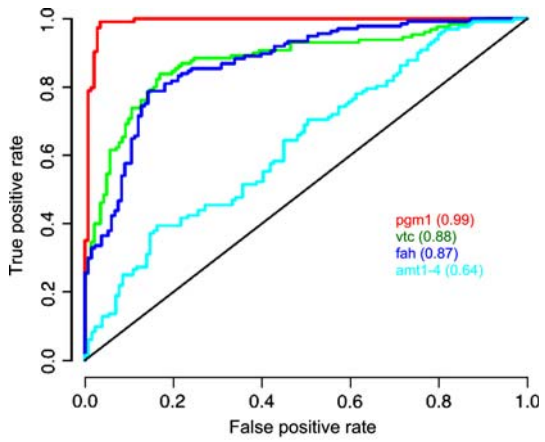

Fig. 2 Receiver operator characteristics curve (ROC) for a selected number of linear Support Vector Machines (SVM) models. Area under the curve is given alongside the genotype name. The diagonal line is the ROC curve equivalent to random guesses

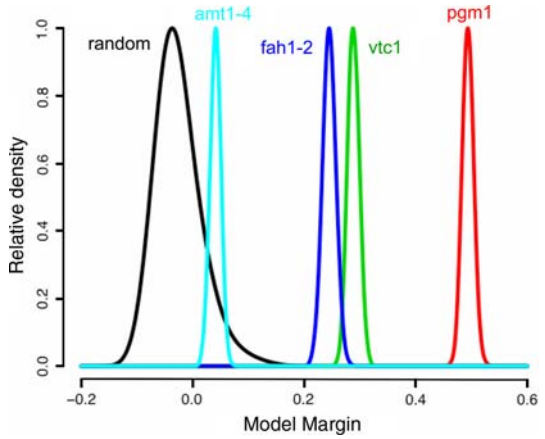

Fig. 3 Illustration of significance testing of the RF margin by class label permutation

be seen in both figures that the model statistical measures are ordered according to the perturbation level expected in the biological outcome: thus the *pgm1-Col2* comparison exhibits logically the highest 'dissimilarity' statistics as this mutant is effectively starch deficient and transcriptomic studies have highlighted expression increases in over 4000 genes. *Vtc*1 or *fah1*-2 represent mutations in two discrete secondary metabolic pathways and as such relatively fewer metabolome differences are detectable by metabolome fingerprinting, resulting in slightly weaker, but none the less significant, modeling statistics. In all three of these comparisons, the statistical measures are higher than the *null hypothesis* comparison between *Col-2* and *amt1-4* in which the ammonium salt transporter mutation did not affect the measured metabolome (Enot et al. 2006). In the comparison of *Col-2* and *amtl-4* the RF model margins overlap significantly the reference distribution

(labeled random in Fig. 3) obtained by permuting the class labels and the area under the ROC curve (AUC = 0.64) is below the accepted limit of 0.8 for defining a good model. In addition, to relate to models in a more meaningful manner, this simple example demonstrates that the statistical "zero difference" (i.e. AUC $\approx$ 0.5, margin $\approx$ 0) regardless of sample size can be reformulated within a biological context without major mathematical redefinition of the problem under study (Bickel 2004).

6 Concluding remarks

Although metabolomics level analytical chemistry tools for assessing substantial equivalence between GM and progenitor genotypes exist, there have been few attempts to explore what kinds of statistical metrics are suitable to quantify compositional similarity. One caveat to such studies is that when transgenes code for novel enzymic activity some differences in the levels of specific metabolites should be expected. Thus to ask questions regarding substantial equivalence, it is important to use data mining approaches which display any discriminatory variables in a discrete and explicit way in order to determine which areas of metabolism are effected. The effect of data dimensionality, often combined with sample paucity, can generate problems for model validation; classifier accuracy alone can be misleading unless validated by sufficient resampling of the available data and AUC assessments. We do not advocate that "one data mining technique fits all" because of the variety of applications and biological questions addressed by a metabolomics approach. A range of significance metrics are important to evaluate in order to decide whether a model is worth pursuing and in the future there is a need for standardised metrics so that everyone can compare results. Currently we suggest that margin measures and scatter matrices eigenvalues in conjunction with estimates of classification accuracy and model sensitivity provide complimentary and appropriate metrics in any specific compositional comparisons.

7 Appendix

Generation of the biological materials and metabolome FIE-MS fingerprint data is described in (Enot et al. 2006). Only results from the ESI positive ion mode data are presented here. All calculations were carried out in the R environment (R 2.4.0) on a PowerPC G5 (dual 1.8 GHz, 2GB SDRAM). Linear Discriminant Analysis was implemented in R according to (Thomaz et al. 2004). Three additional R packages randomForest (Liaw and Wiener 2002), ROCR (Sing et al. 2005), e1071 were used to perform RF, SVM and ROC analyses. 20 bootstraps of training data were employed to compute the .632+ accuracy and area under curve. Note that identical partitioning was executed to allow direct comparisons between RF, LDA and SVM statistics. 2000 permutations of the class labels were performed to get an estimate of the reliability of the RF average margins and LDA eigenvalues. Data, scripts and complete results can be made available upon request from the authors.

References

Baker, J. M., Hawkins, N. D., Ward, J. L., Lovegrove, A., Napier, J. A., Shewry, P. R., & Beale, M. H. (2006). A metabolomic study of substantial equivalence of field-grown genetically modified wheat. *Plant Biotechnology Journal, 4*, 381–392.

Berrar, D., Bradbury, I., & Dubitzky, W. (2006). Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics, 22*, 1245–1250.

Bickel, D. R. (2004). Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics, 20*, 682–688.

Braga-Neto, U. M., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics, 20*, 374–380.

Braga-Neto, U., & Dougherty, E. R. (2005). Exact performance of error estimators for discrete classifiers. *Pattern Recognition 38*, 1799–1814.

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*, 171–196.

Catchpole, G. S., Beckmann, M., Enot, D. P., Mondhe, M., Zywicki, B., Taylor, J., Hardy, N., Smith, A., King, R. D., Kell, D. B., Fiehn, O., & Draper, J. (2005). Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 14458–14462.

Charlton, A., Allnutt, T., Holmes, S., Chisholm, J., Bean, S., Ellis, N., Mullineaux, P., & Oehlschlager, S. (2004). NMR profiling of transgenic peas. *Plant Biotechnology Journal, 2*, 27–35.

Choi, H. K., Choi, Y. H., Verberne, M., Lefeber, A. W., Erkelens, C., & Verpoorte, R. (2004). Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique. *Phytochemistry, 65*, 857–864.

Cockburn, A. (2002). Assuring the safety of genetically modified (GM) foods: the importance of an holistic, integrative approach. *Journal of Biotechnology, 98*, 79–106.

Dıaz-Uriarte, R. (2005). Supervised methods with genomic data: A review and cautionary view. *Data analysis and visualization in genomics and proteomics* (pp. 193–214). New York: Wiley.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science, 1857*, 1–15.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78*, 316–331.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association, 92*, 548–560.

Enot, D. P., Beckmann, M., Overy, D., & Draper, J. (2006). Predicting interpretability of metabolome models based on

behavior, putative identity, and biological relevance of explanatory signals. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 14865–14870.

Fawcett, T. (2003). ROC Graphs: Notes and practical considerations for data mining researchers. *HP Laboratories technical report.*

Fu, W. J., Carroll, R. J., & Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. *Bioinformatics, 21*, 1979–1986.

Fukusaki, E., & Kobayashi, A. (2005). Plant metabolomics: Potential for practical operation. *Journal of Bioscience Bioengineering, 100*, 347–354.

Garratt, L. C., Linforth, R., Taylor, A. J., Lowe, K. C., Power, J. B., & Davey, M. R. (2005). Metabolite fingerprinting in transgenic lettuce. *Plant Biotechnology Journal, 3*, 165–174.

Good, P. (2000). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer series in statistics.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR), 31*, 264–323.

Konig, A., Cockburn, A., Crevel, R. W., Debruyne, E., Grafstroem, R., Hammerling, U., Kimber, I., Knudsen, I., Kuiper, H. A., Peijnenburg, A. A., Penninks, A. H., Poulsen, M., Schauzu, M., & Wal, J. M. (2004). Assessment of the safety of foods derived from genetically modified (GM) crops. *Food and Chemical Toxicology, 42*, 1047–1088.

Kuiper, H. A., Kleter, G. A., Noteborn, H. P., & Kok, E. J. (2001). Assessment of the food safety issues related to genetically modified foods. *The Plant Journal, 27*, 503–528.

Kuiper, H. A., Kleter, G. A., Noteborn, H. P., & Kok, E. J. (2002). Substantial equivalence–an appropriate paradigm for the safety assessment of genetically modified foods? *Toxicology, 181–182*, 427–431.

Kuiper, H. A., Kok, E. J., & Engel, K. H. (2003). Exploitation of molecular profiling techniques for GM food safety assessment. *Current Opinion in Biotechnology, 14*, 238–243.

Le Gall, G., Colquhoun, I. J., Davis, A. L., Collins, G. J., & Verhoeyen, M. E. (2003). Metabolite profiling of tomato (*Lycopersicon esculentum*) using 1H NMR spectroscopy as a tool to detect potential unintended effects following a genetic modification. *Journal of Agricultural and Food Chemistry, 51*, 2447–2456.

Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*, 18–22.

Lyons-Weiler, J., Pelikan, R., Zeh Iii H. J., Whitcomb, D. C., Malehorn, D. E., Bigbee, W. L., & Hauskrecht, M. (2005). Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles, and a prescription for random sampling repeated studies for massive high-throughput genomic and proteomic studies. *Cancer Informatics, 1*, 53–77.

Manetti, C., Bianchetti, C., Bizzarri, M., Casciani, L., Castro, C., D'Ascenzo, G., Delfini, M., Di Cocco, M. E., Lagana, A., Miccheli, A., Motto, M., & Conti, F. (2004). NMR-based metabonomic study of transgenic maize. *Phytochemistry, 65*, 3187–3198.

Manetti, C., Bianchetti, C., Casciani, L., Castro, C., Di Cocco, M. E., Miccheli, A., Motto, M., & Conti, F. (2006). A metabonomic study of transgenic maize (*Zea mays*) seeds revealed variations in osmolytes and branched amino acids. *Journal of Experimental Botany, 57*, 2613–2625.

Manly, B. F. J. (2004). *Multivariate statistical methods: A primer*. Chapman & Hall/CRC.

Martinez, A. M., & Kak, A. C. (2001). PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 228–233.

Massart, D. L. (1988). *Chemometrics*. Amsterdam: Elsevier.

Mattoo, A. K., Sobolev, A. P., Neelam, A., Goyal, R. K., Handa, A. K., & Segre, A. L. (2006). Nuclear magnetic resonance spectroscopy-based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen–carbon interactions. *Plant Physiology, 142*, 1759–1770.

Shepherd, L. V., McNicol, J. W., Razzo, R., Taylor, M. A., & Davies, H. V. (2006). Assessing the potential for unintended effects in genetically modified potatoes perturbed in metabolic and developmental processes. Targeted analysis of key nutrients and anti-nutrients. *Transgenic Research, 15*, 409–425.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics, 21*, 3940–3941.

Singh, S. (2003). Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*, 1534–1539.

Somorjai, R. L., Dolenko, B., Baumgartner, R. (2003). Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions. *Bioinformatics, 19*, 1484–1491.

Tan, C. S., Ploner, A., Quandt, A., Lehtio, J., & Pawitan, Y. (2006). Finding regions of significance in SELDI measurements for identifying protein biomarkers. *Bioinformatics, 22*, 1515–1523.

Thomaz, C. E., Boardman, J. P., Hill, D. L. G., Hajnal, J. V., Edwards, D. D., Rutherford, M. A., Gillies, D. F., & Rueckert, D. (2004). Using a Maximum Uncertainty LDA-Based Approach to Classify and Analyse MR Brain Images. *Lecture Notes In Computer Science, 3216*, 291–300.

Windeatt, T. (2003). Vote counting measures for ensemble classifiers. *Pattern Recognition, 36*, 2743–2756.

Yang, J., & Yang, J. (2003). Why can LDA be performed in PCA transformed space? *Pattern Recognition, 36*, 563–566.