Data analysis standards in metabolomics

Chair: Prof. Roy Goodacre (School of Chemistry, University of Manchester, UK, roy.goodacre@manchester.ac.uk).

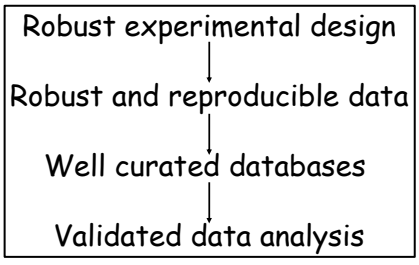
Working group:
Dr J. David Baker (Pfizer, Inc., Ann Arbor, MI, USA, David.Baker@Pfizer.com)
Dr Richard Beger (National Center for Toxicological Research, Jefferson, AR, USA, Richard.Beger@fda.hhs.gov)
Dr David Broadhurst (School of Chemistry, University of Manchester, UK, David.Broadhurst@manchester.ac.uk)
Dr Giorgio Capuani (Chemistry Department, "La Sapienza" University", Rome, Italy, giorgio.capuani@uniroma1.it)
Dr Andrew Craig (BlueGnome LTD, Cambridge, UK, andrew.craig@cambridgebluegnome.com)
Prof Douglas Kell (School of Chemistry, University of Manchester, UK, dbk@manchester.ac.uk)
Dr Bruce Kristal (Department of Neuroscience, Weill Medical College of Cornell University, and Dementia Research Service, Burke Medical Research Institute, USA, kristal@burke.org)
Dr Cesare Manetti (Chemistry Department, "La Sapienza" University", Rome, Italy, manetti@caspur.it)
Dr Jack Newton (Chenomx Inc, Edmonton, Alberta, Canada, jnewton@chenomx.com)
Dr Giovanni Paternostro (Burnham Institute for Medical Research, La Jolla, CA, USA, giovanni@burnham.org).
Prof. Michael Sjöström (University of Umea, Sweden, michael.sjostrom@chem.umu.se)
Prof. Age Smilde (Swammerdam Institute for Life Sciences, Nieuwe Achtergracht 166, 1018 WV Amsterdam, asmilde@science.uva.nl)
Dr Johan Trygg (University of Umea, Sweden, johan.trygg@chem.umu.se)
Dr Florian Wulfert (School of Biosciences, University of Nottingham, UK, Florian.Wulfert@nottingham.ac.uk)

Aims and goals

It is clear that algorithms do not drive metabolomics investigation, but rather the question one seeks to answer with metabolomics influences the data analysis strategy. The goal of this group is to define the reporting requirements associated with statistical and chemometric analysis of metabolite data. This will include identifying the type of algorithm that will be required, and where a model is built, its construction and its validation. These points must be reported so that the data analysis is as objective and unbiased as possible.

Scene setting

The figure opposite identifies the clear flow of information (pipeline) in a typical metabolomics experiment. Whilst multivariate analysis (MVA; also referred to as chemometrics and machine learning) features at the end of the flow, in order for the analysis to be valid there must be robust experimental design. For MVA this particularly refers to the sample type the numbers needed and obviously using the correct control and test groups.



Although experimental data capture and data storage and retrieval are also important, these are dealt with by other working groups.

Design of experiments (DOEs) require that the biological space is adequately populated prior to data capture and subsequent analysis. This is clearly determined by the experiment in question but, for example, if one was interested in the childhood disease leukaemia the control set of healthy individuals must not include adults. Most MVA algorithms are only capable of interpolation, that is to say they give answers within their knowledge realm and can not extrapolate beyond this. Therefore to account for this the DOE would span the metadata that were collected in terms of e.g. sex, age, height, BMI (body mass index) etc, and include suitable sample numbers to account for inherent biological variability. There are approaches to accomplish the former based on space filling algorithms including full or fractional factorial design, Plackett-Burman, Taguchi arrays, to name the most popular ones. The latter requires some preliminary metabolite data collection of the same samples, nominally under identical conditions, where the variation in metabolite data can be assessed in terms of biological reproducibility. Power laws, ANOVA (analysis of variance) and MANOVA (multivariate ANOVA) can then be used to decide on the minimum number of samples required.

Reporting structure:

The number of samples per class should be reported along with the relevant metadata capture, and how accurately these are spanned in the calibration, validation and test sets (*vide infra* for definitions of these data sets).

Pre-processing

Before any analysis is performed metabolite data must be normalised and/or scaled and cleaned up if there is any removable noise or any missing values to be imputed. There are many approaches to normalisation and scaling that can be used and the most popular include scaling to total response, scaling to individual metabolite (or peak), log transformation, scaling to unit variance (autoscale), Pareto scaling, derivatisation, mean centring, vector normalisation. The way in which the data were scaled prior to analysis must be explained. In most instances this will have been optimised, and if this is the case then this must be performed objectively as described under validation below.

At this stage it is useful to draw a distinction between row and column pre-processing operations. Row operations tend to be described as "normalisation" e.g. for a *given sample* dividing each of its feature values by some value such as their sum or mean. Whilst Column operations, tend to be referred to as "scaling", e.g. log scaling, scaling to unit variance and are as such *dependent on all of the samples collected*. This has implications for reporting as row normalisations are sample independent, whilst column scalings are analysis specific, and as such should be reported separately.

For NMR-based metabolomics pre-processing causes dimensionality reduction either by using "spectral binning", where the spectrum line in a bin of a fixed width (typically 0.04 ppm) is integrated and represented as a single variable. Alternatively one may adopt a so-called "targeted profiling" approach where a library of reference spectra is used to fit a NMR spectrum to retrieve actual concentration values for metabolites. In a similar fashion, one must make a choice as to how to process hyphenated data derived from some chromatographic (GC or HPLC) separation prior to mass spectrometry. One can work

directly on the data or after some deconvolution. If deconvolution to metabolite lists is used then these aspects should be detailed in the *chemical analysis reporting structure*.

Reporting structure:

The way in which the data are scaled prior to analysis must be explicitly detailed.

Algorithm selection

The sort of question that one wants to answer drives the selection of the most relevant algorithm (or set of algorithms). It is not feasible to discuss the pros and cons of each method as this is often subjective, but we can define a reporting structure based on the biological application.

Whilst metabolomics experiments do generate multivariate data (*vide infra*) one can employ univariate methods to test for significant metabolites that are increased or decreased between different groups. These include parametric methods for data that are normally distributed, and the most common being ANOVA (analysis of variance), t-tests, z-tests. When normal distribution of data cannot be assumed then non-parametric methods can be used and these include for example Kruskal-Wallis analysis. The significance of these can result in a probability value and data may be visualised directly by using for example box-and-whisker plots.

Multivariate data consist of the results of observations of many different metabolites (variables) for a number of individuals (objects). Each variable may be regarded as constituting a different dimension, such that if there are *n* variables (metabolites) each object may be said to reside at a unique position in an abstract entity referred to as *n*-dimensional hyperspace. This hyperspace is necessarily difficult to visualise, and the underlying theme of multivariate analysis (MVA) is thus *simplification* or dimensionality reduction. This dimensionality reduction occurs in one of two ways; either using an unsupervised or supervised learning algorithms (see the figure below for a summary of the main methods).

Objects (1-*n*) going down in different rows		X-var 1 Metabolite 1	X-var 2 Metabolite 2	X-var 3 Metabolite 3	Y-var 1 First trait to be predicted	Y-var 2 Second trait to be predicted
Sample 1						
Sample *i*, …						
Sample n						

Input data Output data

Unsupervised [use X data only]

- Hierarchical clustering
- Principal components analysis
- Independent components analysis
- Kohonen neural networks

Abbreviations: multilayer perceptrons (MLPs), radial basis functions (RBFs), support vector machine (SVMs), LDA (linear discriminant analysis, PLS (partial least squares), CVA (canonical variates analysis), DFA (discriminant function analysis), MLR (multiple linear regression), PCR (principal components regression), GA (genetic algorithm), genetic programming/computing (GP/GC), evolutionary algorithm (EA), evolutionary programming (EP), classification and regression tree (CART), multivariate adaptive regression splines (MARS).

Supervised [use X & Y data]

- Artificial neural networks
  - MLPs, RBFs, SVMs
- Discriminant analysis
  - LDA, PLS-DA, CVA, DFA
- Regression analysis
  - MLR, PCR, PLS
- Evolutionary-based algorithms
  - GA, GP (GC), EA, EP
- Regression trees
  - CART, MARS, Random Forests
- Inductive logic programming

Unsupervised learning

When learning is unsupervised, the algorithm is shown a set of inputs and then left to *cluster* the metabolite data into groups. For MVA this optimization procedure is usually *simplification* or dimensionality reduction. This means that a large body of metabolite data (*x*-data) are summarised by means of a few parameters with minimal loss of information. The most used approaches are principal components (PCA) and hierarchical cluster analyses (HCA), and after clustering the ordination plots or dendrograms then have to be interpreted. Alternative approaches include nested algorithms (eg, PCA followed by HCA), and SIMCA (soft independent modeling of class analogy) and kNN (k-nearest neighbours).

Reporting structure:

As PCA and HCA are unbiased analysis and describe the natural variation in the input *x*-data what needs to be reported is the percent explained variance generated for each principal component plotted and the specific way in which HCA has been generated. This includes construction of the similarity matrix and whether an agglomerative or divisive clustering algorithm is used.

Supervised learning

When one knows the desired responses (*y*-data, or traits or classes) associated with each of the metabolite data inputs (*x*-data) then the system may be supervised. The goal is to find a mathematical transformation (model) that will correctly associate all or some of the inputs with the target traits. This trait can be categorical (e.g., disease *vs*. healthy) or quantitative

(e.g., grade of cancer, response to therapy). In its conventional form this is achieved by minimising the error between the known target and the model's response (output). In addition there exist special types of supervised learning that effect explanatory analyses; that is to say the mathematical transformation from input to output data is transparent. Such inductive methods allow one to discover which metabolites (inputs) are key for the separation of the traits to be predicted. These approaches may help in the validation of the model in terms of its biological relevance that can be tested by a complementary approach using transcriptomics and proteomics.

Although not prescriptive there is some sort of logical work-flow in supervised data analysis that goes from univariate analyses (ANOVA, t-tests, Kruskal-Wallis, etc) through to relatively simple linear multivariate analyses (CVA, PLS etc) to more complex non-linear multivariate approaches (ANNs, GP, etc). Obviously, the computations get more complex as one progresses through these stages and one should stop once a satisfactory and validate conclusion is made.

Validation: *As these supervised learning methods use both input x-data and output y-data in model formation the analysis must be fully validated*. All these methods require optimisation. For regression based approaches (MLR, PCR, PLS) and discriminant analysis (LDA, CVA and DFA) the number of latent variables that are used in the model must be optimised. For neural networks the optimisation will be in terms of the number of iterations (epochs) in model formation, and for evolutionary-based algorithms the number of population cycles. Whilst for pre-processing, the effect on model performance must also be assessed objectively.

In general, there are two strategies to model validation; (a) leave-*n* out and (b) the use of fixed training and validation sets.

> In (a) leave-one or leave-*n* out for model calibration is where a single or *n* sample(s) are iteratively left out, the model is then reconstructed, the omitted sample projected into model space, and its location used to assess the predictive ability of the model. Once this process is optimised then one should try to use some independent assessment of predictive ability on data that has not been seen by the model.
>
> In (b) the approach is similar to the above but one uses three 'fixed' data sets (these are defined as following as these terms sometimes vary between laboratories):
> *Training set*: refers to the *x*-data and *y*-data pairs used to construct a model.
> *Validation set*: refers to the *x*-data and *y*-data pairs used to validate model construction. This is used during training where the *y*-data predicted and *y*-data known are compared.
> *Test set*: refers to the *x*-data used to test the model. These *x*-data are only used after the model has been constructed with the training and validation data sets.

When supervised analyses are used, or pre-processing optimisation employed, the above validation approach must be conducted and be included in the report.

Finally, there is a choice to be made between building a model for screening; that is to say a robust model where one is not interested in which metabolites are important (so called semi-supervised approach), and a model for bio-marker selection, where the aim is to find the smallest subset of variables that 'adequately' model the hypothesis.

Reporting structure:

The exact details of how the metabolite data were objectively split into training and cross validation sets (either separate data or re-sampling of training data) must be given long with details of any independent test data.

Metric used for choosing the number of latent variables, number of iterations or populations must be given based on the above three data sets.

Software

One for discussion – I have not included any at this stage as we need to be fully inclusive and the analysis strategy is more important than software X *vs*. software Y.

Design of reporting structure

Most data analyses start with the production of the initial data table which will then be analysed. It is worth considering that the reporting structure should be split into two. The first part comprises all the steps which are taken in order to turn the raw analytical data into the initial data table (each row being one sample and each column being one feature) and this will likely be derived from the *chemical analysis reporting structure*. The second part should then describe the analysis done on the table, including all pre-processing operations as well as the algorithms/analyses. Dividing the reporting structure in this way allows it to be very modular and flexible.

References

Beavis, R.C., Colby, S.M., Goodacre, R., Harrington, P.B., Reilly, J.P., Sokolow, S. and Wilkerson, C.W. (2000) Artificial intelligence and expert systems in mass spectrometry in Meyers, R.A. (Ed), *Encyclopedia of Analytical Chemistry*. pp. 11558-11597.

Beebe, K.R., Pell, R.J. and Seasholtz, M.B. (1998) *Chemometrics: a practical guide*. Wiley, New York.

Brown, M. Dunn, W.B., Ellis, D.I., Goodacre, R., Handl, J., Knowles, J.D., O'Hagan, S., Spasić, I. & Kell, D.B. (2005) A metabolome pipeline: from concept to data to knowledge. *Metabolomics* **1**, 39-51.

Chatfield, C. and Collins, A.J. (1980) *Introduction to Multivariate Analysis*. Chapman & Hall, London.

Duda, R.O., Hart, P.E. and Stork, D.E. (2001) *Pattern classification, 2nd ed.* John Wiley, London.

Everitt, B.S. (1993) *Cluster Analysis*. Edward Arnold, London.

Goodacre, R., Vaidyanathan, S., Dunn, W.B., Harrigan, G.G. and Kell, D.B. (2004) Metabolomics by numbers - acquiring and understanding global metabolite data. *Trends in Biotechnology* **22,** 245-252.

Hall, R.D. (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytologist* **169,** 453-468.

Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26,** 99-105.

Krzanowski, W.J. (1988) *Principles of Multivariate Analysis: A User's Perspective*. Oxford Univeristy Press, Oxford.

Manly, B.F.J. (1994) *Multivariate Statistical Methods : A Primer*. Chapman & Hall, London.

Martens, H. and Næs, T. (1989) *Multivariate Calibration*. John Wiley, Chichester.

Weckwerth, W. and Morgenthal, K. (2005) Metabolomics:from pattern recognition to biological interpretation. *Drug Discovery Today* **10,** 1551-1558.

ENDS