Contents lists available at ScienceDirect

Progress in Nuclear Magnetic Resonance Spectroscopy

journal homepage: www.elsevier.com/locate/pnmrs

Bioinformatic methods in NMR-based metabolic profiling

Timothy M.D. Ebbels *, Rachel Cavill

Biomolecular Medicine, Division of Surgery, Oncology, Reproductive Biology and Anaesthetics, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 17 June 2009
Accepted 16 July 2009
Available online 22 July 2009

Key words:
Metabonomics
Metabolomics
Metabolic profiling
Bioinformatics
Statistical methods
Modelling
Machine learning
Pattern recognition

Contents

1. Introduction

The last decade has seen a revolution in the practice of biological research, primarily due to the rise of techniques that are able to profile levels of molecular organisation at a global level. This wide coverage of the so-called ‘omics’ techniques has been inspired, and is often made possible by, the completion of genome sequences which allow interpretation of events in terms of the full complement of biomolecules present in a cell or organism. The field of metabolic profiling (also known as metabonomics [1] or metabolomics [2]) studies the myriad small molecular weight

Abbreviations: PCA, principal components analysis; PLS, partial least squares; PLS-DA, partial least squares-discriminant analysis; O-PLS or O2-PLS, orthogonal partial least squares; GA, genetic algorithm; GP, genetic programming; SVM, support vector machine; kNN, k-nearest neighbour; STOCSY, statistical total correlation spectroscopy; SHY, statistical heterospectroscopy; LC, liquid chromatography; JRES, J-resolved; HSQC, heteronuclear single quantum coherence; MS, mass spectrometry; DIGE, differential in gel electrophoresis.

* Corresponding author. Tel.: +44 20 75943160; fax: +44 20 75943226.
E-mail address: t.ebbels@imperial.ac.uk (T.M.D. Ebbels).

doi:10.1016/j.pnmrs.2009.07.003

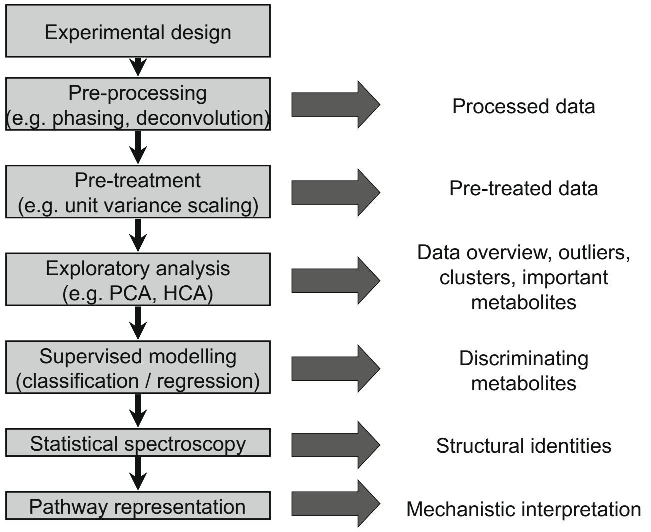
Experimental design
Pre-processing (e.g. phasing, deconvolution)
Processed data
Pre-treatment (e.g. unit variance scaling)
Pre-treated data
Exploratory analysis (e.g. PCA, HCA)
Data overview, outliers, clusters, important metabolites
Supervised modelling (classification / regression)
Discriminating metabolites
Statistical spectroscopy
Structural identities
Pathway representation
Mechanistic interpretation

Fig. 1. A schematic diagram of the steps involved and information recovered in the statistical analysis of NMR metabolic profiles.

molecules involved in metabolism, with a particular emphasis on how the levels of these molecules change in response to different biological conditions. The ultimate goal of this field is to achieve improvements in understanding of metabolic processes that may lead to advances in many areas including clinical diagnosis, therapeutics, functional genomics and toxicology. NMR spectroscopy has been employed as an analytical tool in studies of metabolism for many years [3–6] principally because it is non-selective in the analytes studied and because of the high degree of structural information obtainable in a short time. It is also favoured because it is non-destructive and highly reproducible with minimal sample preparation. The complex biological samples analysed in metabolic profiling studies result in correspondingly complex NMR spectra; consequently the extraction of useful information from such metabolic profiles is a difficult task and this is the subject of this review.

Early studies employing NMR to profile biofluid mixtures [5,7] did not attempt to model the spectra statistically and relied on visual analysis of the spectra for their interpretation. In the early 1990s the potential of statistical pattern recognition techniques to extract useful information from the data became apparent and began to be applied [8–11]. With the advent of other omics techniques in the late 1990s, the statistical analysis of metabolic profiles has become one of several specialities within the field of post-genomic statistical modelling and has received wider attention [12–19]. A distinction is usually made between global profiling, in which as many metabolites as possible are assayed in a single step, and targeted profiling in which a subset of interesting metabolites are identified *a priori* and the analytical methods optimised for measurement of this group. We are also particularly interested in screening approaches in which large numbers of samples are rapidly profiled, necessitating the use of automatic processing algorithms. The goals of statistical modelling in metabolic profiling can be briefly summarised as (1) visualisation of the overall similarities and differences between samples and variables, (2) determination of whether or not there is a significant difference between groups or trends related to the effect of interest, (3) discovery of which metabolic signals are responsible for these patterns, (4) structural characterisation of the metabolites involved, and (5) analysis of metabolic effects at a pathway level. Until recently, aim 4 was addressed exclusively by further experimental analytical procedures, but recent developments in correlation analysis have yielded useful statistical tools [20–24] which can aid this process.

The analysis of NMR metabolic profiles typically involves a number of stages, conceptualised in the flow diagram of Fig. 1. One of the most important stages, but one which is often overlooked, is the initial design of the experiment. Here, an interaction of the data analyst with the other scientists involved (analytical chemists, biological scientists etc.), is essential to ensure that the desired goals are met. For example, it is important to ensure that any extraneous variables are not confounded with the effect of interest (e.g. having systematic differences in the ages of control and treated groups in studies of age related diseases). Once the biological experiment has been run and the NMR spectra obtained, standard data processing techniques are applied to the raw free induction decays (FIDs) to obtain correctly phased, baseline corrected and chemical shift referenced spectra. This process itself is non-trivial when the samples are complex and numerous, but software for automatically processing metabolic profile spectra has matured greatly over the past 10 years and is now generally sufficient for this task. Once the basic spectra are obtained, they must undergo a preprocessing step which transforms the data to a table of N samples (rows) by M variables (columns). Each row will represent the metabolic profile of a particular biological sample and each column represents a given metabolite or metabolic signal. There follows an optional pre-treatment step in which the rows or columns of the table may be rescaled. For example, normalisation procedures scale each row by a factor, for example to account for overall variation in sample dilution. An example column operation is scaling each variable to unit variance. The resulting table forms the data input to the next two stages of analysis – exploratory and predictive modelling. The exploratory stage is characterised by so-called *unsupervised* methods, those in which the algorithm takes no account of *a priori* information about the structure of the data (e.g. clusters, trends etc.) This stage addresses the first goal mentioned above and typically involves visualisation of the overall distribution of samples and variables, and assessment of data quality including detection and removal of outliers. Once the global structure of the data has been assessed in such a way, the analyst may wish to proceed to a predictive modelling stage in which one attempts to build mathematical rules that use the metabolic profile to predict an external response variable, such as biological class. At this stage, *supervised* methods are typically used in which a part of the data (the *training set*) is used to fit the model, and a separate part (the *test* set) is used to estimate the predictive accuracy of the model. The importance of this latter step cannot be overstated. As with all high dimensional data, it is very easy to obtain models appearing to explain the data well, yet which are over-fitted or otherwise not predictive outside the particular sample set for which they have been developed. Thus the process of *model validation* is a critical step in the whole modelling process.

The nature of the metabolome and the NMR methods used in its interrogation give particular characteristics to the data obtained, thus influencing the types of modelling methods used. In contrast to the genome or even the proteome, a large proportion of the metabolome is still unknown, even for model organisms. The degree of incompleteness varies between organisms, biofluids/tissues and conditions and is hard to estimate. This is one reason why it is not easy to convert raw spectral data to concentrations of known metabolites; in a typical NMR profile, a large number of the resonances may be unassigned, particularly for low level or partially resolved signals. Even when resonances are assigned, it is difficult to obtain truly quantitative concentration information in high throughput profiling applications. Some important reasons for this include peak crowding and overlap, changes in chemical shift position of resonances due to differential matrix effects, complexation