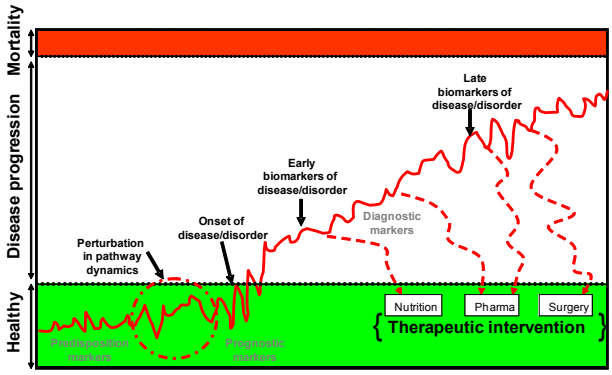
Biomarkers and Biomarker Discovery



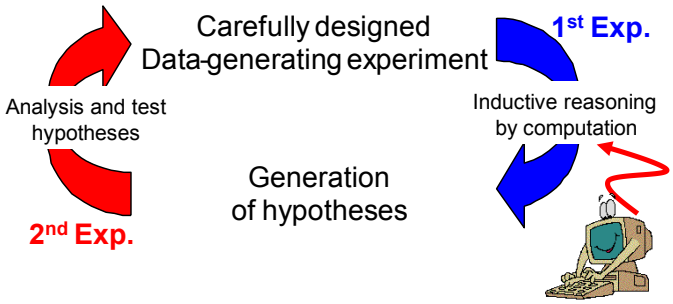
Ellis, D.I. *et al*. (2007) *Pharmacogenomics* **8**, 1243-1266.

The NIH definition of a biomarker is:
"A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention."

You will have heard of some biomarkers as these are what a GP or hospital may use. For example: a fasted glucose test for diabetes; prostate specific antigen (PSA is a protein) for cancer of the prostate; or troponins (proteins) for diagnosing cardiovascular disorders such as myocardial infarction and heart failure; or you may even have had a dipstick test carried out on urine which tests for nitrite which is an indicator of a bacterial urinary tract infection.

These biomarkers are very valuable for diagnosing disease and there is an urgent need to have many more that are predictive with good accuracy. There are many more biomarkers waiting to be discovered and this is a very active area of research in metabolomics where hypothesis generation is performed.

In this approach metabolite profiling is used to generate quantitative lists of metabolites from control populations and test subjects with the disease of interest. Data analysis is then used to mine the metabolites and determine which are discriminatory for the disease and which of these could be used in predictive medicine. The figure illustrates the idea of using two experiments: one to generate the hypothesis of which metabolites are important with respect to disease and experiment two to test this hypothesis.



There are many different multivariate or chemometric analyses methods that could be used and the key objective is to *make the analysis as valid as possible so stringent statistical validation is needed*. It is essential that a good data generating experiment is designed so as to encompass all possible types of variation inherent in these experiments. These include:

(1) At the biological level, this means controlling the diet immediately before sampling and making sure that the control and test subjects are balanced; that is to say there is no bias in the two cohorts so gender, age, BMI ranges etc must be equal in the two populations.

(2) Sample preparation level, both with respect to quenching metabolism and metabolite extractions. This can be tested by performing multiple extracts.

(3) Variation introduced by the analytical instrument itself.

Thus lots of data are needed and, moreover, a good strategy to adopt is to collect three sets of data:

(1) The first set is called the metabolomics 'training data' and is used to construct a mathematical model that will relate metabolite data with disease status; this can be categorical (diseased *versus* healthy) or quantitative (severity of disease or grade of cancer).

(2) The second set of data is usually terms the metabolomics 'validation data' and is used to cross-validate the model generated by the training data; for example, some modelling processes (e.g., discriminant analysis or partial least squares) involve the extraction of a certain number of scores (latent variables), too few are inadequate whilst too many may cause the metabolite data to be over fitted (that is to say the model will include unwanted noise). Thus the second set of validation data are used to tune this selection process.

(3) The final set of data is the metabolomics 'test data' and this is used to assess how well the modelling process described above has done. These data are *independent* and so can be used to assess the predictive nature of the analysis.

For the detection of biomarkers one can of course start off with the easiest method of 'stare and compare!' Whilst unlikely to give rise to any markers it is always wise to actually look at the data as this can also be used as a quality check. The next method is to use difference profiles where the average metabolite profile from the diseased subjects is subtracted from the average metabolite profile from healthy individuals. Finally, univariate analysis of variance (ANOVA), student t-test or non-parametric (i.e., if the data are not normally distributed) equivalents can be used to ascertain is there is any statistically significant differences between individual metabolites for healthy *versus* diseased individuals.

The next stage would be to use some data analysis process that is multivariate. The concept of multivariate biomarker profiles has become reality and so more powerful supervised learning methods are needed. In supervised learning an algorithm is used to transform the multivariate data from metabolite profiles into something of biological interest, usually of much lower dimensionality, which as discussed above can be categorical (diseased *vs*. healthy) or quantitative (severity of disease). In supervised learning both metabolite data (inputs) and disease status (outputs or targets) are used, and these two types of data form pairs which are used in the calibration of the model. The goal of supervised learning is to find a model or mapping that will correctly associate the inputs with the targets. Once calibrated the information on the important features (metabolites) can be extracted.

Below highlights the various algorithms that are currently employed to discover biomarkers using supervised learning:

Discriminant analysis (DA) is a cluster analysis-based method and involves projection of test data into scores space. This is a categorical method and *loadings matrices* can give an indication of important inputs (metabolites).

Partial least squares (PLS) is a very popular linear regression based method. The algorithm can be programmed in a quantitative way (PLS1 for severity of disease) or categorical (PLS-DA for classification of diseased from healthy), and as for DA *loadings matrices* can give an indication of important metabolites.

Classification and regression trees (CART), a type of rule induction algorithm, can also be used. These are categorical algorithms based on the growth of a decision tree, with predictive segregation of the data at the branches of the tree, and the leaves being the classifications (diseased or healthy). These branches are the *decision boundaries which can be used to discover which metabolites are important* for separation.

Machine learning (ML) approaches are also used and some of these are based on concepts of Darwinian selection to generate and to select the most important metabolites that are predictive for the output (disease *vs*. healthy). These evolutionary computation (EC) algorithms are include genetic algorithms (GAs) and genetic programming (GP). GAs give you an indication of which metabolites are important, whilst GPs can give some information on the relationships amongst those metabolites for the classification.

All of these methods generate hypotheses of which metabolites may serve as potential biomarkers. These of course need to be validated by a separated study and ultimately in larger populations.