Evaluation of molecular profiling platforms in clinical pharmacology

Doorn, M.B.A. van

CHAPTER 2

Introduction to molecular profiling platforms

Gene expression profiling using cDNA microarray technology

Gene expression

The complex process of coordinated expression of a large number of genes is a requirement for the maintenance of a proper health status [1]. With little exceptions, all cells of the human body have a full set of chromosomes and harbour identical genes. Of these genes, only a few are turned on and it is the subset that is 'expressed' that gives the distinctive properties to each cell type and thus determines its phenotype and biological function [1].

Gene expression is defined as the transcription of information stored within the DNA into messenger RNA (mRNA) molecules that are subsequently translated into proteins that perform most of the cellular functions [1]. The road from 'gene' to 'function' is controlled at many levels including the transcriptional level (transcription initiation, elongation and termination), the posttranscriptional level (RNA translocation, RNA splicing, RNA stability), the translational level (translation initiation, elongation and termination), and the posttranslational level (protein splicing, translocation, stability and covalent modifications like glycosylation) [2]. Disruptions in any of these steps can result in changes to the physical structure and activity of the various DNA, RNA, and protein complexes, thus leading to an altered health or disease status [2].

Studying gene expression

The type and quantity of mRNA produced by a particular cell or tissue are studied to learn which genes are expressed. This can provide important insights into how the cell or tissue responds to its changing needs in i.e. a disease state or as a result of a (pharmacological) intervention. As indicated above, gene expression is a highly complex and tightly regulated process. It enables cells to react dynamically both to environmental stimuli and to their own changing requirements [1]. This mechanism can act as both an 'on/off' switch to control which genes are expressed as well as a 'volume control' that increases or decreases the level of expression of particular genes as necessary [1]. Global gene expression levels can be measured using a wide variety of low and high throughput technologies. Using traditional low throughput methods (i.e. northern blotting [3] or real-time (quantitative) polymerase chain reaction (RT-PCR) analysis [4]) to study gene expression, one can only assess a relatively small number of genes (usually less than one hundred) at a time. In addition, these methods are considered to be very labour- and time-intensive [5;6]. In contrast, newly emerged high throughput technologies allow the rapid analysis of the (semi-quantitative) expression of multiple genes in a single experiment [7-9]. These technologies can be classified into two categories: 'open' and 'closed' systems [10]. Open systems do not require any a priori knowledge of the sequence of the gene or genome being studied. Closed systems, however, always require some knowledge of the gene sequences [2]. Examples of an open system include serial analysis of gene expression (SAGE) [11] and massively parallel signature-sequencing (MPSS) [12]. An example of a closed system is the DNA microarray (or GeneChip™), which is currently by far the most widely employed technique for global gene expression profiling [10].

Microarray technology

Microarray technology, developed in parallel by Southern et al. [13] and Fodor et al. [14] was pioneered at the Brown lab at Stanford University [8], and is currently the most widely applied high throughput technology that allows the rapid analysis of global gene expression profiles [7-9]. Although it has proven to be valuable in many areas of biomedical research (see further), the effective application of this technology is reliant on the required level of quantification needed to answer the scientific question posed by the researcher.

The most straightforward microarray experiment design compares two distinct biological conditions, i.e. 'disease' state *vs.* 'healthy' state. Genes that are up- or down regulated in the disease state are more likely than random chance to be potential drug targets for that particular disease and can collectively (i.e. the expression profile) offer a detailed molecular disease 'fingerprint' [15].

In recent years, microarray technology has been employed to study basic biological processes, such as the temporal changes in gene expression during the cell cycle and has shed more light on the underlying mechanisms of both physiological and pathophysiological processes, including aging, tumorigenesis, rheumatoid arthritis, and inflammatory bowel disease [16-19]. Furthermore, the technology has shown to be quite useful in clinical oncology research where methods for molecular diagnostics [20;21] (classification of tumours according to gene expression profile) and prediction [22] (prediction of patient

prognosis based on gene expression profile) have made their first steps towards a clinically applicable tool. Moreover, the Oncotype Dx tex [23], which measures the tumour expression of 21 genes, can predict how likely certain breast cancers are to recur and respond to pharmacological treatment and similar Mammaprint®test [24] (based on the expression of 70 genes), were recently marketed in the United States. In addition, the use of microarray technology is promising for toxicology assessments [25] as well as for monitoring the effects (i.e. mode of action prediction) of *in vitro* [26] and *in vivo* [27-29] pharmacological interventions, and may therefore have considerable impact on the process of drug development [30-32].

Besides being useful tools for global gene expression profiling, microarrays can be employed in high throughput genotyping, i.e. single nucleotide polymorphism (SNP) detection [33]. Recently, the US Food and Drug Administration approved the first DNA microarray test, the AmpliChip cytochrome P450 Genotyping Test (Roche Molecular Systems) for use as a diagnostic [34]. This test can be used to predict how quickly a certain individual will metabolize a range of drugs, thereby assisting physicians to select the optimal drug and dosage. Another interesting application was reported by Joseph DeRisi, who used a microarray to identify the SARS virus [35]. Finally, new applications of the technology are continuously emerging, such as genome-wide epigenetic analysis and on-chip synthesis [36].

Gene expression profiling using DNA microarrays

DNA microarrays are small, solid supports (usually glass microscope slides, but can also be silicon chips or nylon membranes) onto which the sequences from thousands of different genes (i.e. probes) are immobilized, or attached, at pre-determined locations [1]. The DNA is printed, spotted, or actually synthesized directly onto the support. Key to the microarray system is that the gene sequences are fixed to their support in an orderly fashion, since the position of each spot in the array is used to identify a specific gene sequence [1]. The probes can be DNA, complementary DNA (cDNA), or oligonucleotides. An oligonucleotide is a short fragment of a single-stranded DNA that is typically 5 to 50 nucleotides long [1;5].

The basic principle of DNA microarrays is complementary base pairing. Complementary nucleotide strands interact non-covalently allowing subsequent detection [37;38]. By the application of microarrays the contemporary gene expression status of a cell or tissue is used to generate a molecular 'fingerprint' [37]. Consequently, it is possible to relate this fingerprint to distinct time points, disease states or any type of intervention. However, because of the large number of data points generated, and the required data normalisation procedures, this processing requires specific, sophisticated statistical algorithms [39-45]. A schematic overview of the steps involved in DNA microarray-based gene expression profiling is provided in figure 1 (p. 202).

Bioinformatics, data management, and functional annotation

Microarray data is typically characterized by many measurements per patient sample [37;46]. Interpretation is quite difficult, since the number of genes vastly surpasses the number of samples and the data correlation structure is poorly understood [37]. The key challenge is to make a distinction between random ('noise') and significant patterns of gene expression [37;46].

Several critical steps need to be taken before the raw data are ready for biological interpretation. After quality control and hybridization of the RNA sample to the microarray, image processing, translation of images into signal intensities and normalization, the data are finally ready for data mining [37;44].

Data mining

Data mining, defined as the discovery of non-obvious information in the data set, often uses mathematical techniques that have traditionally been used to identify patterns in complex data [37;46]. Recently, these techniques have been adapted to suit the needs of gene expression (i.e. 'transcriptome') analyses [37]. In essence, there are two different approaches to analyze the data, i.e. the *unsupervised* and the *supervised* approach [47]. An unsupervised analysis does not use any pre-defined class definition, and simply seeks to determine what structure is inherent in the data. In contrast, a supervised analysis aims at finding putative associations between genes or gene profiles and a pre-defined class.

Unsupervised analysis

Unsupervised analysis of microarray data seeks to identify (sub)groups of samples or outliers without using information of the clinical features of each sample [37]. For instance new subgroups of a disease with a

distinct 'molecular signature' can be identified [48]. A commonly used unsupervised method is hierarchical clustering [49]. Although this type of analysis is quite useful for data visualization, it has a number of drawbacks including the lack of statistically valid quantitative information about the genes that are differentially expressed between classes [50]. In addition, this method does not enable the investigator to deal with multiple comparison issues in a statistically valid manner. Consequently, cluster analyses are more appropriate for 'class discovery', but are often less suitable for 'class comparison' or 'class prediction' [51].

Supervised analysis

Typically, it is of great interest to correlate microarray data directly with clinical data [37]. In supervised analyses the classification of each sample is used (i.e. diseased *vs.* healthy or active treated *vs.* placebo treated). The main objective of this evaluation is to identify significant differentially expressed genes between two classes i.e. to perform 'class comparison' [51]. There are a large number of parametric and nonparametric tests available to assess gene significance in the analysis [52-58]. However, because of the vast number of genes, results should be adjusted for multiple hypothesis testing to exclude random patterns [51;59]. A commonly used approach is the estimation of the 'false discovery rate' (FDR). This rate will provide an indication of the number of false positive genes that are present in a given list of differentially expressed genes between two classes [60]. The desired FDR can be determined by the investigator and is featured in i.e. the Significance Analysis of Microarrays (SAM) [61;62] and BrBarraytools [63] software applications. In each type of supervised analysis, a list of differentially expressed genes associated with the (clinical) response variable is generated. Ultimately, the number of genes in this list is reliant on the (arbitrary) significance level set by the investigator.

Functional annotation

Functional clusters of genes (based on their similarity among gene expression profiles) derived from i.e. hierarchical clustering analyses, can be identified using detailed gene annotation [37]. This information can be used to provide new insights into (patho)physiological pathways and may also provide a simple means of gaining leads to the functional relationship of many genes for which detailed information is currently unavailable [49]. A structured annotation of genes and gene products, essential for all biological interpretations, is available from the Gene Ontology consortium [64].

Furthermore, relevant gene information can be found in various other (freely internet accessible) databases such as the Online Mendelian Inheritance in Man (OMIM; more clinically oriented) [65], UniProtKB/Swiss-Prot Protein Knowledgebase [66], NCBI Genbank [67] and EBI Ensembl [68] thereby substantiating the knowledge about a particular gene and its gene product and thus aid in the biological interpretation of microarray experiment results.

Molecular network and pathway analysis

To evaluate the role of significant differentially expressed genes in the pathogenesis of a certain disease or in response to a certain intervention, the question arises whether some of the genes are interconnected and share biological functions or are involved in common pathways [37]. There are several collections of (metabolic) pathways that can help to answer these questions. Valuable information can for instance be found in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [69] and Encyclopaedia of Metabolic Pathways (METACYC) [70]. In addition, biological networks can be generated (and pathways visualised) through the use of various commercial and non-commercial applications [71]. One of the commercial applications, i.e. Ingenuity Pathway Analysis™ (IPA), is a web-delivered application that enables scientists to discover, visualize and explore therapeutically relevant (gene and protein) networks significant to their experimental results [72]. Since this specific software application was used in the analysis of the microarray studies in this thesis it will be discussed in more detail in the following paragraph.

Ingenuity Pathway Analysis™

To allow for a meaningful pathway analysis, first a sufficient number of genes have to be identified that is differentially expressed between the two defined classes. For generating molecular networks that indicate how these genes may be interrelated, a certain cut-off of (i.e. 5% FDR or p-value < 0.001) can be set. However, since these analyses are used for exploratory purposes, relatively non-stringent cut-off criteria are usually applied. After uploading the expression data file in the IPA application, each probe set is mapped to its corresponding database gene object to

designate so-called 'focus genes' [72]. Focus genes are genes from the analysis input data file that have a level of significance below the chosen cut-off criterion and interact directly with other genes in the IPA global molecular network. To start building the networks, the application queries the IPA Knowledge Base for interactions between focus genes and all other gene objects stored in the knowledge base, and generates a set of networks with a network size of 35 genes / gene products [72]. The application will then calculate a score for each network according to the fit of the user's set of significant genes. This score ('Z-score') is derived from a p-value and random chance. A score of 2 indicates that there is a 1 in 100 chance that the focus genes are together in a network due to chance. Consequently, every score of 2 or above has at least a 99% confidence of not being generated by random chance alone. Biological functions and pathways are then calculated and assigned to each network [72]. Finally, the results are ready for biological interpretation by the investigator.

Limitations of global gene expression profiling

Although microarray-based gene expression profiling offers the advantage of being able to monitor the expression many genes simultaneously, this enormous set of data only provides one aspect of the disease or (pharmacological) intervention being studied [10]. In fact, it is well known that not all biological processes are regulated at the transcriptional level. Proteins are the key players in many processes and post-translational modification (i.e. phosphorylation, prenylation, sulfation etc.) of proteins is also very important for the regulation of many biological processes and pathways [10]. Therefore, a more complete understanding of the disease process or (drug)mechanism being investigated can only be obtained when it is studied at the different molecular levels of mRNA (transcriptomics), protein (proteomics) and metabolite (metabolomics) simultaneously [73]. However, since the field of 'proteomics' is beyond the scope of this thesis it will not be discussed here. The following section provides an overview of the background, measurements, and statistical methods employed in the metabolomics studies of this thesis.

Metabolic profiling using ^{1}H NMR spectroscopy

Introduction

Along with its counterparts in the '-omics' revolution, metabolomics has arisen as an important and gripping discipline in biomedical research [73]. It involves the identification and quantitation of large amounts of metabolites from cells or biofluids and their changes due to physiological and nonphysiological processes [74]. A schematic overview of the steps involved in NMR-based metabolomics studies is provided in figure 2 (p. 202).

Metabolomics is less well-defined than i.e. transcriptomics or proteomics, illustrated by the fact that many synonyms (for virtually the same approach) i.e. 'metabonomics' [75], 'metanomics' [76] and 'metabolite profiling' [77] can be found in the pertaining literature [78]. The definition for 'metabonomics' previously proposed by Nicholson et al.: 'the quantitative measurement of the multiparametric response of living organisms to pathophysiological stimuli or genetic modification' [75] will be used to cover the term 'metabolomics' in this introduction.

When foreign compounds (i.e. drugs) interact with cells and tissues they disturb the ratios, concentrations and fluxes of endogenous metabolites by interfering with key intermediary metabolic pathways [73;78;79]. When slightly perturbed (i.e. as a consequence of disease processes or intervention), cells will try to maintain homeostasis and metabolic control by changing the composition of their intra- and extra cellular body fluids [79]. Hence, as a result of this 'metabolic disturbance', distinctive organ- and mechanism-specific alterations in biofluid (i.e. urine, plasma, saliva, etc.) composition can be observed [78]. To explore the multifaceted metabolic reactions to this disturbance, which can be disease processes, reactions to a toxin or drug, or even genetic manipulation, non-selective, specific 'information-rich' investigative approaches are essential [73;78;79]. In addition to ^{1}H NMR spectroscopy, which was employed in the studies described in this thesis, several other analytical methods can serve as powerful means of acquiring multivariate metabolic data [80], including; Gas Chromatography / Mass Spectrometry (GC/MS) [77;81], High Performance Liquid Chromatography / Mass Spectrometry (HPLC/MS) [82], HPLC-coupled with a diode array detector (DAD) [83] and optical spectroscopic techniques [75;76] (figure 3; p. 203).

Even though the field of clinical metabolomics is still in its infancy, many reports have been published involving a great variety of disease states and biofluids, illustrating the vast prospective for development and growth of this research field [78;84-87].

In (pre-clinical) animal experiments, the precise biological history of the studied animals is known and experimental conditions (i.e. diet, temperature and compound dose) as well as sample timing are well controlled. However, in the case of human studies, there is a surfeit of variables in the medical history and experimental conditions that are much more difficult to control [78]. The intrinsic variability of human metabolic and biochemical parameters makes human biofluids more complex to analyse and even more care must be given to rigorous statistical analysis (see further) [78].

Background to NMR

The NMR phenomenon was first discovered in 1945 and published in 1946 [88;89]. With the advent of chemical shift and spin coupling it has developed into a key analytical technique that still continues to evolve [78]. Over the last 20 years, an enormous increase in technological innovations and knowledge in this research area has resulted in a wide range of applications for NMR, ranging from the characterization of prospective protein drugs and protein drug targets [90] to whole-body imaging using magnetic resonance imaging (MRI) devices [91].

The basic principle of the conventional NMR technique can be explained as follows. The test-sample, contained in a special glass tube of diameter of 5 or 10 mm, usually dissolved in a particular solvent, is placed in a high uniform magnetic field. Subsequently, from a variable radio frequency (RF) source, RF of appropriate power is sent through a shielded cable to a coil that is wound around the tube holding the test-sample. The frequency is varied stepwise using small increments (or so-called 'scanning') through a preset frequency range. During scanning, the absorption of the RF is monitored and measured using appropriate electronic circuitry. Subsequently, the frequency of the RF and its corresponding absorption value are plotted on the x- and y- axes. The resulting graph, is characteristic of the compound(s) in the test sample and referred to as the 'NMR spectrum' [92;93].

The absorption of the RF energy occurs because of the presence of some nuclei (i.e. ^{1}H, ^{13}C, ^{31}P, ^{19}F etc.) all of which exhibit non-zero nuclear spin and are called magnetic nuclei. The NMR spectrum arising due to the presence of ^{1}H is generally termed 'proton NMR' or 'PMR'. When a compound containing such an atom is kept in the magnetic field it behaves in two or more different ways that are associated with different energy levels. Protons split into two different energy levels and the energy gap between these levels depends upon many factors, but primarily the magnetic field itself. The higher the magnetic field, the bigger the energy gap and hence higher RF will be absorbed [92;93].

However, during NMR measurements the field is kept constant. Even in a constant magnetic field, the energy level gap for the same nuclei within a molecule can be different due to the difference in shielding of the nuclei by the electrons surrounding the nuclei. In other words, since electrons around the nuclei of interest are involved in bond formation, they shield or de-shield the nuclei to different extents resulting in different energy gaps and in turn different RF absorptions. These shifts in the absorption lines are called 'Chemical Shifts' and are one of the most fundamental parameters of NMR [92].

Generally, NMR signals are reported in 'chemical shifts', which are the ratio of the frequency difference between a reference standard and the test signal to the frequency of the spectrometer. This is a dimensionless quantity is expressed in ppm to make the ratio a convenient number. This dimensionless chemical-shift quantity is generally termed as 'd' [92;93].

Occasionally, the chemical shift reference standard is added to the solvent itself so that one can directly measure/calculate the chemical shift. Thus, Tetramethyl silane (TMS), which is an inert volatile compound, is frequently used as ppm marker for ^{1}H, ^{13}C and ^{29}Si NMR measurements. The NMR signal arising out of the TMS is set to zero ppm in each case. Similarly for other nuclei such convenient standards are chosen. Sometimes, it is not possible to have these primary ppm markers or reference in the solvents. In such cases, secondary standards can be used, for which the ppm value is known [92;93].

When one is not too much concerned about the exact ppm value, one can use external referencing. In this case, the spectrum of a known compound of known ppm value is measured first and the spectral chart is calibrated. Then without changing any instrumental experimental parameters, the NMR spectrum of the test compound is measured. By comparing the 'external' standard spectrum and the sample spectrum, one can get a reasonably accurate ppm value of the signals [92;93].

Another important phenomenon in NMR is the interaction between neighbouring magnetic nuclei. This is called spin-spin coupling [92;93]. This causes the otherwise single resonance line of each chemically

shifted nucleus to be split into multiple lines in an ordered fashion. Though this can increase the complexity of the spectrum, one can, if correctly assigned, get a wealth of information of the group of atoms present in the immediate vicinity and their through-bond connectivity [92;93].

Instrumentation

A NMR instrument comprises a Magnet (permanent magnet / electromagnet or iron magnet / superconducting magnet). The sample is inserted into the centre of the magnetic field. The sample within the sample tube is surrounded by RF coils as mentioned earlier. Signals are detected by either the same coil or another coil depending upon the manufacturer's design. The main spectrometer consists of the RF generator, magnetic field monitor, RF signal receiving and detecting system-recording device, and necessary power supplies. Optionally, there can be computer connected to the spectrometer to collect, store and process the spectral data [92;93].

There are two classes of NMR spectrometers; continuous wave (CW) and pulsed Fourier Transform (FT) spectrometers. The CW technique is quite old and is almost obsolete whereas almost all current manufacturers produce only FT NMR spectrometers, which are much more versatile. Though both type of spectrometers contain the magnet and the sample probe, the design of the RF transmitter and the detection circuitry are different. In FT NMR, the sample in a resonant coil is subjected to an intense and short RF pulse. Since the short pulse contains a broad band of frequencies (Fourier components of the RF pulse) this system is subjected to a broad band excitation. Irrespective of the differences in the chemical shift spread all nuclei of a given kind are simultaneously excited and give a time-response, which is acquired by a fast digital computer to give the so-called Free Induction Decay (FID). Upon a mathematical Fourier transform, the FID produces a frequency spectrum, which will be identical to the spectrum obtained by the conventional CW sweep of the frequency [92].

The advantage is that several thousand responses from a sample can be coherently added, increasing the signal to noise ratio enormously. This makes it possible to address by NMR any Magnetic nucleus in the Periodic Table irrespective of the magnetic moment and natural abundance [92;93].

The role of Liquid-State ^{1}H NMR in metabolomics

An important point of metabolomics studies is to have one or more generic analytical methods at ones disposal that allow rapid biological sample profiling [78]. High resolution NMR spectroscopy is now a firmly established method in this area, and it is well documented that it can generate useful qualitative and quantitative data on the metabolic status of a person, animal or cell system [78;94-96]. A number of studies have reported on the usefulness of ^{1}H NMR spectroscopic for the analysis of a broad array of biofluids, including i.e. urine [97], blood plasma [98;99] or serum [85;100], seminal fluid [101], and spinal fluid [102]. The characteristics of NMR spectroscopy that are useful in this type of study are summarized in table 1.

TABLE 1 Summary of useful features of ^{1}H NMR spectroscopy in biofluid studies.

Feature	Comment
Selectivity	'Hypothesis-free' pre-selection of analytical conditions based on the chemical properties of the analyte(s), or postulation of the metabolites affected by a disease or intervention is not necessary. Therefore, a broad array of low molecular weight metabolites and macromolecules can be rapidly monitored in parallel, without a priori knowledge or expectation of the results.
Non-Invasive	The non-invasive, non-destructive and non-equilibrium perturbing nature of NMR spectroscopy allows the subsequent analysis of a sample with other techniques.
Speed	A typical single pulse ^{1}H NMR biofluid spectrum can be obtained in les than 10 minutes.
Sample Volumes	Only 50 µl of sample is needed for routine spectroscopy with a standard NMR probe and only 1-2 µl with a micro-volume NMR probe.
Sample Preparation	The only requirement for analysis is the addition of deuterium oxide.
Dynamic Information	The technique can provide data on dynamic biochemical processes in complex matrices and molecular interactions, i.e. protein-ligand binding.
Structural Information	Data provides qualitative structural information, which, with the addition of internal standards at known concentrations, can also be quantitative.

Different from other investigative methods, pre-selection of the analytes of interest is not required for NMR spectroscopy [78]. In fact, its strength lies within its potential to rapidly provide data relating to low molecular weight compounds and macromolecules without a priori knowledge or expectation of the results [78]. Moreover, the latest advances in sensitivity and the non-destructive nature of the method have led to its increased popularity for biomedical studies, illustrated by the increasing number of scientific papers being published on the subject [78;98;99;103-110].

Although conventional enzymatic and chromatographic methods are more sensitive than ^{1}H NMR spectroscopy in detecting low levels of metabolites, these methods entail consideration of the key biochemical pathways and pre-selection of metabolites of interest [78]. Consequently, the subjective and time-consuming process of selection of a series of biochemical methods is required. Moreover, if an inappropriate or too narrow range of biochemical methods and variables are selected, important metabolic changes may be missed [78]. In contrast, the use of NMR spectroscopy to study (changes in) metabolite profiles does not involve such a pre-selection of analytes and thus permits subsequent 'hypothesis-free' multivariate analysis without any bias imposed by the scientist's expectations [78].

Multivariate Data Analysis

^{1}H NMR spectra of biofluids capture a wealth of information on endogenous processes in both health and disease [78]. These biofluids (e.g. plasma or urine) contain numerous compounds, which in turn produce thousands of NMR signals and consequently, signal overlap and crowding of the spectra can occur [78]. Subtle metabolic differences or alterations can therefore be easily overlooked when examining these spectra by eye [78]. Moreover, in most cases it is the global metabolite profile, rather than the absolute concentrations of specific individual metabolites, which appears to provide most of the information on the metabolic status [73;78].

This highly complex spectral profile can be more efficiently analysed using advanced data reduction and pattern recognition methods for classification (diagnostic) purposes [78].

Metabolomics studies (in analogy to microarray experiments) generally produce large amounts of complex data. The datasets, containing NMR spectra and clinical data (i.e. disease or treatment status), hold biological information that encompasses many correlated variables or so-called 'peaks' [78]. The correlation between the individual peaks is derived from the fact that a single metabolite will characteristically have more than one peak in an NMR spectrum and that there are biological relationships between metabolites [78].

To identify these relationships, several multivariate analysis methods can be used [111-113], however, a detailed description is beyond the scope of this introduction. Here, we will focus on the most clear-cut approach (used in the metabolomics studies of this thesis) being Principal Component Analysis (PCA) [114]. For practical reasons, PCA coupled with discriminant analysis (PC-DA) approaches are extremely relevant in the context of high dimensional (NMR) data [74]. Typically, PC-DA analysis will be performed on triplicate analytical measurements on each sample, after which each sample is defined by a category (e.g. active treated *vs.* placebo treated). Then, after selection of a number of principal components, a discriminant analysis is performed [115]. Any observed clustering or separation of samples resulting from PCA or PC-DA provides strong evidence for statistically significant differences between or among sample datasets [74]. Another (related) method entails the partial least-squares (PLS) analysis [116]. This method is basically a generalization of PCA where a projection model is developed predicting Y (i.e., an outcome) from X (the variables used in PCA as described above) via the scores of X [116]. When PLS is combined with discriminant analysis (DA) the Y's are defined by the user and the algorithm allocates observations (a set of samples) to one class of a given set of classes [116].

Usually, the PCA or PC-DA clustering is the first step in pattern recognition, followed by the interpretation of the peaks within the dataset that largely account for an observed pattern [74]. Such interpretation identifies the variables (metabolites) that are responsible for the differences between separate clusters (e.g. disease *vs.* healthy) and thus provides the basis for subsequent biological interpretation [74]. Unfortunately, many of these variables are unknown and can only be identified by their chemical shift values [74]. In-house proprietary databases and public domain databases can assist in the identification, but especially at lower concentrations, the number of unknown variables increases progressively [74]. Proper identification of these unknown variables thus becomes one of the key challenges in metabolomics research.

Why are the '-omics' technologies important?

The '-omics' technologies provide the means for the identification of distinct molecular phenotypes or 'fingerprints' through the analysis of (human) tissues, cells and biofluids. Furthermore, it is anticipated that the application of the '-omics' technologies will continue to improve molecular diagnostics and will provide 'deep' insights into the pathogenic alterations in diseases and mechanisms of pharmacological interventions [10]. Especially when the different '-omics' sub disciplines

are integrated, as advocated in the 'systems biology' approach [117], it can provide a broad, more 'holistic' view on disease and drug mechanisms, which could eventually lead to the development of new or improved treatment modalities. However, in the end, the combined strength of the three major '-omics' pillars ultimately determines if 'systems biology' can live up to its expectations and will become a useful tool for future clinical drug evaluation.

REFERENCE LIST

1 Microarrays: Chipping Away At The Mysteries Of Science And Medicine. Http://www.ncbi.nlm.nih.gov/about/primer/microarrays.html (19-1-2006)

2 Lewin DA, Weiner MP. Molecular biomarkers in drug development. *Drug Discov.Today* 2004; **9**:976-983

3 Alwine JC, Kemp DJ, Parker BA *et al.* Detection of specific RNAs or specific fragments of DNA by fractionation in gels and transfer to diazobenzyloxymethyl paper. *Methods Enzymol.* 1979; **68**:220-242

4 Arya M, Shergill IS, Williamson M, Gommersall L, Arya N, Patel HR. Basic principles of real-time quantitative PCR. *Expert.Rev.Mol Diagn.* 2005; **5**:209-219

5 Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat.Genet.* 1999; **21**:10-14

6 Hsu B, Cass L, Williams WV. Application of transcriptome analysis to clinical pharmacology studies. *Curr.Mol Med* 2005; **5**:65-82

7 Lockhart DJ, Dong H, Byrne MC *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 1996; **14**:1675-1680

8 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; **270**:467-470

9 Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996; **6**:639-645

10 Bilello JA. The agony and ecstasy of 'OMIC' technologies in drug development. *Curr.Mol Med* 2005; **5**:39-52

11 Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995; **270**:484-487

12 Brenner S, Johnson M, Bridgham J *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat.Biotechnol.* 2000; **18**:630-634

13 Southern EM, Maskos U, Elder JK. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* 1992; **13**:1008-1017

14 Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science* 1991; **251**:767-773

15 Bailey WJ, Ulrich R. Molecular profiling approaches for identifying novel biomarkers. *Expert.Opin.Drug Saf* 2004; **3**:137-151

16 Heller RA, Schena M, Chai A *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc.Natl. Acad.Sci.U.S.A* 1997; **94**:2150-2155

17 Khan J, Simon R, Bittner M *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res* 1998; **58**:5009-5013

18 Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* 1996; **33**:151-152

19 Perou CM, Jeffrey SS, van de RM *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc.Natl.Acad.Sci.U.S.A* 1999; **96**:9212-9217

20 Bittner M, Meltzer P, Chen Y *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 2000; **406**:536-540

21 Golub TR, Slonim DK, Tamayo P *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286**: 531-537

22 't Veer LJ, Dai H, van de Vijver MJ *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002; **415**:530-536

23 Paik S, Shak S, Tang G *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N.Engl.J Med* 2004; **351**:2817-2826

24 MammaPrint(R). http://www.agendia.com (15-6-2006)

25 Thomas RS, Rank DR, Penn SG *et al.* Identification of toxicologically predictive gene sets using cDNA microarrays. *Mol Pharmacol* 2001; **60**:1189-1194

26 Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc.Natl.Acad.Sci. U.S.A* 2003; **100**:9608-9613

27 DePrimo SE, Wong LM, Khatry DB *et al.* Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC.Cancer* 2003; **3**:3

28 Palotas A, Puskas LG, Kitajka K *et al.* Altered response to mirtazapine on gene expression profile of lymphocytes from Alzheimer's patients. *Eur.J Pharmacol* 2004; **497**:247-254

29 Palotas A, Puskas LG, Kitajka K *et al.* The effect of citalopram on gene expression profile of Alzheimer lymphocytes. *Neurochem.Res* 2004; **29**:1563-1570
30 Braxton S, Bedilion T. The integration of microarray information in the drug development process. *Curr.Opin.Biotechnol.* 1998; **9**:643-649
31 Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nat.Genet.* 1999; **21**:48-50
32 Marton MJ, DeRisi JL, Bennett HA *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat.Med* 1998; **4**:1293-1301
33 Dong S, Wang E, Hsie L, Cao Y, Chen X, Gingeras TR. Flexible use of high-density oligonucleotide arrays for single-nucleotide polymorphism discovery and validation. *Genome Res* 2001; **11**:1418-1424
34 Jain KK. Applications of AmpliChip CYP450. *Mol Diagn.* 2005; **9**:119-127
35 Wang D, Urisman A, Liu YT *et al.* Viral discovery and sequence recovery using DNA microarrays. *PLoS.Biol* 2003; **1**:E2
36 Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. *Nat.Rev.Genet.* 2006; **7**:200-210
37 Schoch, C, Dugas, M, Kern, W, Kohlmann, A, Schnittger, S, and Haferlach, T. 'Deep insight' into microarray technology. http://www.infobiogen.fr/services/chromcancer/deep/microarraysid20045.html (17-05-2005)
38 Southern E, Mir K, Shchepinov M. Molecular interactions on microarrays. *Nat.Genet.* 1999; **21**:5-9
39 Hess KR, Zhang W, Baggerly KA, Stivers DN, Coombes KR. Microarrays: handling the deluge of data and extracting reliable information. *Trends Biotechnol.* 2001; **19**:463-468
40 Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 2003; **19**:649-659
41 Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* 2003; **19**:570-577
42 Nadon R, Shoemaker J. Statistical issues with microarrays: processing and analysis. *Trends Genet.* 2002; **18**:265-271
43 Quackenbush J. Computational analysis of microarray data. *Nat.Rev.Genet.* 2001; **2**:418-427
44 Quackenbush J. Microarray data normalization and transformation. *Nat.Genet.* 2002; **32 Suppl**:496-501
45 Slonim DK. From patterns to pathways: gene expression data analysis comes of age. *Nat.Genet.* 2002; **32 Suppl**:502-508
46 Butte A. The use and analysis of microarray data. *Nat.Rev.Drug Discov.* 2002; **1**:951-960
47 Liu ET, Karuturi KR. Microarrays and clinical investigations. *N.Engl.J Med* 2004; **350**:1595-1597
48 Perou CM, Sorlie T, Eisen MB *et al.* Molecular portraits of human breast tumours. *Nature* 2000; **406**:747-752
49 Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc.Natl.Acad.Sci.U.S.A* 1998; **95**:14863-14868
50 Shannon W, Culverhouse R, Duncan J. Analyzing microarray data using cluster analysis. *Pharmacogenomics.* 2003; **4**:41-52
51 Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl.Cancer Inst.* 2003; **95**:14-18
52 Broberg P. Statistical methods for ranking differentially expressed genes. *Genome Biol* 2003; **4**:R41
53 Brody JP, Williams BA, Wold BJ, Quake SR. Significance and statistical errors in the analysis of DNA microarray data. *Proc.Natl.Acad.Sci.U.S.A* 2002; **99**:12975-12978
54 Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. *Genet.Epidemiol.* 2002; **23**:70-86
55 Pan W. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics.* 2003; **19**:1333-1340
56 Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics.* 2002; **18**:1454-1461
57 Xu R, Li X. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics.* 2003; **19**:1284-1289
58 Zhao Y, Pan W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics.* 2003; **19**:1046-1054

59 Brazeau DA. Combining genome-wide and targeted gene expression profiling in drug discovery: microarrays and real-time PCR. *Drug Discov.Today* 2004; **9**:838-845
60 Bretz F, Landgrebe J, Brunner E. Multiplicity issues in microarray experiments. *Methods Inf. Med* 2005; **44**:431-437
61 Significance Analysis of Microarrays (SAM). http://www-stat.stanford.edu/~tibs/sam (19-1-2006)
62 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc.Natl.Acad.Sci.U.S.A* 2001; **98**:5116-5121
63 BrBarraytools. http://linus.nci.nih.gov/brb-arraytools.html (19-1-2006)
64 Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet.* 2000; **25**:25-29
65 Online Mendelian Inheritance in Man (omim) http://www.ncbi.nlm.nih.gov/omim/ (19-1-2006)
66 Swiss-Prot Protein knowledgebase. http://us.expasy.org/sprot (19-1-2006)
67 NCBI Genbank. http://www.ncbi.nlm.nih.gov/genbank/ (9-12-2006)
68 EBI Ensembl. http://www.ensembl.org/index.html (21-12-2006)
69 Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004; **32**:D277-D280
70 Krieger CJ, Zhang P, Mueller LA *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2004; **32**:D438-D442
71 Guffanti A, Reid JF, Alcalay M, Simon G. The meaning of it all: web-based resources for large-scale functional annotation and visualization of DNA microarray data. *Trends Genet.* 2002; **18**:589-592
72 Ingenuity pathway analysis. http://www.ingenuity.com (19-1-2006)
73 van der Greef J, Davidov E, Verheij E *et al.* Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis. In: *The Role of Metabolomics in Systems Biology : A New Vision for Drug Discovery and Development* eds Harrigan G, Goodacre R, Boston / Dordrecht / London: Kluwer Academic Publishers, 2003: 190-198
74 Witkamp RF. Genomics and systems biology--how relevant are the developments to veterinary pharmacology, toxicology and therapeutics? *J Vet.Pharmacol Ther* 2005; **28**:235-245
75 Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 1999; **29**:1181-1189
76 Watkins SM. Comprehensive lipid analysis: a powerful metanomic tool for predictive and diagnostic medicine. *Isr.Med Assoc.J* 2000; **2**:722-724
77 Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 2000; **18**:1157-1161
78 Daykin, C. A., Vogels, J., Tas, A, and Wulftert, F. Metabolomics Rosiglitazone Proof-of-Principle study report: Comparison of Human Subjects with type II Diabetes and Healthy Volunteers; Introduction. (17-5-2005)
79 Lindon JC, Holmes E, Nicholson JK. So what's the deal with metabonomics? *Anal.Chem* 2003; **75**:384A-391A
80 Sumner LW, Mendes P, Dixon RA. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 2003; **62**:817-836
81 Mills GA, Walker V. Headspace solid-phase microextraction profiling of volatile compounds in urine: application to metabolic investigations. *J Chromatogr.B Biomed.Sci. Appl* 2001; **753**:259-268
82 Tolstikov VV, Fiehn O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal.Biochem* 2002; **301**:298-307
83 Pham-Tuan H, Kaskavelis L, Daykin CA, Janssen HG. Method development in high-performance liquid chromatography for high-throughput profiling and metabonomic studies of biofluid samples. *J Chromatogr. B Analyt.Technol.Biomed.Life Sci.* 2003; **789**:283-301
84 't Hart BA, Vogels JT, Spijksma G, Brok HP, Polman C, van der GJ. 1H-NMR spectroscopy combined with pattern recognition analysis reveals characteristic chemical patterns in urines of MS patients and non-human primates with MS-like disease. *J Neurol. Sci.* 2003; **212**:21-30
85 Brindle JT, Nicholson JK, Schofield PM, Grainger DJ, Holmes E. Application of

chemometrics to 1H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst* 2003; **128**:32-36
86 Lamers RJ, DeGroot J, Spies-Faber EJ *et al.* Identification of disease- and nutrient-related metabolic fingerprints in osteoarthritic Guinea pigs. *J Nutr.* 2003; **133**:1776-1780
87 Lenz EM, Bright J, Wilson ID, Morgan SR, Nash AF. A 1H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J Pharm Biomed.Anal.* 2003; **33**:1103-1115
88 Bloch F. Nuclear induction. *Phys rev* 1946; **70**:460-474
89 Purcell EM, Torrey HC, Pound RV. Resonance absorption by nuclear magnetic moments in a solid. *Phys rev* 1946; **69**:37-38
90 Wishart D. NMR spectroscopy and protein structure determination: applications to drug discovery and development. *Curr.Pharm Biotechnol.* 2005; **6**:105-120
91 Lauenstein TC, Semelka RC. Whole-body magnetic resonance imaging. *Top.Magn Reson. Imaging* 2005; **16**:15-20
92 Nuclear Magnetic Resonance spectroscopy http://saif.iitm.ac.in/facilities/4_nmr.html (19-1-2006)
93 Storey P. Introduction to magnetic resonance imaging and spectroscopy. *Methods Mol Med* 2006; **124**:3-57
94 Bundy JG, Osborn D, Weeks JM, Lindon JC, Nicholson JK. An NMR-based metabonomic approach to the investigation of coelomic fluid biochemistry in earthworms under toxic stress. *FEBS Lett.* 2001; **500**:31-35
95 Griffin JL, Walker LA, Troke J, Osborn D, Shore RF, Nicholson JK. The initial pathogenesis of cadmium induced renal toxicity. *FEBS Lett.* 2000; **478**:147-150
96 Khandelwal P, Beyer CE, Lin Q, McGonigle P, Schechter LE, Bach AC. Nanoprobe NMR spectroscopy and in vivo microdialysis: new analytical methods to study brain neurochemistry. *J Neurosci.Methods* 2004; **133**:181-189
97 Anthony ML, Sweatman BC, Beddell CR, Lindon JC, Nicholson JK. Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Mol Pharmacol* 1994; **46**:199-211
98 Daykin CA, Foxall PJ, Connor SC, Lindon JC, Nicholson JK. The comparison of plasma deproteinization methods for the detection of low-molecular-weight metabolites by (1)H nuclear magnetic resonance spectroscopy. *Anal.Biochem* 2002; **304**:220-230
99 Wevers RA, Engelke U, Heerschap A. High-resolution 1H-NMR spectroscopy of blood plasma for metabolic studies. *Clin Chem* 1994; **40**:1245-1250
100 Brindle JT, Antti H, Holmes E *et al.* Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat.Med* 2002; **8**:1439-1444
101 Tomlins AM, Foxall PJ, Lynch MJ, Parkinson J, Everett JR, Nicholson JK. High resolution 1H NMR spectroscopic studies on dynamic biochemical processes in incubated human seminal fluid samples. *Biochim.Biophys Acta* 1998; **1379**:367-380
102 Coen M, O'Sullivan M, Bubb WA, Kuchel PW, Sorrell T. Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis. *Clin Infect.Dis.* 2005; **41**:1582-1590
103 Daykin CA, Corcoran O, Hansen SH *et al.* Application of directly coupled HPLC NMR to separation and characterization of lipoproteins from human serum. *Anal.Chem* 2001; **73**:1084-1090
104 Daykin CA, Van Duynhoven JP, Groenewegen A, Dachtler M, Van Amelsvoort JM, Mulder TP. Nuclear magnetic resonance spectroscopic based studies of the metabolism of black tea polyphenols in humans. *J Agric.Food Chem* 2005; **53**:1428-1434
105 Deprez S, Sweatman BC, Connor SC, Haselden JN, Waterfield CJ. Optimisation of collection, storage and preparation of rat plasma for 1H-NMR spectroscopic analysis in toxicology studies to determine inherent variation in biochemical profiles. *J Pharm Biomed.Anal.* 2002; **30**:1297-1310
106 Messana I, Forni F, Ferrari F, Rossi C, Giardina B, Zuppi C. Proton nuclear magnetic resonance spectral profiles of urine in type II diabetic patients. *Clin Chem* 1998; **44**: 1529-1534
107 Nicholson JK, Sadler PJ, Bales JR, Juul SM, MacLeod AF, Sonksen PH. Monitoring metabolic disease by proton NMR of urine. *Lancet* 1984; **2**:751-752
108 Nicholson JK, O'Flynn MP, Sadler PJ, MacLeod AF, Juul SM, Sonksen PH. Proton-nuclear-magnetic-resonance studies of serum, plasma and urine from fasting normal and diabetic subjects. *Biochem J* 1984; **217**:365-375
109 Nicholson JK, Foxall PJ, Spraul M, Farrant RD, Lindon JC. 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. *Anal. Chem* 1995; **67**:793-811
110 Williams RE, Lenz EM, Evans JA *et al.* A combined (1)H NMR and HPLC-MS-based metabonomic study of urine from obese (fa/fa) Zucker and normal Wistar-derived rats. *J Pharm Biomed.Anal.* 2005; **38**:465-471
111 Beebe KR, Pell RJ, Seasholz MB. *Chemometrics: a practical guide* New York: John Wiley & Sons, 1998
112 Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi- and Megavariate Data Analysis; Principles and Applications* Umea, Sweden: Umetrics AB, 2006
113 Jackson JE. *A user's guide to principal components* New York: John Wiley & Sons, 2006
114 Windig W, Haverkamp J, Kistemaker PG. Interpretation of Sets of Pyrolysis Mass-Spectra by Discriminant-Analysis and Graphical Rotation. *Analytical Chemistry* 1983; **55**:81-88
115 Hoogerbrugge R, Willig SJ, Kistemaker PG. Discriminant-Analysis by Double Stage Principal Component Analysis. *Analytical Chemistry* 1983; **55**:1710-1712
116 Leo G, van Hoogdalem E, van Doorn M. NMR-Based Metabolomics of Urine from An Exploratory Study of Ciprofibrate in Healthy Volunteers and Patients with Type 2 Diabetes Mellitus.In: *Frontiers in Drug Design & Discovery* eds Caldwell G, Atta-ur-Rahman, Springer B, Bentham Science Publishers, 2006
117 Davidov E, Holland J, Marple E, Naylor S. Advancing drug discovery through systems biology. *Drug Discov.Today* 2003; **8**:175-183.