Metabolomics (2011) 7:329–343
DOI 10.1007/s11306-010-0270-3

REVIEW ARTICLE

Bioinformatics tools for cancer metabolomics

Grigoriy Blekherman · Reinhard Laubenbacher · Diego F. Cortes · Pedro Mendes · Frank M. Torti · Steven Akman · Suzy V. Torti · Vladimir Shulaev

Received: 19 August 2010 / Accepted: 20 December 2010 / Published online: 12 January 2011

Abstract It is well known that significant metabolic change take place as cells are transformed from normal to malignant. This review focuses on the use of different bioinformatics tools in cancer metabolomics studies. The article begins by describing different metabolomics technologies and data generation techniques. Overview of the data pre-processing techniques is provided and multivariate data analysis techniques are discussed and illustrated with case studies, including principal component analysis, clustering techniques, self-organizing maps, partial least squares, and discriminant function analysis. Also included is a discussion of available software packages.

Keywords Metabolomics · Cancer · Metabolite profiling · NMR · Mass spectrometry · Bioinformatics

G. Blekherman · R. Laubenbacher · D. F. Cortes · P. Mendes · V. Shulaev
Virginia Bioinformatics Institute, Washington St. 0477, Blacksburg, VA 24061, USA

R. Laubenbacher · P. Mendes · F. M. Torti · S. Akman · S. V. Torti · V. Shulaev
Comprehensive Cancer Center, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

P. Mendes
School of Computer Science and Manchester Centre for Integrative Systems Biology, The University of Manchester, 131 Princess St, Manchester, M1 7DN, UK

F. M. Torti · S. Akman
Department of Cancer Biology, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

S. V. Torti
Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC 27157, USA

V. Shulaev (✉)
Department of Biological Sciences, College of Arts and Sciences, University of North Texas, 1155 Union Circle #305220, Denton, TX 76203, USA
e-mail: shulaev@unt.edu

1 Introduction

A significant role in cancer initiation and progression is attributed to changes in RNA and protein expression levels and regulation (Byrum et al. 2010; Chari et al. 2010; Fink-Retter et al. 2009; Korkola and Gray 2010; Larkin et al. 2010). However, changes in small molecules also provide important mechanistic insights into cancer development. There is a strong body of evidence supporting the important role of metabolic regulation in cancer. Malignant cells undergo significant changes in metabolism including a redistribution of metabolic networks (Boros et al. 2003). These metabolic changes result in different metabolic landscapes in cancer cells versus normal cells. Metabolomics, as a global approach, is especially useful in identifying overall metabolic changes associated with a particular biological process and finding the most affected metabolic networks. Moreover, metabolomics provides an additional layer of information that can be linked with transcriptomics and proteomics data to obtain a comprehensive view of a biological system.

Metabolomics is a relatively new field in genomics research but it is gaining broader recognition in the cancer community. Most cancer metabolomics studies to date have been done using metabolic fingerprinting or profiling with NMR spectroscopy of tissue extracts or in vivo magnetic resonance spectroscopy. Using NMR spectroscopy techniques it is possible to differentiate several tumor

types in humans and in animal models (Beckonert et al. 2010; Cheng et al. 1996; Devos et al. 2004; Lukas et al. 2004; Merz and Serkova 2009; Tate et al. 1998, 2000). But while techniques based on magnetic resonance have the advantage of being non-invasive, they have low sensitivity and cannot detect molecules at low concentrations. Mass spectrometry methods provide advantage of higher sensitivity and are more appropriate for in vitro studies (Patterson et al. 2010; Qiu et al. 2010; Sugimoto et al. 2010; Urayama et al. 2010; Want et al. 2010).

Similar to transcriptomics and proteomics, metabolomics generates large amounts of data. Storing, pre-processing and multivariate statistical analysis of these data provide a significant challenge and require specialized mathematical, statistical and bioinformatics tools (reviewed in (Katajamaa and Orešič 2007; Madsen et al. 2010; Shulaev 2006)). Metabolomics experiments generate a large volume of specialized data that are complex and multi-dimensional. Storing, organizing and retrieving the data and associated metadata requires properly designed databases. The analysis of these data sets is equally challenging and new analysis algorithms are still being developed. Multivariate statistical analysis of the metabolomics data in many cases utilizes the same approaches as the analysis of other genomic data. However, metabolomics has unique bioinformatics needs in addition to others common in microarray or proteomics data due to the fact that it is generated by multiple analytical platforms and requires extensive data pre-processing. Major areas where developments in data analysis techniques are crucial for further progress of metabolomics include: data and information management, raw analytical data processing, metabolomics standards and ontology, statistical analysis and data mining, data integration, and mathematical modeling of metabolic networks within the framework of systems biology.

This article aims at providing a basic overview of metabolomics data analysis, including practical applications and metabolomics software, as it is being used for cancer research. We discuss major multivariate data analysis techniques, including principal component analysis, clustering techniques, self-organizing maps, partial least squares, and discriminant function analysis., and illustrate them with case studies.

2 Metabolomics technology and data generation

There are three major approaches used in metabolomics studies: targeted analysis, metabolite profiling and metabolic fingerprinting (Fiehn 2002; Shulaev 2006). Targeted analysis is used to measure the concentration of limited numbers of known metabolites precisely. It is a truly quantitative approach and provides low limits of detection for known metabolites. Metabolite profiling usually is an untargeted high throughput measurement of the levels of a large number of metabolites. The advantage of metabolite profiling is that, unlike targeted analysis, it does not require the compounds of interest to be known a priori and therefore it can be used to identify novel metabolic biomarkers or survey the global metabolic changes in the system. Metabolic fingerprinting considers the total metabolite profile, or fingerprint, as a unique pattern characterizing a snapshot of the metabolism in a particular cell line or tissue.

Analytical techniques used for metabolite profiling include nuclear magnetic resonance (NMR) (Jordan and Cheng 2007; Serkova and Glunde 2009), Gas Chromatography–Mass Spectrometry (GC–MS) (Asiago et al. 2010; Qiu et al. 2010; Yang et al. 2007), Liquid Chromatography–Mass Spectrometry (LC–MS) (Chen et al. 2009; Kind et al. 2007; Yang et al. 2010), Capillary Electrophoresis–Mass Spectrometry (CE–MS) (Soga 2007), and Fourier transform infrared spectroscopy (FT-IR) (Johnson et al. 2003; Kim et al. 2010). Advantages and downside of each techniques for metabolomics were reviewed elsewhere (Fiehn 2002; Shulaev 2006; Sumner et al. 2003). Due to diversity in cellular metabolites' chemical and physical properties there is no single analytical technique that can analyze all metabolites simultaneously. Usually a combination of analytical techniques is used to cover as broad a range of metabolites as possible. This presents a unique challenge for data analysis since each analytical technique generates a specific data structure and has to be processed with a specialized informatics tool.

3 Metabolomics standards and metadata

Metabolomics, like other genomics research areas, requires standards of data management, analysis and reporting to be adopted by community (Bino et al. 2004; Castle et al. 2006; Griffin and Steinbeck 2010; Jenkins et al. 2004; Lindon et al. 2005; Sansone et al. 2007). This is critical considering the growing number of metabolomics studies, the urgent need to store metabolomics data in genomics databases, and the ability to compare data generated in different labs (Castle et al. 2006). Metabolomics Society (http://www.metabolomicssociety.org/) working group is currently working on the development of metabolomics standards to be adopted by the community (http://www.metabolomicssociety.org/mstandards.html). The first step in devising standards for metabolomics experiments was the development of the Minimum Information about a Metabolomics Experiment (MIAMET) (Bino et al. 2004). MIAMET defines the minimum information required to

report from metabolomics experiment including experimental data and metadata (data about the experiment).

4 Metabolomics data analysis

Major steps in metabolomics data analysis include raw data pre-processing, spectral deconvolution and component detection, data normalization and multivariate statistical analysis.

4.1 Raw data pre-processing

The first step in metabolomics data analysis is the processing of the raw data and it involves several steps depending on the methodology used (Hansen 2007). Due to the complex nature of metabolomics data, when the objective is to identify and measure as many different metabolites as possible, raw data processing is a very important step in data analysis. Raw data processing for different techniques used in metabolomics have been extensively reviewed (Hansen 2007; Katajamaa and Orešič 2007; Scalbert et al. 2009; Schripsema 2010; Spraul et al. 1994). Therefore, here we will just introduce the major concepts on how the raw data can be processed and transformed into the format suitable for multivariate statistical analysis or machine learning techniques. More detail on the individual raw data processing techniques the reader can find in the referenced manuscripts.

Typically NMR data show variation in peak width, shape and position due to differences in sample matrix (i.e., pH or ionic strength) or variations in instrument performance. Therefore, raw data should be pre-processed to correct these variations. NMR data pre-processing usually include correction of line width using line broadening parameter (i.e., using tuned exponential multiplication), Fourier transformation, phase correction with user predefined phase constants, and positioning and scaling (Lommen et al. 1998). NMR data processing techniques include binning, peak picking, and spectrum deconvolution (Schripsema 2010). Binning or bucketing is the most common routine for NMR data processing prior to multivariate statistical analysis or fingerprinting (Beckwith-Hall et al. 1998; Spraul et al. 1994). Binning is achieved by separating the spectra into multiple discrete regions (hixels or buckets) which are than averaged and integrated. This leads to substantial information loss but corrects the data for peak shifts due to pH or ionic strength variation among samples. The other advantage of binning is significant data reduction which simplifies subsequent data analysis. Usually the bucket width is fixed to 0.04 ppm resulting in the reduction of the high resolution NMR spectrum from of 16 to 64 K data points to on average of 250 data points. Binned data can be directly imported into numerous statistical packages for multivariate analysis. The other approach to NMR data pre-processing is spectra alignment and peak picking. Several NMR spectral alignment algorithms can be found in the literature, including linear fit method, described by Vogels et al. (1996), or automatic removal of frequency shifts in spectra by PCA, described by Brown and Stoyanova (1996).

The other approach to NMR data processing is to deconvolute the spectra into individual components (see review by Schripsema (2010)). This approach has an advantage over other pre-processing approaches as it allows the identification and quantitation of individual components in the mixture from the complex NMR spectrum. Weljie et al. (2006) described a novel technique of deconvoluting complex spectra based on mathematical modeling of individual NMR resonances from pure compound spectra to create a component database followed by database search to identify and quantify metabolites in complex spectra of biological matrices. Authors defined this approach as "targeted profiling" and validated it against a "spectral binning" analysis. The technique proved to be very stable in PCA-based pattern recognition analysis, insensitive to water suppression effects, relaxation times, and scaling factors. As unambiguous compound identification from the complex one-dimensional ^{1}H-NMR spectrum can be complicated; several two-dimensional NMR techniques, including 2D-COSY (Correlation Spectroscopy), 2D-HMBC (Heteronuclear Multiple Bond Coherence), 2D-HSQC (Heteronuclear Single Quantum Coherence), 2D-TOCSY (Total Correlation Spectroscopy) (Ludwig et al. 2009), 2D-HRMAS (High Resolution Magic Angle Spinning)(Bayet-Robert et al. 2010) and 2D-JRES (*J*-resolved spectroscopy) (Fonville et al. 2010), were used in metabolomics studies to increase metabolite specificity and improve quantitation (Ludwig and Viant 2010; Schripsema 2010). Application of two-dimensional techniques in metabolomics studies is rather limited due to a long time required to acquire 2D spectrum (i.e., it requires close to 20 h to acquire typical 2D TOCSY spectrum to achieve the same sensitivity as 1D spectrum acquired in few minutes (Tang et al. 2004). 2D-JRES NMR spectroscopy provides analysis speed advantage over other two-dimensional techniques. It takes about 20 min to acquire a 2D spectrum of a metabolite mixture with relatively little convolution of signals at the same time significantly improving spectral assignment and accurate quantitation (Ludwig and Viant 2010).

Mass spectrometry data processing involves noise reduction, spectrum deconvolution, peak detection and integration, chromatogram alignment, component detection, identification and quantitation (Katajamaa and Orešič 2007). Raw mass spectrometry data, in addition to real

spectral data, contains background and noise. Background is a slowly varying part of the spectral signal, while noise includes rapid spikes in the intensity of the signal. To remove noise from data several algorithms have been developed and are currently implemented in both commercial and publicly available software. Typically, moving window filters are most often used to remove the noise from the data (Hansen 2007). Other noise filtering algorithms include median filtering (Hastings et al. 2002), Savitsky-Golay filter based on polynomial regression (Savitzky and Golay 1964), and wavelet transform (Chen et al. 2010; Coombes et al. 2005). Next step in raw data processing is spectral deconvolution. Deconvolution or component detection is used to separate two or more co-eluting (or overlapping) components in the mass spectral data. Several deconvolution algorithms are implemented in both commercial and public software. GC–MS data are often processed using AMDIS (Automated Mass Spectral Deconvolution and Identification System, http://chemdata.nist.gov/mass-spc/amdis/) software that utilizes well described algorithms, has proven to be extremely useful in processing of the GC–MS data (Halket et al. 1999). ESI-LC–MS data can be processed using variety of algorithms, including component detection algorithm (CODA) (Windig et al. 1996), "windowed mass selection method" (WMSM) (Fleming et al. 1999), singular value decomposition (SVD), sequential paired covariance (SPC) (Muddiman et al. 1995), or higher order sequential paired covariance (HO-SPC) (Muddiman et al. 1997).

Following the deconvolution process it is necessary to define, integrate and quantitate the peaks corresponding to individual components. Peak detection approaches in mass spectrometry data were recently reviewed by Katajamaa and Orešič (2007) who reviewed three major strategies to feature detection: (1) peak detection performed separately in two dimensions (retention time and m/z), (2) by extracting each individual ion chromatogram and processing them independently, and (3) fitting a model to the data. The first method identifies the features with the intensities above the defined threshold independently in the retention time and m/z directions and the ones that meat both threshold criteria are defined as peaks. Second method relies on identifying peaks in a discrete set of extracted ion chromatograms, each representing a small range of m/z values. The third strategy is based on fitting two- or three-dimensional model of isotope pattern to raw signal. All these approaches allow generating peak lists with quantitative information on individual metabolites/components for subsequent multivariate analysis of datasets.

A typical metabolomics experiment involves a large number of samples. Due to many variations, such as instrument or chromatographic column performance, buffer composition, matrix complexity, or environmental conditions, retention time fluctuates over a set of chromatograms. To correct for this retention time fluctuations, chromatogram should be aligned to compare same features in a dataset. Detailed review of the different alignment strategies and algorithm can be found in recent review of mass-spectrometry data processing in metabolomics by Katajamaa and Orešič (2007). Alignment allows comparing large datasets where samples were analyzed over period of time and sometime on different instruments.

4.2 Multivariate analysis

Metabolomics data can be analyzed with a wide range of statistical and machine-learning algorithms. These can be classified in two major classes: unsupervised and supervised algorithms (Mendes 2002).

Unsupervised methods that have been used in analyzing metabolomics data are principal component analysis (PCA) (Odunsi et al. 2005), hierarchical clustering (Eisen et al. 1998), and self-organizing maps (Tamayo et al. 1999). Supervised methods include ANOVA (Churchill 2004), partial least squares (PLS) (Musumarra et al. 2003), support vector machines, k-Nearest neighbors, and discriminant function analysis (DFA) (Raamsdonk et al. 2001). Several excellent textbooks (Hastie et al. 2001; Johnson and Wichern, 2007; Quinn and Keough 2002; Tettamanzi and Tomassini 2001; Wilcox 2005) provide extensive discussion on both the mathematical and practical aspects of the different statistical and machine learning algorithms for data analysis, mining, inference and prediction; therefore, here we will only outline major methods widely used for metabolomics data analysis and illustrate them with case studies from cancer research.

4.2.1 Data normalization, scaling and dimensionality reduction

Before statistical analysis metabolomics data have to be normalized to account for differences in metabolite recoveries during the extraction process or systematic errors due to instrument performance. Normalization can be achieved by either using single or multiple internal standards spiked into the sample prior to or during the extraction or by using various normalization factors (Sysi-Aho et al. 2007).

Metabolomics data, like other "omic" data, are underdetermined, meaning that they contain many more variables than samples (Kohane et al. 2003). In a typical "omic" experiment an average of several hundreds to tens of thousands of variables are measured (i.e., all the genes in a microarray experiment, or hundreds of metabolites in a metabolomics study), but only a relatively small number of samples are collected to examine this high-dimensional

space. For proper statistical analysis of these data it is necessary to reduce the number of variables in order to obtain uncorrelated features in the data. This can be best achieved either through significance methods in ANOVA and *t*-tests, through linear combinations of variables in PCA, or by using evolutionary algorithms such as genetic algorithms or genetic programming. Evolutionary algorithms are usually carried out in combination with a second analysis algorithm (e.g., PLS or DFA) that search for combinations of variables most effective in the secondary algorithm, and are guided by principles of evolution and selection of species (reviewed by Pena-Reyes and Sipper 2000). Evolutionary algorithms have been successfully applied to metabolomics data (Kell 2002).

4.2.2 *Principal component analysis (PCA)*

PCA is an unsupervised statistical data analysis method that is used for dimension reduction and visualization of the data. The goal is to find a way to represent high dimensional data by a projection into a small dimensional subspace, without losing the important features of the data.

PCA finds a small dimensional subspace such that the orthogonal projection of the data into this subspace moves the data points as little as possible. This produces a small dimensional representation of the original data that can be used for visualization or more sophisticated methods of data analysis. Since we are minimizing the displacement of the points, we can hope that the small dimensional representation captures some important features of the data.

An equivalent way of looking at PCA is that it finds a small dimensional subspace such that the orthogonal projection of the data into this subspace captures as much variance of the original data as possible.

More precisely, PCA starts by finding the one dimensional subspace that captures the most variance. This subspace is called the first Principal Component (PC). We think of the variance in the data that is not captured by the first PC as left-over variance. Next PCA finds the one dimensional subspace that captures as much left-over variance as possible. This subspace is called the second Principal Component. The process can be repeated to generate as many PCs as desired. It is also possible to have PCA generate as many principal components as necessary to capture a certain percentage of the variance in the data, as opposed to generating a fixed number of PCs.

Each Principal Component is orthogonal (perpendicular) to all other PCs. The best two dimensional subspace for capturing variance is the span of the first two PCs. The best three dimensional subspace is the span of the first three PCs and so on. Equivalently, these are the best subspaces for minimizing the displacement that orthogonal projection onto this subspace causes.

Advantages of using PCA: PCA has been extensively used in metabolomics data analysis and it is a well established method and can be performed with a variety of statistical analysis packages. PCA has a very natural geometrical interpretation in terms of minimizing the displacement caused by projection of the data onto a small dimensional subspace. It provides a good visualization tool for the data, by looking at the projection into the best two (or three) dimensional subspace. Since the Primary Components are orthogonal to each other, this is often referred to as plotting PCs against each other.

Downside of using PCA: The principal components are linear combinations of variables that explain the most variance. As such PCA is inherently biased toward selecting (assigning large coefficients/weights) the variables with large variance. The variables that are good differentiators but have relatively small variance are unlikely to be picked up by PCA. Since metabolomics strives to analyze the whole metabolome, it is actually likely that there are many variables unrelated to the problem in question that possess large variance. These variables will obscure the true differentiating variables when using PCA.

Another important disadvantage is that the Primary Components are linear combinations of all the variables. Therefore, PCA is not effective in singling out a small group of important metabolites. It is possible to use VARIMAX rotation to find new orthogonal axis that have the same span as PCs, but are combinations of fewer variables. However, this is not guaranteed to substantially improve results.

4.2.2.1 Example of using PCA: detecting epithelial ovarian cancer In Odunsi et al. (2005) metabolic profiles of human serum were used in detecting Epithelial Ovarian Cancer (EOC). Samples were obtained from 53 individuals with EOC, 12 patients with benign cysts and 38 healthy women. The samples were analyzed using ^{1}H-NMR spectroscopy and each spectrum was reduced to 200–250 integral segments of equal width. Each variable was Pareto-scaled to dampen the tendency of PCA to select the variables with highest variance. In Pareto scaling a variable is divided by the square root of its variance. This does not eliminate high variance entirely, which may be undesirable, but it gives variables with lower variances a better chance to be detected by PCA.

The data was separated into three different subsets: premenopausal women and cancer patients, postmenopausal women and cancer patients and one with benign cyst and cancer patients. PCA was applied to each of the subsets and two dimensional plots of PC1 versus PC2 were generated and analyzed.

The plots showed that the projection onto the first two Primary Components was effective in separating cancer

samples from non-cancer samples in each of the three cases. When all samples were analyzed together the patients with benign cysts overlapped with healthy patients, but there was separation between cancer and non-cancer samples. This suggests that metabolomics may be useful in early detection of epithelial ovarian cancer. The loadings (coefficients) used in PCA were studied in search of markers that distinguish EOC from normal samples. Several potential markers with high loadings such as sugar hydrogens and 3-hydroxybutyrate were identified, suggesting how the metabolic profile of cancer patients is altered by the disease.

4.2.3 Partial least squares (PLS)

PLS is a regression based method of data analysis. The underlying idea of PLS is that although we collect highly multidimensional data, the phenomenon under investigation can be explained using a relatively small number of factors. PLS computes these factors to which linear regression is then applied.

In data analysis linear regression is used to find the best linear predictor of some variables Y based on the dependent variables X (sample readings). For example, we can try to build a regression model to predict the risk of developing heart disease based on cholesterol readings and weight of the patient, along with other factors.

When we are concerned with classification problems, such as cancer versus non-cancer or distinguishing different types of cancer, we do not have quantitative measurements for the variable that we are trying to predict, only the division of samples into different classes. In this case we set the variable Y to have entry *0* for all samples in the first class, entry *1* for all samples in the second class and so on. In case of PLS, this is called PLS Discriminant Analysis or PLS-DA. Unlike PCA, this is a supervised method of data analysis, the separation of samples into different groups is crucial to building the model.

PLS differs from the usual linear regression in that PLS does not just predict independent variables Y based on the dependent variables X. Instead, PLS tries to find a small dimensional subspace, such that the projection into this subspace does not change the dependent variables X very much, and at the same time, the coordinates of this new subspace are good predictors of Y. This corresponds to the idea that the variables Y can be predicted using a small number of factors (thus a projection to a small dimensional subspace).

In this way PLS is related to the Principal Component Regression, where Principal Components of X from PCA are used to predict the independent variables Y. This also allows PLS to handle data sets with highly correlated variables, which is often the case in metabolomics. Linear Regression tends to not work well in these instances.

The output of PLS is similar to the output of PCA. PLS generates a list of orthogonal vectors (components), which we can think of as PLS version of Principal Components of PCA. The PLS components are the factors that "best explain" the behavior of independent variables Y and they span the subspace onto which the dependent variables X are projected. We can choose the number of PLS components to be used.

The projection of the data on the first several components can be used for visualization or dimension reduction in the same way as with PCA. The PLS components are further used to construct a linear regression model for the independent variables Y. The regression model is used in predicting the classes of unknown samples.

Advantages of using PLS: PLS is a well established method in the field of chemometrics and it has also been applied in bioinformatics, social sciences and other fields. There are available software packages that perform PLS, although it is not as ubiquitous as PCA. PLS does not have the same tendency to gravitate toward high variance variables that PCA does. However, it does not have the simple geometrical interpretation of PCA. PLS provides a tool for visualization and dimension reduction in the same way as PCA.

Downside of using PLS: PLS shares the other disadvantage of PCA in that its components are linear combinations of all the variables. It is often not possible to single out a small group of variables that are responsible for classification into different groups.

4.2.3.1 Example of using PLS: using ^{1}H-NMR-based metabolomics for prognosis of high risk leukemia patients Chronic lymphocytic leukemia (*CLL*) is a disease with varying clinical course and survival rates. Roughly one third of the patients require immediate treatment while another third do not require treatment and have long survival rates. The remaining third exhibits a passive phase followed by disease progression. An early diagnostic method for predicting disease progression is therefore of utmost importance.

The mutational status of the immunoglobulin heavy chain variable region (IGHV) of CLL cells has been shown to provide useful prognostic information. However, the cost and difficulty of IGHV sequencing led to a search for alternative prognostic markers.

MacIntyre and co-authors (MacIntyre et al. 2010) used metabolic profiles of serum of leukemia patients to provide an alternative method of predicting IGHV mutation. Twenty nine samples were gathered from untreated early stage leukemia patients along with nine control samples. Of the 29 leukemia samples 19 came from patients with

mutated IGHV region and ten from patients with non-mutated IGHV region.

The samples were analyzed using ^{1}H-NMR Spectroscopy. As a first step PCA was applied to the resulting data. However, PCA failed to reveal clustering based on disease status. Further analysis of the loadings of first two principal components revealed that majority of the variation came from differences in glucose concentration.

PLS discriminant analysis (PLS-DA) was then applied to the data. The analysis of the first three components revealed clear separation between CLL and healthy samples. The loadings were studied and it was found that increased levels of pyruvate, glutamate, proline, pyridoxine and decreased concentration of isoleucine in CLL patients were mainly responsible.

The possibility of using metabolomics analysis as a predictor of IGHV status was then analyzed. Again PCA was initially applied to the 19 CLL samples and mild clustering was observed. Further analysis with PLS-DA showed clear separation between the two groups. The PLS-DA loadings showed that the unmutated IGHV samples had higher levels of cholesterol, lactate, uridine and lower levels of pyridoxine and glycerol, among others. Further NMR quantification revealed statistically significant differences in concentrations of cholesterol, lactate, methionine and pyridoxine between mutated and unmutated IGHV patients.

4.2.4 Clustering

Cluster analysis, or clustering, refers to a whole host of different algorithms that aims at dividing observations into classes, or clusters based on a distance function. The goal of a clustering algorithm is to partition the data into groups so that the distances between the samples within each group are small when compared to distances between samples from different clusters.

The distance function is thought of as a measure of dissimilarity. Different distance functions may be used; the Euclidean distance is most often used in practice; Manhattan distance is sometimes used as well.

Clustering algorithms can be divided into two types: hierarchical and non-hierarchical. A non-hierarchical clustering algorithm simply divides the data into clusters. An example of this is K-means Clustering. We think of the mean of samples in a cluster as the center of this cluster. The K-means clustering algorithm divides the data into a prescribed number of clusters (K) so that for each sample the closest cluster center is the center of the cluster that it belongs to. The number of clusters K comes from the data, for example, if we want to divide into disease and healthy groups, we would use two clusters, but if we have different species or phenotypes then we can use a higher K.

In hierarchical clustering we start by thinking of each sample as its own cluster. Then two closest clusters are merged together and the process is repeated until all samples are in the same cluster. The way of deciding which two clusters are closest depends on the hierarchical clustering method used. For example, in Single Linkage hierarchical clustering the distance between cluster A and cluster B is the shortest distance between any sample in cluster A and any sample in cluster B. Many other measures are possible, some of which involve distance between centers of clusters as well distances between individual samples.

A hierarchical clustering method outputs a dendrogram, which records which clusters were joined together and at what distance. This provides a visual description of the evolution of the clusters.

Clustering methods are typically unsupervised. The samples are divided into groups based on the distance between them, without taking into account class labels. If the data divides into groups along the lines of class labels then this is a strong indication that the class division is reflected in the data.

Clustering has been extensively used in genomics studies and, therefore, can be readily adopted to study metabolomics data.

Advantages of using clustering: Clustering is a well established method and its various incarnations are available in a great variety of packages.

Downside of using clustering: Metabolomics data typically has variables of very different values. If the variables are not rescaled then the various clusters will be determined by variables with large variance. The problem will be similar to the one we encountered with using PCA: by its nature metabolomics data have many extraneous variables, many of which will possess large variance. These variables will strongly influence the performance of the clustering algorithms.

Genomics data is typically log-transformed, thereby greatly reducing the difference in values and variance between the variables. However, log-transformation is usually not performed with metabolomics data. A rescaling, such as Pareto scaling we have seen in the PCA application example, often needs to be performed to at least partially eliminate the difference in variance between the variables.

4.2.4.1 Example of using clustering: recognizing different cancer cell lines from metabolomics data In Cuperlovic-Culf et al. (2009) fuzzy K-means clustering was applied to metabolomics data from breast cancer cell lines gathered with ^{1}H NMR spectroscopy. “Fuzzy” refers to the fact that samples are not simply partitioned into K clusters, but instead each sample is assigned a vector of K numbers,

with each number denoting the probability that the sample belongs to the corresponding cluster. We can also think of these numbers as denoting the membership value between the sample and each of the K clusters; the higher the membership value the more likely the sample is to belong to that cluster.

Five cell lines were used for the experiment, two cell lines were grown from normal cells and three from cancer cells; two of the cancer cell lines were from invasive metastatic cancer and one from non-invasive cancer. Five replicates were gathered for each cell line and the samples were analyzed using ^{1}H NMR spectroscopy. The spectrum for each sample was separated into major peaks and the peaks were matched with compounds using existing literature. Each peak was then integrated to calculate the total presence for each compound.

When PCA was applied to these data and the two first principal components were plotted against each other it was observed that there was clear separation between the two normal cell lines and the three cancer cell lines. The two normal cell lines were also separated from each other. However the three cancer cell lines were mixed. Similar results were observed when applying hierarchical clustering: the two normal cell lines formed crisp separated clusters but the three cancer cell lines formed mixed clusters. When regular K-means clustering (here with $K = 5$ for the five cell lines) was applied the results were inconclusive with different cell lines mixing between clusters.

However, when the fuzzy K-means clustering was used the two normal cell lines as well as the non-invasive cancer cell line formed separate clusters based on the top membership value. The two invasive cell lines were mixed based on the top membership value. However, the invasive cell lines were clearly separated into two clusters based on the second highest membership value. This shows the potential of fuzzy clustering to not only divide the data into a fixed number of clusters but further subdivide it based on phenotypes or other relevant characteristics.

4.2.5 Self organized maps

Self-Organized Map is a method of two-dimensional visualization of the data. It is based on a specific kind of neural network, where the neurons are arranged in a planar or toroidal grid. We have already seen two methods of two dimensional visualization: plotting two Principal Components of PCA or PLS against each other. Both of these methods are simply linear projections of the data.

Self-Organizing Map also generates a two dimensional representation of the data, but it is capable of more complex pattern recognition. The goal of SOM is to find a good two-dimensional representation of the data that is capable of exploiting non-linear phenomena in the data. Since we want not to just build a classification "black box", but to also understand the underlying biology of the problem, a visualization of the data can provide very valuable insights.

SOM is an unsupervised method since it does not use class labels in the construction of the map. A rectangular or hexagonal grid of appropriate size needs to be chosen. A typical SOM procedure consists of learning, where the network is fed data from representative samples and visualization where the network arranges new data based on learned patterns.

An output of SOM is a 2-dimensional "map" where samples that are similar to each other according to the data analysis are placed in a similar region.

Advantages of using SOM: Self-Organizing Map provides a good visualization tool for the data. Unlike plots of Principal components of PCA and PLS it can be used to detect non-linear relationships in the data. It also does not suffer from the tendency to select variables based on high variance.

SOM is a very well established method with applications in many fields. The original application of SOM was to voice recognition, but it has been extensively applied in bioinformatics and medicine. There are a variety of software packages that will build SOM from data.

Downside of using SOM: While SOM is capable of exploiting non-linear relationships in the data, the individual variables (metabolites) that are most responsible for the classification are not always easily read off. While GA methods are specifically geared toward producing small classification models and it is possible to study the loadings in PCA or PLS, the most important metabolites for SOM are more hidden.

4.2.5.1 Example of SOM: examination of metabolic changes in breast cancer tissue In study by Beckonert and co-authors (Beckonert et al. 2003) metabolic changes in human breast cancer tissue were examined. A total of 88 samples were collected consisting of 49 cancer tissue of varying grade and 39 healthy samples. The samples were analyzed using ^{1}H NMR Spectroscopy. The resulting NMR data for each sample was a vector of 1,057 components, with 655 components corresponding to water-soluble metabolites and 402 components corresponding to lipids.

Subsequently, feature selection was performed using 3-Nearest Neighbor Clustering to identify 62 metabolites (features) that were promising in separating different grades of cancer and healthy tissue. Three Nearest Neighbor Clustering was used instead of PCA in hopes that some low variance metabolites would be selected and prove successful in differentiating different grades of breast cancer and healthy tissue.

A Self Organizing Map on a rectangular grid was built with the 62 selected parameters. The map provided distinct separation between the various grades of cancer. The upper part of the map was taken up by the control samples, with intermediate malignancy samples (grade 2) on the lower left and high malignancy samples (grade 3) on the lower right.

Furthermore, the 62 underlying metabolites were analyzed to see whether each individually provides a picture similar to the map built using the combination of all 62. For each metabolite a different "concentration map" was built. The concentration of a compound in a sample was shown on the map as shade of gray, with white corresponding to low concentration and black corresponding to high concentration. The compounds whose "concentration map" provided a picture similar to the SOM were judged to most responsible for the classification.

In water soluble metabolites it was seen that glucose and myo-inositol were among the best differentiators between healthy and malignant tissues, which was consistent with earlier findings. Other potentially interesting differentiating metabolites such as UDP sugar derivatives were identified.

In lipids there was a general increase in fatty acid concentration with tumor grade. Concentration of cholesterol and glycerol, for example, showed a strong propensity to increase in grade 3 tumors. The general picture of increase in fatty acids from healthy to tumor cells was again consistent with earlier findings and gives evidence to the hypothesis that fatty acids are synthesized within the tumor cells.

4.2.6 Genetic algorithms based methods

Genetic Algorithms (GA) methods are used to find small subsets of metabolites that are promising for distinguishing between groups of samples, such as identifying cancer versus healthy samples. Unlike the methods discussed previously, we not only want to build a classification model, but we want to require that the model only uses a few metabolites. The goal of these methods is more toward "biomarker discovery", rather than just a classification system.

When using PCA and PLS even if the data analysis method is effective in classifying samples into groups, it is not easy to single out a small subset of metabolites that is responsible for the separation. The Principal Components of PCA and PLS are linear combinations of all the variables, while in fact we expect to measure many parameters that are unrelated to our problem when looking at the entirety of the metabolome.

To address this problem, we can try to pick small subsets of variables, and build a classification model using just these variables. The method used for building the small model is up to the modeler; in practice Discriminant Function Analysis (DFA) has been popular. However, regardless of the method used, it is not computationally possible to exhaustively search even all four or five variable subsets, since the total number of variables we have is in hundreds or thousands. This is where Genetic Algorithms come in.

We do not need to try all the possible subsets, we just need to produce the subsets that perform best for classification purposes. Genetic Algorithms provide an optimization heuristic when exact optimization or exhaustive search are computationally impossible.

GA based methods start with a small population of subsets. For each subset a classification model is built, using a method such as DFA, and the accuracy of the model or some other related property is used as its score. Then the population is changed according to Genetic Algorithm methods to try to optimize the choice of subsets to provide the best classifying models. The algorithms typically run for a fixed number of steps after which the top scoring subset is produced.

Genetic algorithms consist of three major steps: selection, crossover and mutation. During selection a subset of the existing population is selected for reproduction. The selection is based on a fitness function, so in our case, the best performing classifying small subsets are selected. Typically, there is some randomness built in the selection process to keep the algorithm from settling on a globally poor local optimum.

In crossover a pair of "parent" subsets produces a new "child" subset. Then the subsets undergo mutation, where a random change is made to the subset with a very small probability. Once crossover and mutation are done, the algorithm is repeated, until a fixed number of iterations (generations) is reached.

Due to the stochastic nature of Genetic Algorithms we are unlikely to produce the same small subset of variables every time. Instead the algorithm is run many times and we analyze what variables are repeatedly chosen for the small classification models. The subsets of variables that are chosen together in a high percentage of models are deemed important for classification.

Advantages of using GA methods: The main advantage of GA based methods is that they produce classifying models that use a small number of distinguishing variables (metabolites). GA methods also do not suffer from the propensity toward selecting variables with high variance, which affects PCA.

Genetic Algorithm methods have been applied in many fields, including analysis of microarray data, sequence alignment, food science and civil engineering.

These methods also offer a lot of flexibility; the Genetic Algorithms can be coupled to any other supervised data analysis method to produce a classification model based on a small subset of variables. The scoring of the models can also be customized to improve performance.

Downside of using GA methods: Due to the large number of variables in metabolomics data it is typical that many small subsets are well suited for classification. Some subsets are present due to sheer randomness and small sample size, while some are biologically meaningful. This presents an added challenge in metabolomics since the majority of peaks are unidentified. Therefore, it is not always possible to decide whether a small subset has biological meaning without a major effort in terms of compound detection. However, this is still a much better situation than having a linear combination of all the variables, in which case we cannot hope to identify all of the metabolites.

The flexibility of GA models also has a downside since there is no standard protocol and particular choices in the implementation of the Genetic Algorithm, the classification method to be used, or method used to score the models will influence performance of the data analysis. There is available software that can perform, for example, GA-DFA, but to our knowledge there are no packages that would allow users to actually changes some of the parameters we listed above and to see what impact these changes will have.

4.2.6.1 Example of using genetic algorithms combined with clustering: diagnosing ovarian cancer from serum In Petricoin et al. (2002) Genetic Algorithms were combined with clustering techniques to produce a method for detection of early stage ovarian cancer based on patients' serum. The method was constructed based on a data set of 100 patients, with 50 having various stages of ovarian cancer and 50 control patients. None of the control samples came from patients with gynecological disease or non-gynecological inflammatory disorders. The serum was analyzed using a SELDI-TOF mass spectrometer and for each of the 100 samples a range of 15,200 m/z ratios with corresponding intensities was calculated.

A genetic algorithm was applied to find a small subset of m/z ratios that best separates the cancer samples from controls. The algorithm started with hundreds of random choices of discriminating subsets consisting of five to twenty m/z ratios each. The discriminating power of each subset was analyzed using a clustering fitness function. The best fit subsets were selected and their m/z ratios were reshuffled to form new subsets until a fully discriminating subset emerged.

The validation was performed using a set of 116 masked samples, with 50 coming from patients with ovarian cancer and 66 controls. The clustering technique, using the discriminating m/z ratios constructed from the initial set, was applied to the classification of the samples in the masked set. The samples were classified as either healthy, cancer or new cluster if they were outside the margin error of either cancer or healthy clusters of the initial set. The above method correctly identified all 50 patients with ovarian cancer. Of the 66 control samples 9 out of 10 patients with gynecological disease were put into a new cluster, while the 10th was put into the healthy cluster. Seven out of seven patients with non-gynecological inflammatory disorder were placed into a new cluster. For the patients with benign ovarian cysts, 24 out of 25 were correctly classified as healthy and one classified as cancer. In the remaining 24 control samples 22 were correctly identified as healthy and two were identified as cancer. The method achieved 100% sensitivity on both the initial set and the masked set and 95% specificity. This example shows the power of genetic algorithm methods, combined with supervised data analysis, to find a small discriminating subset and build a predictive model based on this subset.

5 Data integration

The proper tools to integrate data from different "omics" platforms is important as "omics" research is more widely used as part of a systems biology approach and high throughput platforms generate data for mathematical modeling biochemical networks (for a recent review see Mehrotra and Mendes 2006). The ability to analyze data obtained at different levels, including transcripts, proteins or metabolites, can provide deeper mechanistic insight into biological systems. Integrated analysis of metabolite and transcript or metabolite and protein levels has been used in many systems and already identified important features of metabolic regulation on different levels. Currently, most metabolomics studies use largely either one, or a combination of two approaches, while integrated studies using a combination of all three approaches are just appearing. This, in many ways, is limited by the lack of proper data analysis tools for integrated analysis of data from multiple levels.

6 Software for metabolomics

By its nature metabolomics requires automated data processing solutions. Although a series of commercial and public tools exist, none of them provides a comprehensive solution to meet the challenges of metabolomics. Selected commercial and freely available software for metabolomics is listed in Tables 1 and 2. More detail on each software package can be found in a recent review by (Katajamaa and Orešič 2007) and in the references listed in the tables.

Many commercial software packages provide tools for basic raw data processing as well as some kind of statistical data analysis. Some packages incorporate unique and

Table 1 Selected commercial software for metabolomics

Software	Vendor	Vendor web site
ACD MS Manager with IntelliXtract	ACD/Labs	http://www.acdlabs.com
ChromaTOF	LECO	www.leco.com
Genespring-MS	Agilent	http://www.opengenomics.com/
Ion Signature Quantitative Deconvolution Software for Mass Spectrometry	Ion Signature Technology, Inc.	http://ionsigtech.com/index.php
Ingenuity Pathways Analysis (IPA)-IPA-Metabolomics™ Analysis	Ingenuity Systems	http://www.ingenuity.com/products/ipa-metabolomics.html
MarkerLynx	Waters	http://www.waters.com/
MarkerView	AB Sciex	http://www3.appliedbiosystems.com/i
Mass Frontier	ThermoFisher	http://www.thermo.com/
Metabolomics Edition	Bio-Rad	http://www.bio-rad.com
SIEVE	ThermoFisher	http://www.thermo.com/

Table 2 Selected open access tools for complex LC/MS data analysis

Software name	Reference	Web address
BL-SOM	(Kanaya et al. 2001)	http://prime.psc.riken.jp/?action=blsom_index
Chrompare	(Frenzel et al. 2003)	http://www.chrompare.com/chrompare/
COMSPARI	(Katz et al. 2004)	http://www.biomechanic.org/comspari/
MathDAMP	(Baran et al. 2006)	
MeMo	(Spasic et al. 2006)	http://dbkgroup.org/memo/
MET-IDEA	(Broeckling et al. 2006)	http://noble.org/
MSFACTs	(Duran et al. 2003)	http://noble.org/
MZmine	(Katajamaa and Orešič 2005)	http://mzmine.sourceforge.net/
TagFinder	(Luedemann et al. 2008)	http://www-en.mpimp-golm.mpg.de/03-research/researchGroups/01-dept1/Root_Metabolism/smp/TagFinder/index.html
XCMS	(Smith et al. 2006)	http://metlin.scripps.edu/download/
XCMS2	(Benton et al. 2008)	http://mathdamp.iab.keio.ac.jp/

powerful algorithms for data analysis not found in other commercial or public software. The major limitation of most commercial software supplied by equipment manufacturers is that it only works with a specific data format. LECO Corporation's ChromaTOF software package, for example, provides superior deconvolution algorithms, but it only works with the proprietary file format generated by the LECO GC-TOF and LC-TOF instruments. This is a major limitation for many metabolomics laboratories that often utilize multiple analytical platforms and employ instrumentation from different vendors. The other limitation of the commercial software is that often the description of a particular data processing algorithm is not available.

Recent development of publically available software packages, including MZmine (Katajamaa et al. 2006), XCMS (Smith et al. 2006), XCMS^2 (Benton et al. 2008), MathDAMP (Baran et al. 2006), and Met-IDEA (Broeckling et al. 2006), expands vendor independent bioinformatics solution for metabolomics data analysis. MZmine software (Katajamaa et al. 2006) employs a modular infrastructure with the ability to integrate new algorithms and applications. Another publically available package, XCMS (Smith et al. 2006), is implemented in R language and is also available for download under GNU General Public License. It provides noise filtering, peak detection, and non-linear spectral alignment algorithms as well as statistical analysis of the data. The software can process both GC–MS and LC–MS data. Recent extension of the XCMS package called XCMS^2 added the capability of automated searching of high quality MS–MS data against METLIN database (Benton et al. 2008). In addition to specialized software packages that were designed for metabolomics application, several mass spectrometry data processing software packages that were developed for proteomics can also be used to process metabolomics data. These include, among others, SpecArray (Li et al. 2005,

http://tools.proteomecenter.org/wiki/index.php?title=Software:SpecArray), MSight (Palagi et al. 2005, http://www.expasy.org/MSight/), and MapQuant (Leptos et al. 2006, http://genepath.med.harvard.edu/mw/MapQuant).

7 Summary

Metabolomics, being a relatively new area of genomics research, is rapidly gaining acceptance in many areas of biomedical research, including cancer research. Recent studies on metabolome changes during cancer development and progression have already shown the feasibility of using metabolomics for cancer diagnostics and prognosis and identifying new targets for anticancer therapy (Bathe et al. 2010; Borgan et al. 2010; Cascante et al. 2010; Catchpole et al. 2009; Howell, 2010; Madhok et al. 2010; Mamas et al. 2010; Slupsky et al. 2010; Sreekumar et al. 2009; Wang et al. 2010; Zitvogel et al. 2010). Despite the availability of many chemometric, statistical, and machine learning tools for the analysis of metabolomics data, many of them have important limitations, and, therefore, there is an urgent need for better tools and software. Further progress in cancer metabolomics greatly depends on the improvement of analytical and bioinformatics platform to improve sensitivity, specificity, metabolome coverage and provide spatial and temporal resolution for important metabolic changes in normal in diseased state.

Acknowledgments The research in the authors' lab is financially supported by the NIH (National Institutes of Health) NCI (National Cancer Institute) grant R01CA120170. FMT and SVT were supported in part by NIH NIDDK (National Institutes of Diabetes and Digestive and Kidney Diseases) grant R37 DK42412.

Bibliography

Asiago, V. M., Alvarado, L. Z., Shanaiah, N., Gowda, G. A. N., Owusu-Sarfo, K., Ballas, R. A., et al. (2010). Early detection of recurrent breast cancer using metabolite profiling. *Cancer Research, 70*, 8309–8318.

Baran, R., Kochi, H., Saito, N., Suematsu, M., Soga, T., Nishioka, T., et al. (2006). MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics, 7*, 530.

Bathe, O. F., Shaykhutdinov, R., Kopciuk, K., Weljie, A. M., McKay, A., Sutherland, F. R., Dixon, E., Dunse, N., Sotiropoulos, D., & Vogel, H. J. (2010). Feasibility of identifying pancreatic cancer based on serum metabolomics. *Cancer Epidemiology Biomarkers Prevention*. doi:10.1158/1055-9965.EPI-10-0712.

Bayet-Robert, M., Loiseau, D., Rio, P., Demidem, A., Barthomeuf, C., Stepien, G., et al. (2010). Quantitative two-dimensional HRMAS 1H-NMR spectroscopy-based metabolite profiling of human cancer cell lines and response to chemotherapy. *Magnetic Resonance in Medicine, 63*, 1172–1183.

Beckonert, O., Coen, M., Keun, H. C., Wang, Y., Ebbels, T. M., Holmes, E., et al. (2010). High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. *Nature Protocols, 5*, 1019–1032.

Beckonert, O., Monnerjahn, J., Bonk, U., & Leibfritz, D. (2003). Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. *NMR in Biomedicine, 16*, 1–11.

Beckwith-Hall, B. M., Nicholson, J. K., Nicholls, A. W., Foxall, P. J., Lindon, J. C., Connor, S. C., et al. (1998). Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins. *Chemical Research in Toxicology, 11*, 260–272.

Benton, H. P., Wong, D. M., Trauger, S. A., & Siuzdak, G. (2008). XCMS(2): Processing tandem mass spectrometry data for metabolite identification and structural characterization. *Analytical Chemistry, 80*, 6382–6389.

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science, 9*, 418–425.

Borgan, E., Sitter, B., Lingjaerde, O. C., Johnsen, H., Lundgren, S., Bathen, T. F., et al. (2010). Merging transcriptomics and metabolomics—advances in breast cancer profiling. *BMC Cancer, 10*, 628.

Boros, L. G., Brackett, D. J., & Harrigan, G. G. (2003). Metabolic biomarker and kinase drug target discovery in cancer using stable isotope-based dynamic metabolic profiling (SIDMAP). *Current Cancer Drug Targets, 3*, 445–453.

Broeckling, C. D., Reddy, I. R., Duran, A. L., Zhao, X., & Sumner, L. W. (2006). MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics. *Analytical Chemistry, 78*, 4334–4341.

Brown, T. R., & Stoyanova, R. (1996). NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts. *Journal of Magnetic Resonance. Series B, 112*, 32–43.

Byrum, S., Montgomery, C. O., Nicholas, R. W., & Suva, L. J. (2010). The promise of bone cancer proteomics. *Annals of the New York Academy of Sciences, 1192*, 222–229.

Cascante, M., Benito, A., Zanuy, M., Vizan, P., Marin, S., & de Atauri, P. (2010). Metabolic network adaptations in cancer as targets for novel therapies. *Biochemical Society Transactions, 38*, 1302–1306.

Castle, A. L., Fiehn, O., Kaddurah-Daouk, R., & Lindon, J. C. (2006). Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Briefings in Bioinformatics, 7*, 159–165.

Catchpole, G., Platzer, A., Weikert, C., Kempkensteffen, C., Johannsen, M., Krause, H., Jung, K., Miller, K., Willmitzer, L., Selbig, J., & Weikert, S. (2009). Metabolic profiling reveals key metabolic features of renal cell carcinoma. Journal of Cellular and Molecular Medicine. doi:10.1111/j.1582-4934.2009.00939.x.

Chari, R., Thu, K. L., Wilson, I. M., Lockwood, W. W., Lonergan, K. M., Coe, B. P., et al. (2010). Integrating the multiple dimensions of genomic and epigenomic landscapes of cancer. *Cancer and Metastasis Reviews, 29*, 73–93.

Chen, H. P., Liao, H. J., Huang, C. M., Wang, S. C., & Yu, S. N. (2010). Improving liquid chromatography-tandem mass spectrometry determinations by modifying noise frequency spectrum between two consecutive wavelet-based low-pass filtering procedures. *Journal of Chromatography. A, 1217*, 2804–2811.

Chen, Y., Zhang, R., Song, Y., He, J., Sun, J., Bai, J., et al. (2009). RRLC-MS/MS-based metabonomics combined with in-depth analysis of metabolic correlation network: Finding potential biomarkers for breast cancer. *Analyst, 134*, 2003–2011.

Cheng, L. L., Lean, C. L., Bogdanova, A., Wright, S. C., Jr., Ackerman, J. L., Brady, T. J., et al. (1996). Enhanced resolution of proton NMR spectra of malignant lymph nodes using magic-angle spinning. *Magnetic Resonance in Medicine, 36*, 653–658.

Churchill, G. A. (2004). Using ANOVA to analyze microarray data. *Biotechniques, 37*, 173–175.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M. C., & Kuerer, H. M. (2005). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics, 5*, 4107–4117.

Cuperlovic-Culf, M., Belacel, N., Culf, A. S., Chute, I. C., Ouellette, R. J., Burton, I. W., et al. (2009). NMR metabolic analysis of samples using fuzzy K-means clustering. *Magnetic Resonance in Chemistry, 47*(Suppl 1), S96–S104.

Devos, A., Lukas, L., Suykens, J. A., Vanhamme, L., Tate, A. R., Howe, F. A., et al. (2004). Classification of brain tumours using short echo time 1H MR spectra. *Journal of Magnetic Resonance, 170*, 164–175.

Duran, A. L., Yang, J., Wang, L., & Sumner, L. W. (2003). Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics, 19*, 2283–2293.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA, 95*, 14863–14868.

Fiehn, O. (2002). Metabolomics-the link between genotypes and phenotypes. *Plant Molecular Biology, 48*, 155–171.

Fink-Retter, A., Czerwenka, K., Gschwantler-Kaulich, D., Hudelist, G., Pischinger, K., Manavi, M., et al. (2009). Proteomics in mammary cancer research. *European Journal of Gynaecological Oncology, 30*, 635–639.

Fleming, C. M., Kowalski, B. R., Apffel, A., & Hancock, W. S. (1999). Windowed mass selection method: A new data processing algorithm for liquid chromatography-mass spectrometry data. *Journal of Chromatography. A, 849*, 71–85.

Fonville, J. M., Maher, A. D., Coen, M., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2010). Evaluation of full-resolution J-resolved 1H NMR projections of biofluids for metabonomics information retrieval and biomarker identification. *Analytical Chemistry, 82*, 1811–1821.

Frenzel, T., Miller, A., & Engel, K. H. (2003). A methodology for automated comparative analysis of metabolite profiling data. *European Food Research and Technology, 216*, 335–342.

Griffin, J. L., & Steinbeck, C. (2010). So what have data standards ever done for us? The view from metabolomics. *Genome Medicine, 2*, 38.

Halket, J. M., Przyborowska, A., Stein, S. E., Mallard, W. G., Down, S., & Chalmers, R. A. (1999). Deconvolution gas chromatography/mass spectrometry of urinary organic acids-potential for pattern recognition and automated identification of metabolic disorders. *Rapid Communications in Mass Spectrometry, 13*, 279–284.

Hansen, M. A. E. (2007). Data analysis. In S. G. Villas-Boas & U. Roessner (Eds.), *Metabolome analysis: An introduction* (pp. 146–187). Hoboken, NJ: Wiley.

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

Hastings, C. A., Norton, S. M., & Roy, S. (2002). New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Communications in Mass Spectrometry, 16*, 462–467.

Howell, A. (2010). Can metabolomics in addition to genomics add to prognostic and predictive information in breast cancer? *BMC Medicine, 8*, 73.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A. R., Taylor, J., et al. (2004). A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology, 22*, 1601–1606.

Johnson, H. E., Broadhurst, D., Goodacre, R., & Smith, A. R. (2003). Metabolic fingerprinting of salt-stressed tomatoes. *Phytochemistry, 62*, 919–928.

Johnson, R., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Pearson Prentice Hall.

Jordan, K. W., & Cheng, L. L. (2007). NMR-based metabolomics approach to target biomarkers for human prostate cancer. *Expert Review of Proteomics, 4*, 389–400.

Kanaya, S., Kinouchi, M., Abe, T., Kudo, Y., Yamada, Y., Nishi, T., et al. (2001). Analysis of codon usage diversity of bacterial genes with a self-organizing map (SOM): Characterization of horizontally transferred genes with emphasis on the E. coli O157 genome. *Gene, 276*, 89–99.

Katajamaa, M., Miettinen, J., & Orešič, M. (2006). MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics, 22*, 634–636.

Katajamaa, M., & Orešič, M. (2005). Processing methods for differential analysis of LC/MS profile data. *BMC Bioinformatics, 6*, 179.

Katajamaa, M., & Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of Chromatography. A, 1158*, 318–328.

Katz, J. E., Dumlao, D. S., Clarke, S., & Hau, J. (2004). A new technique (COMSPARI) to facilitate the identification of minor compounds in complex mixtures by GC/MS and LC/MS: Tools for the visualization of matched datasets. *Journal of The American Society for Mass Spectrometry, 15*, 580–584.

Kell, D. B. (2002). Metabolomics and machine learning: Explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Molecular Biology Reports, 29*, 237–241.

Kim, D. H., Jarvis, R. M., Xu, Y., Oliver, A. W., Allwood, J. W., Hampson, L., et al. (2010). Combining metabolic fingerprinting and footprinting to understand the phenotypic response of HPV16 E6 expressing cervical carcinoma cells exposed to the HIV anti-viral drug lopinavir. *Analyst, 135*, 1235–1244.

Kind, T., Tolstikov, V., Fiehn, O., & Weiss, R. H. (2007). A comprehensive urinary metabolomic approach for identifying kidney cancerr. *Analytical Biochemistry, 363*, 185–195.

Kohane, I. S., Kho, A. T., & Butte, A. J. (2003). *Microarrays for integrative genomics. A Bradford book*. Cambridge: The MIT Press.

Korkola, J., & Gray, J. W. (2010). Breast cancer genomes-form and function. *Current Opinion in Genetics and Development, 20*, 4–14.

Larkin, S. E., Zeidan, B., Taylor, M. G., Bickers, B., Al-Ruwaili, J., Aukim-Hastie, C., et al. (2010). Proteomics in prostate cancer biomarker discovery. *Expert Review of Proteomics, 7*, 93–102.

Leptos, K. C., Sarracino, D. A., Jaffe, J. D., Krastins, B., & Church, G. M. (2006). MapQuant: Open-source software for large-scale protein quantification. *Proteomics, 6*, 1770–1782.

Li, X. J., Yi, E. C., Kemp, C. J., Zhang, H., & Aebersold, R. (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics, 4*, 1328–1340.

Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C., Craig, A., Pearce, J. T., et al. (2005). Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology, 23*, 833–838.

Lommen, A., Weseman, J. M., Smith, G. O., & Noteborn, H. P. J. M. (1998). On the detection of environmental effects on complex matrices combining off-line liquid chromatography and 1H-NMR. *Biodegradation, 9*, 513–525.

Ludwig, C., & Viant, M. R. (2010). Two-dimensional J-resolved NMR spectroscopy: Review of a key methodology in the metabolomics toolbox. *Phytochemical Analysis, 21*, 22–32.

Ludwig, C., Ward, D. G., Martin, A., Viant, M. R., Ismail, T., Johnson, P. J., et al. (2009). Fast targeted multidimensional NMR metabolomics of colorectal cancer. *Magnetic Resonance in Chemistry, 47*(Supp 1), S68–S73.

Luedemann, A., Strassburg, K., Erban, A., & Kopka, J. (2008). TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics, 24*, 732–737.

Lukas, L., Devos, A., Suykens, J. A., Vanhamme, L., Howe, F. A., Majos, C., et al. (2004). Brain tumor classification based on long echo proton MRS signals. *Artificial Intelligence in Medicine, 31*, 73–89.

MacIntyre, D. A., Jimenez, B., Lewintre, E. J., Martin, C. R., Schafer, H., Ballesteros, C. G., et al. (2010). Serum metabolome analysis by 1H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. *Leukemia, 24*, 788–797.

Madhok, B. M., Yeluri, S., Perry, S. L., Hughes, T. A., & Jayne, D. G. (2010). Targeting glucose metabolism: An emerging concept for anticancer therapy. *American Journal of Clinical Oncology*. doi:10.1097/COC.0b013e3181e84dec.

Madsen, R., Lundstedt, T., & Trygg, J. (2010). Chemometrics in metabolomics—a review in human disease diagnosis. *Analytica Chimica Acta, 659*, 23–33.

Mamas, M., Dunn, W. B., Neyses, L., & Goodacre, R. (2010). The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology*. doi:10.1007/s00204-010-0609-6.

Mehrotra, B., & Mendes, P. (2006). Bioinformatics approaches to integrate metabolomics and other systems biology data. In K. Saito, R. A. Dixon, & L. Willmitzer (Eds.), *Plant metabolomics* (pp. 105–115). Berlin, Heidelberg: Springer-Verlag.

Mendes, P. (2002). Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics, 3*, 134–145.

Merz, A. L., & Serkova, N. J. (2009). Use of nuclear magnetic resonance-based metabolomics in detecting drug resistance in cancer. *Biomarkers in Medicine, 3*, 289–306.

Muddiman, D. C., Huang, B. M., Anderson, G. A., Rockwood, A., Hofstadler, S. A., WeirLipton, M. S., et al. (1997). Application of sequential paired covariance to liquid chromatography mass spectrometry data—Enhancements in both the signal-to-noise ratio and the resolution of analyte peaks in the chromatogram. *Journal of Chromatography A, 771*, 1–7.

Muddiman, D. C., Rockwood, A. L., Gao, Q., Severs, J. C., Udseth, H. R., Smith, R. D., et al. (1995). Application of sequential paired covariance to capillary electrophoresis electrospray-ionization time-of-flight mass-spectrometry—Unraveling the signal from the noise in the electropherogram. *Analytical Chemistry, 67*, 4371–4375.

Musumarra, G., Barresi, V., Condorelli, D. F., & Scire, S. (2003). A bioinformatic approach to the identification of candidate genes for the development of new cancer diagnostics. *The Journal of Biological Chemistry, 384*, 321–327.

Odunsi, K., Wollman, R. M., Ambrosone, C. B., Hutson, A., McCann, S. E., Tammela, J., et al. (2005). Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *International Journal of Cancer, 113*, 782–788.

Palagi, P. M., Walther, D., Quadroni, M., Catherinet, S., Burgess, J., Zimmermann-Ivol, C. G., et al. (2005). MSight: An image analysis software for liquid chromatography-mass spectrometry. *Proteomics, 5*, 2381–2384.

Patterson, A. D., Lanz, C., Gonzalez, F. J., & Idle, J. R. (2010). The role of mass spectrometry-based metabolomics in medical countermeasures against radiation. *Mass Spectrometry Reviews, 29*, 503–521.

Pena-Reyes, C. A., & Sipper, M. (2000). Evolutionary computation in medicine: An overview. *Artificial Intelligence in Medicine, 19*, 1–23.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet, 359*, 572–577.

Qiu, Y., Cai, G., Su, M., Chen, T., Liu, Y., Xu, Y., et al. (2010). Urinary metabonomic study on colorectal cancer. *Journal of Proteome Research, 9*, 1627–1634.

Quinn, G. P., & Keough, M. J. (2002). *Experimental design and data analysis for biologists*. Cambridge: Cambridge University Press.

Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., et al. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology, 19*, 45–50.

Sansone, S. A., Fan, T., Goodacre, R., Griffin, J. L., Hardy, N. W., Kaddurah-Daouk, R., et al. (2007). The metabolomics standards initiative. *Nature Biotechnology, 25*, 846–848.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*, 1627–1639.

Scalbert, A., Brennan, L., Fiehn, O., Hankemeier, T., Kristal, B. S., van Ommen, B., et al. (2009). Mass-spectrometry-based metabolomics: Limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics, 5*, 435–458.

Schripsema, J. (2010). Application of NMR in plant metabolomics: Techniques, problems and prospects. *Phytochemical Analysis, 21*, 14–21.

Serkova, N. J., & Glunde, K. (2009). Metabolomics of cancer. *Methods in Molecular Biology, 520*, 273–295.

Shulaev, V. (2006). Metabolomics technology and bioinformatics. *Briefings in Bioinformatics, 7*, 128–139.

Slupsky, C. M., Steed, H., Wells, T., Dabbs, K., Schepansky, A., Capstick, V., Faught, W., & Sawyer, M. B. (2010). Urine metabolite analysis offers potential early diagnosis of ovarian and breast cancers. *Clinical Cancer Research*. doi:10.1158/1078-0432.CCR-10-1434.

Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry, 78*, 779–787.

Soga, T. (2007). Capillary electrophoresis-mass spectrometry for metabolomics. *Methods in Molecular Biology, 358*, 129–137.

Spasic, I., Dunn, W. B., Velarde, G., Tseng, A., Jenkins, H., Hardy, N., et al. (2006). MeMo: A hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics, 7*, 281.

Spraul, M., Neidig, P., Klauck, U., Kessler, P., Holmes, E., Nicholson, J. K., et al. (1994). Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples. *Journal of Pharmaceutical and Biomedical Analysis, 12*, 1215–1225.

Sreekumar, A., Poisson, L. M., Rajendiran, T. M., Khan, A. P., Cao, Q., Yu, J., et al. (2009). Metabolomic profiles delineate

potential role for sarcosine in prostate cancer progression. *Nature, 457*, 910–914.

Sugimoto, M., Wong, D. T., Hirayama, A., Soga, T., & Tomita, M. (2010). Capillary electrophoresis mass spectrometry-based saliva metabolomics identified oral, breast and pancreatic cancer-specific profiles. *Metabolomics, 6*, 78–95.

Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: Large-scale phytochemistry in the functional genomics era. *Phytochemistry, 62*, 817–836.

Sysi-Aho, M., Katajamaa, M., Yetukuri, L., & Orešič, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics, 8*, 93.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., et al. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences USA, 96*, 2907–2912.

Tang, H., Wang, Y., Nicholson, J. K., & Lindon, J. C. (2004). Use of relaxation-edited one-dimensional and two dimensional nuclear magnetic resonance spectroscopy to improve detection of small metabolites in blood plasma. *Analytical Biochemistry, 325*, 260–272.

Tate, A. R., Foxall, P. J., Holmes, E., Moka, D., Spraul, M., Nicholson, J. K., et al. (2000). Distinction between normal and renal cell carcinoma kidney cortical biopsy samples using pattern recognition of (1)H magic angle spinning (MAS) NMR spectra. *NMR in Biomedicine, 13*, 64–71.

Tate, A. R., Griffiths, J. R., Martinez-Perez, I., Moreno, A., Barba, I., Cabanas, M. E., et al. (1998). Towards a method for automated classification of 1H MRS spectra from brain tumours. *NMR in Biomedicine, 11*, 177–191.

Tettamanzi, A., & Tomassini, M. (2001). *Soft computing: Integrating evolutionary, neural, and fuzzy systems*. Berlin: Springer.

Urayama, S., Zou, W., Brooks, K., & Tolstikov, V. (2010). Comprehensive mass spectrometry based metabolic profiling of blood plasma reveals potent discriminatory classifiers of pancreatic cancer. *Rapid Communications in Mass Spectrometry, 24*, 613–620.

Vogels, J., Tas, A. C., Venekamp, J., & VanderGreef, J. (1996). Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *Journal of Chemometrics, 10*, 425–438.

Wang, H., Tso, V. K., Slupsky, C. M., & Fedorak, R. N. (2010). Metabolomics and detection of colorectal cancer in humans: A systematic review. *Future Oncology, 6*, 1395–1406.

Want, E. J., Wilson, I. D., Gika, H., Theodoridis, G., Plumb, R. S., Shockcor, J., et al. (2010). Global metabolic profiling procedures for urine using UPLC-MS. *Nature Protocols, 5*, 1005–1018.

Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted profiling: Quantitative analysis of 1H NMR metabolomics data. *Analytical Chemistry, 78*, 4430–4442.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Burlington, MA: Elsevier Academic Press.

Windig, W., Phalp, J. M., & Payne, A. W. (1996). A noise and background reduction method for component detection in liquid chromatography/mass spectrometry. *Analytical Chemistry, 68*, 3602–3606.

Yang, C., Richardson, A. D., Smith, J. W., & Osterman, A. (2007). Comparative metabolomics of breast cancer. *Pacific Symposium on Biocomputing, 12*, 181–192.

Yang, Q., Shi, X., Wang, Y., Wang, W., He, H., Lu, X., et al. (2010). Urinary metabonomic study of lung cancer by a fully automatic hyphenated hydrophilic interaction/RPLC-MS system. *Journal of Separation Science, 33*, 1495–1503.

Zitvogel, L., Kepp, O., Aymeric, L., Ma, Y., Locher, C., Delahaye, N. F., et al. (2010). Integration of host-related signatures with cancer cell-derived predictors for the optimal management of anticancer chemotherapy. *Cancer Research, 70*, 1–6.