Chapter 9
NMR-Based Metabolomics

9.1 Introduction

Metabolomics is a relatively new and emerging field of the omics research compared to other well-established platforms (genomics, transcriptomics, and proteomics) and is rapidly growing as evidenced by the increasing number of publications in this field (Fig. 9.1). It studies the global profiles of metabolites in a biological system (cell, tissue, or organism) under a given set of conditions (Goodacre et al. 2004). Arguably, its history can be traced back to ancient times (1500–2000 BC) when traditional Chinese doctors used ants to detect high concentrations of glucose in patient's urine for diagnosing diabetes (Van der Greef and Smilde 2005). The concept that individuals might have different "metabolic patterns" that can be detected in their biological fluids was first introduced by Roger Williams in the late 1940s (Williams 1956; Gates and Sweeley 1978). Since then, several terms (or definitions) have been proposed to describe the field of metabolomics. "Metabolic profile" was introduced by Horning and Horning (1971) to describe the quantitative measurement of metabolite concentrations in urine. "Metabolome" was proposed by Oliver et al. (1998) as referring to the complete set of small-molecule ($<$1 kDa) endogenous metabolites in an organism, and "metabonomics" by Nicholson et al. as "the quantitative measurement of the dynamic multiparametric response to living systems to pathophysiological stimuli or genetic modification" (Nicholson et al. 1999). Fiehn subsequently extended "metabolome" terminology to metabolomics as the comprehensive and quantitative analysis of all metabolites of an organism (Fiehn 2001). Although these terms are frequently used interchangeably, there is a growing consensus that the field is named as metabolomics, as reflected by the establishment of the Metabolomics Society (an international society) in 2004 and its official journal *Metabolomics* in 2005.

Metabolomics is considered as a complementary tool to other omics platforms because it more directly reflects cellular physiological states as being the most downstream in the omics family (Fig. 9.2). While a genome contains all genetic information of a living organism, the information of genes is expressed to produce
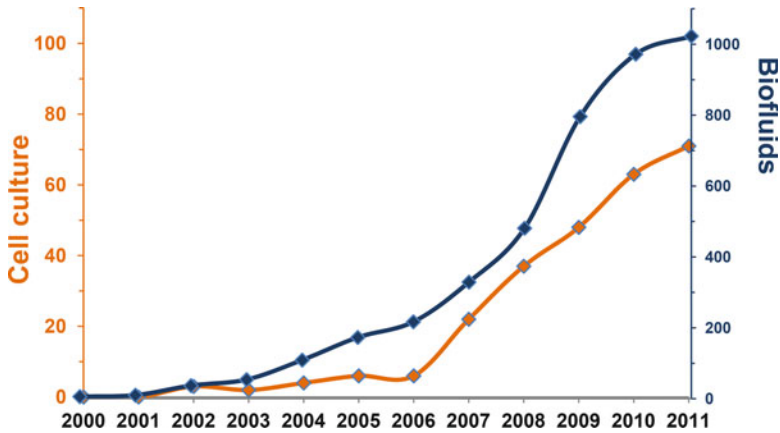
Fig. 9.1 Annual numbers of publications in the field of metabolomics from search results of PubMed database. The numbers present the rapidly growing rate of metabolomics research
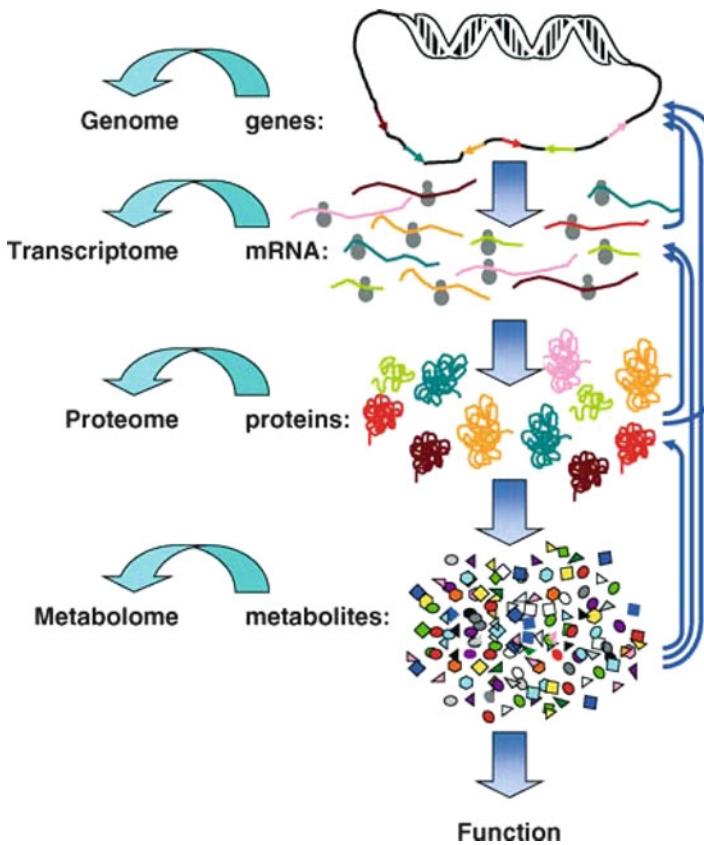

Fig. 9.2 Schematic overview of omics technologies. The information flows from genes to transcripts to proteins to metabolites to function—phenotype. *Blue arrows* indicate interactions regulating the respective translations (reproduced with permission from Goodacre (2005). Copyright © 2005 Springer)

transcriptome—a complete set of RNA transcripts of the genes (transcriptomics) and levels of mRNA determine the concentration of expressed proteins (proteomics). The translations of genes and mRNA are complex processes and sometimes errors are introduced during the translations or chemical modifications occur after the translations. Moreover, many proteins are not enzymes and only enzymes regulate the concentrations of metabolites. Therefore, it is necessary to conduct a comprehensive and quantitative analysis of the metabolic profile of a living organism in order to fully study its systemic responses to environmental or genetic perturbations (Fiehn 2001; Khoo and Al-Rubeai 2007). Metabolomics can thus provide a snapshot of the cellular biochemistry and physiology taking place at any instant. Metabolomics combined with system biology can finally bridge the gap between genotype and phenotype (Goodacre 2005).

In this chapter, methodologies of multivariate statistical analysis used in metabolomics and several examples of metabolomics applications will be described. These examples constitute only a small portion of enormous current and potential applications of metabolomics.

9.2   Fundamentals of Multivariate Statistical Analysis for Metabolomics

Multivariate data analysis (MVDA; Rencher 2002) has been applied to the analysis of complicated chemical and biochemical data with efficient and robust statistical methods and has provided meaningful and reliable models to analyze the metabolomic data obtained by NMR spectroscopy, mass spectrometry, Raman spectroscopy, FTIR, etc. The MVDA methods routinely used in metabolomics include principal component analysis (PCA; Wold et al. 1987), partial least squares discriminant analysis (PLS-DA; Wold et al. 1984), and orthogonal PLS-DA (OPLS-DA; Wold et al. 1998).

In this section, the following key questions will be addressed:

1. What is PCA?
2. What is data scaling and why do we need it?
3. What is preprocessing and how is it used to improve models?
4. How is PCA applied to metabolomics data analysis?
5. What do the results of PCA mean?
6. How can the results be interpreted?
7. What are Q^2 and R^2 factors?
8. How are the statistical models validated?

PCA, invented in 1901 by Karl Pearson (Pearson 1901), is the workhorse in multivariate analysis (MVA) for metabolomics and is the first step in the statistical analysis to obtain an overview of metabolomics data. The main objectives of PCA are to identify the variance in the data and transform high-dimensional data into fewer dimensions. The scatter scores plot of the first two components (PC1 and

PC2) is usually examined to see whether the data is homogeneous, any outliers exist, any grouping is formed, and what the trends of the grouping are. An outlier is typically defined as any observation that has fallen outside of the Hotelling's T2 ellipse (Hotelling 1931) at the 95% confidence interval in the scores plots of the first two components.

PCA maximizes the variance of a linear combination of the variables, based on the assumption that the largest variance in a dataset contains most of information. PCA is carried out with no prior knowledge on groupings of observations or partitioning of variables. For a data matrix X with m rows of observations (e.g., samples) and n columns of variables (e.g., NMR chemical shifts), PCA transforms the data into a new coordinate space according to:

$$X = TP' + E \tag{9.1}$$

where $\mathbf{T}$ is called the scores matrix, $\mathbf{P}$ is called the loadings matrix, $\mathbf{P}'$ is the transposed loadings matrix, and $\mathbf{E}$ is the residual matrix, which contains remaining variation that is not captured by the PCA. Scores and loadings matrices $\mathbf{T}$ and $\mathbf{P}$ can be expressed in the form of:

$$\mathbf{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1k} \\ t_{21} & t_{22} & \cdots & t_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mk} \end{pmatrix} \tag{9.2}$$

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nk} \end{pmatrix} \tag{9.3}$$

in which subscript m represents the number of observations, n represents the number of variables, and k represents the order of principal components, which is smaller than n, and typically no more than 10. For example, $k = 3$ represents the third principal component, PC3. For NMR-based metabolomics, variables are NMR spectral bins, whereas m is the number of NMR spectra. The transposed loadings matrix is defined as:

$$\mathbf{P}' = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{n1} \\ p_{12} & p_{22} & \cdots & p_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1k} & p_{1k} & \cdots & p_{nk} \end{pmatrix} \tag{9.4}$$

By substituting matrices $\mathbf{T}$ and $\mathbf{P}'$ with (9.2) and (9.4), (9.1) becomes:

$$\mathbf{X} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1k} \\ t_{21} & t_{22} & \cdots & t_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mk} \end{pmatrix} \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{n1} \\ p_{12} & p_{22} & \cdots & p_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1k} & p_{2k} & \cdots & p_{nk} \end{pmatrix} + \mathbf{E} \qquad (9.5)$$

Matrices $\mathbf{T}$ and $\mathbf{P}'$ can also be rewritten as a series of vectors, $\mathbf{t}_i$ and $\mathbf{p}_i$. Then the PCA (9.5) has the form of:

$$\mathbf{X} = (\mathbf{t}_1 \quad \mathbf{t}_2 \quad \cdots \quad \mathbf{t}_k) \begin{pmatrix} \mathbf{p}'_1 \\ \mathbf{p}'_2 \\ \vdots \\ \mathbf{p}'_k \end{pmatrix} + \mathbf{E} \qquad (9.6)$$

$$= \mathbf{t}_1 \mathbf{p}'_1 + \mathbf{t}_2 \mathbf{p}'_2 + \cdots + \mathbf{t}_k \mathbf{p}'_k + \mathbf{E} \qquad (9.7)$$

in which $\mathbf{t}_1, \mathbf{t}_2, \ldots \mathbf{t}_k$ are scores vectors of the first component, second component, $\ldots$ and k component in scores matrix $\mathbf{T}$ and $\mathbf{p}'_1, \mathbf{p}'_2, \ldots \mathbf{p}'_k$ are loadings vectors of first component, second component, $\ldots$ and k component of the transposed loadings matrix $\mathbf{P}'$. For $i = 1, 2, \ldots, k$,

$$\mathbf{t}_i = \begin{pmatrix} t_{1i} \\ t_{2i} \\ \vdots \\ t_{mi} \end{pmatrix} \qquad (9.8)$$

$$\mathbf{p}'_i = (\mathbf{p}_{1i} \quad \mathbf{p}_{2i} \quad \cdots \quad \mathbf{p}_{ni}) \qquad (9.9)$$

Note that $\mathbf{t}_i$ and $\mathbf{p}'_i$ are vectors while t_{mi} and p_{ni} are individual elements of the vectors. Equations (9.8) and (9.9) state that for each principal component i, there are m scores and n loadings. Each score is the coordinate of the observation (or sample) on the principal component i (the ith axis of the new coordinate system generated by the PCA transformation). For instance, for $k = 3$,

$$\mathbf{X} = \mathbf{E} + \begin{array}{ccc} \text{PC1} & \text{PC2} & \text{PC3} \\ \downarrow & \downarrow & \downarrow \end{array} \begin{pmatrix} t_{11} & t_{12} & t_{13} \\ t_{21} & t_{22} & t_{23} \\ \vdots & \vdots & \vdots \\ t_{m1} & t_{m2} & t_{m3} \end{pmatrix} \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{n1} \\ p_{12} & p_{22} & \cdots & p_{n2} \\ p_{13} & p_{23} & \cdots & p_{n3} \end{pmatrix} \begin{array}{l} \leftarrow \text{PC1} \\ \leftarrow \text{PC2} \\ \leftarrow \text{PC3} \end{array} \qquad (9.10)$$
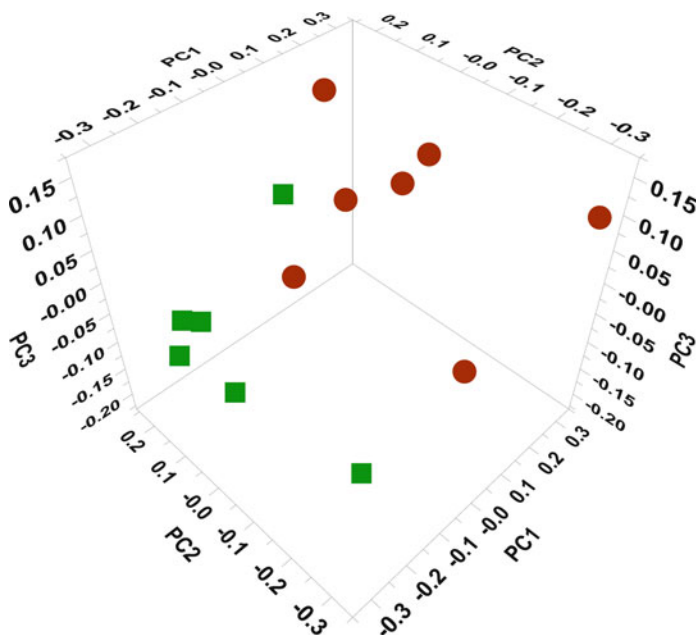
Fig. 9.3 An example of a 3D scores plot of a PCA model. Each score is a linear combination of a complete set of variables according to (9.15). Two classes of scores are represented with two different colors and shapes

Three PCs are three axes of the new coordinate system, and each observation (each sample) has three scores that define the position of the observation in the 3D PCA scores plot (Fig. 9.3). For example, sample m has scores of t_{m1}, t_{m2}, and t_{m3}, corresponding to PC1, PC2, and PC2 components, respectively.

To illustrate the meaning of the loadings matrix **P**, the matrix is rearranged by multiplying both sides of (9.5) by $\mathbf{p}_i$ from the right:

$$
X(p_1 \quad p_2 \quad \cdots \quad p_k) = (t_1 \quad t_2 \quad \ldots \quad t_k) \begin{pmatrix} p'_1 \\ p'_2 \\ \vdots \\ p'_k \end{pmatrix} (p_1 \quad p_2 \quad \cdots \quad p_k)
$$

$$
+ E(p_1 \quad p_2 \quad \cdots \quad p_k) \tag{9.11}
$$

Because loadings matrix **P** is orthonormal ($\mathbf{p}'_i \mathbf{p}_j = 0$ for $i \neq j$, $\mathbf{p}'_i \mathbf{p}_j = 1$ for $i = j$), scores vector $\mathbf{t}_i$ can be expressed mathematically as

$$
t_i = Xp_i \tag{9.12}
$$

in which for $i = 1, 2, \ldots k$,

$$p_i = \begin{pmatrix} p_{1i} \\ p_{2i} \\ \vdots \\ p_{ni} \end{pmatrix} \tag{9.13}$$

$$\begin{pmatrix} t_{1i} \\ t_{2i} \\ \vdots \\ t_{mi} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \begin{pmatrix} p_{1i} \\ p_{2i} \\ \vdots \\ p_{ni} \end{pmatrix} \tag{9.14}$$

In which subscripts 1, 2, ..., m represent observations or a total of m scores (for each principal component); subscripts 1, 2, ..., n represent n variables or n loadings (for each principal component). A total of n variables are used to describe an observation. In other words, one sample consists of n variables. For a PCA model with the first three principal components, $i = 1, 2, 3$, sample m has three scores and each score has n loadings, whereas one sample has n variables in the original data matrix $\mathbf{X}$. This means that PCA reduces n dimensions of each of m samples in the original data matrix X to three dimensions in PCA coordinate system.

From (9.14), the scores of sample m on PC_i axis can be rewritten as:

$$t_{mi} = \sum_{j=1}^{n} x_{mj} p_{ji} \tag{9.15}$$

The equation tells us that the scores of sample m, t_{mi}, are transformed from the original variables (NMR spectral bins) by the linear combinations of the variables with the corresponding loadings, p_{ji} that serve as coefficients (or contribution) of the transformation. Scores are the coordinates of an observation in the new coordinate system obtained by the PCA, whereas loadings of a score represent how the original variables in the data matrix X contribute to the scores.

Equation (9.15) also implies that a larger value of an original variable, x_{mj}, may contribute more to the model, even if its counterpart loading has a smaller value. It means that large variables weigh more variance in a PCA model. To reduce this drawback of PCA, the original dataset is usually preprocessed by scaling all variables. A scaling approach divides each variable with a scaling factor that is different for each variable:

$$\hat{x}_{ij} = \frac{x_{ij}}{f} \tag{9.16}$$

In (9.16), subscript i represents the ith observation and j represents the jth variable, x_{ij} is an original variable (the jth variable for the ith observation), $\hat{x}_{ij}$ is a scaled variable, and f is a scaling factor. The scaling is used to reduce the importance of metabolites with high concentrations in a statistical model.

There is a variety of scaling methods used for metabolomics, among which Pareto scaling is a popular one. Pareto method scales each column (variable) of the dataset by the square root of standard deviation (Eriksson et al. 2006):

$$\hat{x}_{ij} = \frac{x_{ij}}{\sqrt{s_j}} \tag{9.17}$$

in which standard deviation s_j is defined by

$$s_j = \sqrt{\frac{\sum_i^m (x_{ij} - \bar{x}_j)^2}{m-1}} \tag{9.18}$$

The mean $\bar{x}_j$ is determined by

$$\bar{x}_j = \frac{1}{m} \sum_i^m x_{ij} \tag{9.19}$$

Pareto scaling increases the relative importance of small values (low metabolite concentrations) while it maintains the original data structure relatively intact and keeps the model closer to the original measurement than unit variance (UV) scaling.

UV scales variables of the dataset by standard deviation (Jackson 1991). As a result, each variable (in each column of the data matrix) has a variance with value 1.0:

$$\hat{x}_{ij} = \frac{x_{ij}}{s_j} \tag{9.20}$$

UV scaling makes all metabolites become equally important, and hence the changes in metabolite concentration are compared based on correlations. However, UV scaling amplifies the experimental errors and noise.

Other available scaling methods include level scaling (van den Berg et al. 2006) that scales the data variables with the mean value:

$$\hat{x}_{ij} = \frac{x_{ij}}{\bar{x}_j} \tag{9.21}$$

and Poisson scaling (also known as square root mean scaling; Keenan and Kotula 2004) that scales variables with square root of the mean:

$$\hat{x}_{ij} = \frac{x_{ij}}{\sqrt{\bar{x}_j}} \tag{9.22}$$

Log transformation (Baxter 1995) is a scaling method that scales data by log function:

$$\hat{x}_{ij} = \log(x_{ij}) \tag{9.23}$$

It reduces heteroscedasticity of the dataset (random variables), but it has difficulty to scale values with large standard deviations and zeros, and inflates experimental errors.

In order to remove the offset from the dataset, variables need to be mean-centered. Centering converts all metabolite concentration (NMR peak integrations) to the changes with respect to the mean instead of to zero of the concentrations.

$$\dot{x}_{ij} = \hat{x}_{ij} - \bar{x}_j \tag{9.24}$$

Shown in Fig. 9.4 are the influences of different scaling applications on a dataset. Data are centered after scaling is applied. Log transformation, leveling scaling, and UV scaling increase the small values of the data dramatically compared to the Pareto scaling. Pareto scaling keeps the original data structure while it reduces the large values by significant amounts. Hence, for most metabolomics studies, Pareto scaling is a suitable choice to scale the data prior to multivariate statistical analysis.

After a model is obtained, two quantities are used to represent the quality of a model, which are R^2 and Q^2. R^2 describes how well the model explains the dataset—the quality of the model, whereas Q^2 represents how well the model can predict—predictability of the model (Eriksson et al. 2006). For a two-component model, a high R^2 indicates that a large portion of the information generated by a model from the dataset is meaningful and a high Q^2 means that the data predicted by the model will match the original dataset well. By definition, R^2 is the fraction of the sum of squares of all variables explained by the model:

$$R_k^2 = \frac{\sum_{m,n} (t_{mk}p_{kn})^2}{\sum_{m,n} x_{mn}^2} \tag{9.25}$$

$$R^2 = \sum_k R_k^2 \tag{9.26}$$

in which R^2_k is R^2 for the kth PC, t_{mk} and p_{kn} are scores and loadings. R^2 has a value from 0 to 1. Q^2 is the fraction defined by

$$Q_k^2 = 1 - \frac{\sum_{m,n} (x_{mn} - t_{mk}p_{kn})^2}{\sum_{m,n} x_{mn}^2} \tag{9.27}$$
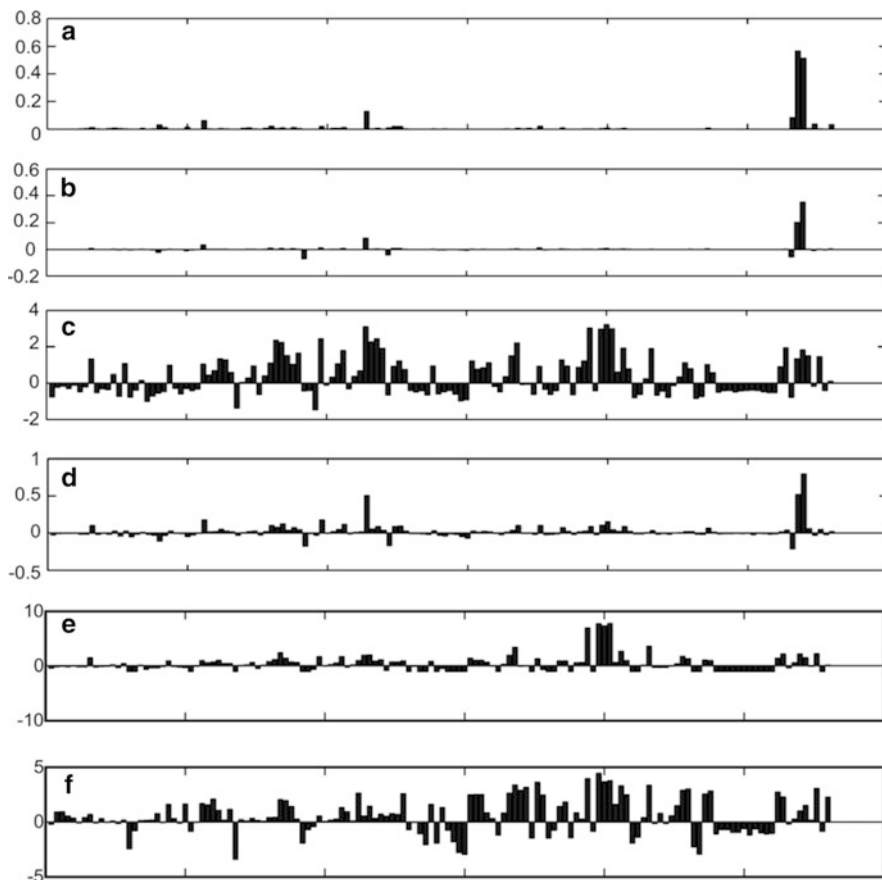
and

$$Q^2 = \sum_k Q_k^2 \tag{9.28}$$

Fig. 9.4 Effects of different scaling techniques on original data. (**a**) Original data, (**b**) after centering, (**c**) UV scaling, (**d**) Pareto scaling, (**e**) level scaling, and (**f**) log transformation (reproduced with permission from van den Berg et al. (2006). Copyright © 2006 BMC Genomics)

in which Q^2_k is Q^2 for the kth PC. A low Q^2 indicates that the PCA model mostly describes noise and does not capture the true data structure.

In summary, PCA is used to identify the variance in the dataset by transforming high-dimensional data into fewer dimensions. The scores are the coordinates of the new "coordinate space" (the model), which are obtained by the linear combination of the original variables factored with the loadings (as coefficients). For a two component PCA model, each sample has two scores (one on PC1 and the other on PC2) and each score has the same number of loadings as that of the original variables. To reduce the influences of large value variables in the model, the dataset is preprocessed with scaling and then centered by the mean. Pareto scaling is best suitable for preprocessing the data because it retains the original data structure. After a model is generated, R^2 is used to describe how well the model explains the

dataset—the quality of the model, and Q^2 represents how well the model can predict—predictability of the model. High R^2 and Q^2 represent a good model in both quality and predictability.

9.3 Sample Preparation

Sample preparation is a crucial step for any research project. This section describes detailed procedures on sample preparation, including proper sample handling of different biological samples (biofluids and tissues), extraction of tissue and biofluid samples, sample preparation for different experiments (solution NMR and high-resolution magic angle spinning [HR MAS]). Quenching and extraction of cell culture samples are also discussed.

Questions to be addressed in this section include the following:

1. How are biofluid samples collected?
2. How are the samples stored?
3. What are the procedures of NMR sample preparation for biological samples?
4. How are adherent mammalian cells cultured?
5. What are the procedures for quenching tissue samples and cultured cells?
6. How are cellular metabolites extracted from tissue and cell samples?
7. How are the lipophilic metabolites extracted from biofluids?

9.3.1 Phosphate Buffer for NMR Sample Preparation

Solution NMR samples for metabolomics are usually made of ^2H_2O phosphate buffer (PB). The PB has a final concentration of 100 mM with pH of 7.4. A standard practice is to first make a stock PB solution with higher concentration (e.g., 500 mM). A 500 mM stock PB solution at pH $= 7.4$ is formulated with 0.1140 g NaH_2PO_4 (monosodium phosphate, 0.95 mM) and 0.5751 g Na_2HPO_4 (disodium phosphate, 4.05 mM) in 10 mL 99.96% ^2H_2O (see Table 9.1). It is recommended that the stock PB is not stored in a refrigerator to avoid precipitation of the phosphates. For urine samples, the buffer solution also contains approximately 0.02% sodium azide (NaN_3, highly toxic) to prevent microbial contamination.

The ^2H_2O before use is required to be degassed to remove as much oxygen as possible, because oxygen can cause line broadening due to its paramagnetic property.

Table 9.1 Formula of phosphate buffer

pH value	Concentration (M)	Compound	Grams/10 mL	mM
7.2	0.5	NaH_2PO_4 (monosodium phosphate)	0.1680	1.40
		Na_2HPO_4 (disodium phosphate)	0.5112	3.60
7.4	0.5	NaH_2PO_4	0.1140	0.95
		Na_2HPO_4	0.5751	4.05

There are a few degas approaches. The simplest one is vacuum degas in which vacuum from a lab vacuum line is applied on a bottle of ^2H_2O for 10 min. The second common one is to place a unsealed, but capped bottle of ^2H_2O in a bath sonicator for 5 min. Caution should be taken to avoid the bottle neck contacting the water. This method is commonly used to remove excess gas (not only O_2) from mobile phase solvent for liquid chromatography (LC) or flow NMR. Thirdly, the traditional way for degassing is a flow-degas procedure that removes O_2 by blowing high-purity argon or nitrogen gas with a 0.1 μm filter into the ^2H_2O for 10 min (see Sect. 3.5.1). Among all these degas procedures, the sonication is most efficient and easiest to use, which requires minimum effort to operate.

9.3.2   Urine Samples

Urine of rodents is collected in wire-bottom cages or metabolic cages (shoebox cages) into tubes with 0.1 mL 1% sodium azide (NaN_3, highly toxic) solution to prevent metabolites in urine from bacterial degradation (Kurien et al. 2004). Note that the final concentration of the NaN_3 solution in the urine should be calculated based on the total volume of the collected urine during the preparation of NMR samples to ensure that the concentration of NaN_3 is not higher than 0.02%. Typically, the addition of NaN_3 will add approximately 10 mM Na^+ to the final NMR samples. The cages are washed after sample collection and at least once daily. There is a waiting period of 48 h before the urine collection is started in order to ease any stress that may be caused by the environment changes.

After collection of urine samples (ideally 250–500 μL), each sample is centrifuged at $3,000 \times g$ for 10 min at 4°C and the supernatants is transferred directly into a cryogenic vial and then frozen in liquid nitrogen or dry ice. Samples are stored at −80°C or colder (Bernini et al. 2011). For preparing NMR samples, a urine sample is mixed with an equal volume of 200 mM PB to obtain a final volume of 220 μL for a 3 mm tube (600 μL for a 5 mm tube, or 150 μL for flow cell), and vortexed until sample is completely mixed (approximately 30 s). The initial concentration of the NMR phosphate buffer (PB) should be 200 mM in 99.8% ^2H_2O to yield the final phosphate concentration of 100 mM.

Because human urine is more diluted than that of rodents, more urine volume may be needed. For a 220 μL NMR sample, 176 μL human urine is mixed with 44 μL 500 mM PB in ^2H_2O in a labeled 2 mL microfuge tube (human urine: PB = 4:1). To achieve better water suppression and flat spectral baseline, 0.5 mL urine sample in a labeled 2 mL microfuge tube is lyophilized (or freeze dry) overnight or dried in a vacuum concentrator for 4 h. Next, 200 μL 100 mM PB is added into the microfuge tube. The solution is vortexed for 1 min and centrifuged briefly. A stock sodium 3-(trimethylsilyl) propionate-2,2,3,3-d$_4$ (TSP) solution is added to each sample as an internal reference with a final concentration of 0.5 mM. Note that the concentration of TSP may be optimized to yield a TSP NMR signal comparable to those of metabolites in the urine sample.

The final step of the sample preparation includes transferring each sample into an individual NMR tube after centrifuging the sample at 10,000 $\times$ g for 10 min at 4°C. For running the samples in flow mode, the samples are filled into a 96-well 0.45 µm filter plate (see Sect. 9.4.2) and centrifuged at 10,000 $\times$ g for 2 min at 4°C and transferred to a 96-well plate filled with 500 µL glass inserts.

9.3.3 Blood Plasma and Serum

Blood plasma is the liquid component of blood without blood cells, which is obtained from anticoagulated blood. Blood serum is the blood plasma without the fibrinogens (coagulation factor proteins). Serum is obtained from coagulated blood (clotted blood) and has all the proteins that plasma has, except those used in blood coagulation. Therefore, serum samples can be prepared for NMR experiments in the same way as for plasma after the anticoagulation process.

Blood samples are collected in labeled Vacutainer® (BD, NJ, USA) serum separator tubes without anticoagulant (Vacutainer® SST™ or Vacutainer® SST™ II plus) for serum isolation or Multivette® (Sarstedt™) Lithium-Heparin plasma tubes for plasma preparation (Deprez et al. 2002; Bernini et al. 2011). When a Vacutainer® tube is used for collecting serum samples, the tube is gently inverted five times. The blood collected without anticoagulant is clotted at room temperature for 30–120 min. For plasma preparation, Multivette® tubes containing blood were immediately placed on a roller for up to 35 min. The Vacutainer® or Multivette® tubes are then centrifuged at 1,500–2,000 $\times$ g for 15 min at 4°C. It has been suggested that both types of tubes do not introduce any impurity to the sample which generates interfering peaks in the ^{1}H NMR spectrum. Plasma preparations using collection tubes with such anticoagulants as citrate, EDTA (e.g., Vacutainer® SST™ K2E 5.4 mg), or an unspecified clotting activator should be avoided for NMR-based metabolomics because these types of anticoagulants give intense NMR signals that interfere with the resonances of metabolites in NMR spectrum.

After the centrifugation, the plasma (supernatant) or serum is carefully transferred into a cryogenic vial and either analyzed immediately or stored at −80°C. It is recommended that the samples should not be stored at −80°C for longer than 1 year to avoid the degradation of samples. For long-term storage, the cryotubes are stored in the vapor phase of liquid nitrogen in a liquid nitrogen dewar.

To prepare a plasma sample for NMR experiments, 200 µL plasma is mixed with 100 µL 300 mM PB with 0.6 mM DSS (or TSP) by vortexing. After they are centrifuged at 10,000 $\times$ g for 10 min at 4°C, the samples are transferred into individual 3 mm NMR tubes. For running the samples in flow mode, the samples are filled into a 96-well 0.45 µm filter plate (see Sect. 9.4.2) after the above centrifugation step and centrifuged at 10,000 $\times$ g for 2 min at 4°C, then transferred to a 96-well plate filled with glass inserts.

Similar to urine samples, plasma samples can also be dried to obtain a better water suppression and better spectral baseline. Plasma samples are first

Table 9.2 Biofluid extraction protocol

Preparation

1. Label and add bearing (3.2 mm) to Eppendorf tubes (SafeLock®). Note that bearing is used to facilitate the mix of solvents, not for breaking tissue mass
2. Place tubes in cold racks
3. Add 280 μL sample (serum, plasma, or urine) to each tube
4. Return to the cold rack
5. Mix ice cold solvents: methanol (HPLC grade), chloroform (HPLC grade), and deionized H_2O (dH$_2$O) in volume ratio of 4:4:2.85

Extraction of biofluid samples

6. Add 1.2 mL of the above solvent mixture into each tube
7. Place in a vortex adapter and vortex for 20 min
8. Centrifuge at 1,000 × g for 15 min at 4 °C. The solutions separate into an upper aqueous phase (polar metabolites) and a lower chloroform phase (lipophilic metabolites) separated by protein debris

Separate and dry samples

9. Using a pipette, carefully remove upper phase leaving a small amount behind (to avoid contamination) and dispense into a labeled 2 mL cylindrical screw-top tube
10. Using a gel loading tip, remove lower phase making sure not to bring any residual upper phase along
11. Dispense into a labeled 2 mL HPLC glass vial (do not use microfuge tube since they are not compatible with chloroform)
12. Dry samples in vacuum concentrator without radiant heat. Four to six hours is usually sufficient

deproteinized by adding 200 μL ice cold methanol (HPLC grade) into 200 μL plasma. After precipitation of proteins for 5 min, the samples are centrifuged at 10,000 × g for 10 min at 4°C. The samples are dried in a vacuum concentrator for 4 h and reconstituted with 200 μL 100 mM PB with 0.2 mM TSP.

Plasma and serum (or urine) samples can also be extracted into polar and lipophilic fractions using a two-phase extraction method. For the extraction, 285 μL biofluid sample is added to an Eppendorf Safelock® 2 mL tube. The samples are extracted with 0.8 mL of an ice cold solvent mixture made with chloroform (HPLC grade) and methanol (1:1 volume ratio). The upper phase is the aqueous phase containing the water soluble metabolites, while lipophilic metabolites are in the lower chloroform phase. Table 9.2 provides a detailed protocol for the extraction. After the extraction, 250 μL CDCl$_3$/Methanol-d$_6$ (2:1 volume ratio) is added to lipophilic samples. The lipophilic samples are then vortexed for 1 min. and transferred into 3 mm tubes for NMR analysis. Aqueous samples are prepared with 220 μL 100 mM phosphate buffer solution and transferred into 3 mm tubes or 150 μL to 500-μL glass inserts on a 96-well plate for NMR analysis.

9.3.4 Tissue Sample Quench, Storage, and Extraction

For a metabolomics study, it is crucial to instantly stop the metabolic activities of biological samples and to prevent degradation of the samples. The most common step

Table 9.3 Tissue or cell extraction protocol

Preparation and weighing tissue samples

1. Label and add bearing (3.2 mm) to Eppendorf tubes (SafeLock®)
2. Weigh tubes and record (skip steps 2 and 5 for cell extraction)
3. Place tubes in cold racks
4. Add sample to each tube
5. Weigh again and record
6. Return to the cold rack

Extraction of tissue samples

7. Add 400 μL ice cold methanol (HPLC grade) and 85 μL ice cold deionized H_2O (dH_2O) to each tube with 100 mg or less (if weight is >100 mg, adjust volumes accordingly)
8. Place in tissuelyser and run for 10 min at 15 s^{-1} (vortex can be used, instead of tissuelyser in steps 8, 12, and 14)
9. Remove and centrifuge at 1,000 × g briefly (~1 min) at 4 °C to bring solvents down from cap
10. Place tubes back into cold racks
11. Add 200 μL ice cold chloroform (HPLC grade)
12. Place in tissuelyser and run for 20 min at 15 s^{-1}
13. Centrifuge briefly again
14. Add 200 μL ice cold chloroform and 200 μL ice cold dH_2O and run the tissuelyser for 10 min at 15 s^{-1}
15. Centrifuge at 1,000 × g for 15 min at 4 °C. The solutions separate into an upper methanol: water phase (polar metabolites) and a lower chloroform phase (lipophilic metabolites) separated by protein debris

Separate and dry samples

16. Using a pipette, carefully remove upper phase leaving a small amount behind (to avoid protein contamination) and dispense into a labeled 2 mL screw-top cylindrical tube
17. Using a gel loading tip, remove lower phase making sure not to bring any protein debris along
18. Dispense into a labeled 2 mL HPLC glass vial (do not use microfuge tube since they are not compatible with chloroform)
19. Dry samples in vacuum concentrator without radiant heat. Four to six hours is usually sufficient

for quenching samples is to instantly freeze the samples in liquid nitrogen. The quenched samples are usually stored in dry ice or −80°C freezer for short term and in liquid nitrogen vapor or −80°C freezer for long term. It is worth noting that for storage in a liquid nitrogen dewar, samples are stored in the vapor phase, not inside the liquid nitrogen, and cryogenic vials are used. When samples must be stored inside the liquid phase (for instance, due to limited space in the dewar), a cryotube is used to seal the vials to prevent samples from suddenly erupting when taken out of the dewar.

The quenching step starts with collecting approximately 100 mg of tissue sample, immediately placing it in a labeled 2 mL cryogenic vial and flash-freezing it in liquid nitrogen. The samples are then subjected to a dual phase methanol/ chloroform extraction for polar and lipophilic metabolites (Table 9.3; Ekman et al. 2008; Viant 2007). Each tissue sample (100 mg or less) is transferred in a labeled Eppendorf Safelock® 2 mL tube with a stainless steel bead (e.g., BioSpec 3.2 mm) in a cold rack. The bead is used for tissue disruption in the next step. The first step of the extraction is to break down the tissue samples using a tissuelyser. After adding 400 μL ice cold methanol (HPLC grade) and 85 μL ice cold deionized water

(dH$_2$O, HPLC grade) to each tube with 100 mg or less tissue (if the weight is greater than 100 mg, adjust volumes accordingly), the tubes are placed in the tissuelyser and grinded at 15 s^{-1} for 15 min. After samples are centrifuged briefly at 4°C (1,000 × g or 3,200 RPM for 1 min), 200 μL ice cold chloroform (HPLC grade) is added to each sample. The samples are again grinded at 15 s^{-1} for 20 min. After briefly centrifuging, 200 μL ice cold chloroform and 200 μL ice cold dH$_2$O are added to each sample, and grinded for last time at 15 s^{-1} for 15 min. The last step is to centrifuge the samples at 1,000 × g for 15 min at 4°C.

After the last centrifuge, each sample tube has two phase extracts. The upper phase is the aqueous phase and contains the water soluble metabolites, while lipophilic metabolites are in the lower organic phase. Proteins and other large molecules are precipitated in the solution and trapped as a thick film in the middle layer between the aqueous and organic phases. The upper aqueous phase is carefully removed using a pipette leaving a small amount behind to avoid protein contamination and dispensed into a labeled 2 mL screw-top cylindrical tube (e.g., Sorenson 12980, Sorenson BioScience), while the lower phase is removed using a gel loading tip and dispensed into a labeled 2 mL HPLC glass vial (note: microfuge tubes are not compatible with chloroform). The samples are dried at room temperature in a vacuum concentrator for 4 h. Each polar sample is reconstituted with 220 μL 100 mM PB with 0.1 mM TSP and transferred into each individual 3 mm NMR tubes after centrifuged at 10,000 × g for 10 min at 4°C. For preparing lipophilic samples, each sample is reconstituted with 250 μL methanol-d$_4$ and CDCl$_3$ mixture (volume ration 1:2) with 1.5 mM TMS.

9.3.5 Culture Adherent Cells

Cultured cell lines from human and model animals have, for a long time, been used for omics research. However, not until recently, adherent tissue cultures have been used for metabolomics research. Cell lines can be cultured either in suspension or in an adherent matrix. This section only includes culturing adherent cells. Adherent mammalian cells are available from tissue culture banks (e.g., American Type Culture Collection; http://www.atcc.org) and can be grown on the surface of culture ware (flask or dish) and microcarriers (plastic beads in the size of a few hundreds of micrometers). The plastic surface is usually treated (called tissue culture treated) to increase the binding of cell membrane proteins to the surface.

There is a variety of culture media for growing mammalian cells lines, of which Dulbecco's Modified Eagle Medium (DMEM) is the most common one. DMEM has different formulations, including low or high glucose, with or without pyruvate, glutamine, bicarbonate, or phenol red (phenolsulfonphthalein). For culturing cells, the medium is also supplemented with 100 units/mL penicillin, 100 μg/mL strepto-mycin, 10% heat-inactivated fetal bovine serum (FBS), and 25 mM HEPES

(for buffering pH). For some hard-to-grow cell lines culture medium is also supplemented with 50 μg/L mouse epidermal growth factor (EGF). Usually, the culture medium is changed twice weekly.

Because the mammalian body contains 5% CO_2, the cells are cultured under 5% CO_2 at 37°C. Note that the CO_2 percentage of a cell culture incubator should be calibrated monthly. Since CO_2 is dissoluble in aqueous solution to form bicarbonate, resulting in pH change, the medium usually contains sodium bicarbonate at a concentration of 3.7 g/L to buffer its pH. When CO_2 is not used such as for fish cell lines, culture media contain much less sodium bicarbonate (0.15 g/L). Phenol red is used as a pH indicator for the medium. It becomes yellow when pH is <6.5 and pink at pH approximately 7.4 (Freshney 2005). When cells consume glucose to produce lactate or acetate for the energetic needs, the acidification of medium turns phenol red into yellowish. Additionally, if the medium has not been changed during a period of 3 days, the color of the medium changes to pink, indicating that nitrogenous byproducts are accumulated in the medium and the cells are under oxidative stress.

To illustrate how adherent cells can be cultured, the following examples are described. Details on culturing mammalian cells can be found in several excellent books (Freshney 2010; Davis 2002). In the first example, human brain cancer (LN229) cells were grown as monolayer cultures in a T75 (75 cm^2) tissue culture-treated flask with 10 mL low glucose DMEM supplemented with 100 units/mL penicillin, 100 μg/mL streptomycin, 25 mM HEPES, and 10% FBS at 37°C in a humidified atmosphere of 5% CO_2. The medium was renewed twice weekly.

The cells can be expanded into additional flasks by stripping the cells from growth surface using 0.25% trypsin solution with 0.01% EDTA. Prior to the addition of 2 mL trypsin solution, medium was removed and the flask was washed twice with phosphate-buffered saline (PBS). After reacting for 5 min at room temperature, the enzyme was inactivated by adding 2 mL culture medium into the trypsin solution. The cell solution was transferred into a 15 mL centrifuge tube and centrifuged at 2,500 RPM for 2–3 min. After removing the medium, 5 mL fresh medium was added into the tube and the cell solution was transferred into the T75 flasks. Next, 10–12 mL medium was added into each individual flask. The cells were incubated in a humidified atmosphere of 5% CO_2.

In the second example, zebrafish liver (ZFL) cells were grown at 28°C in a humidified atmosphere. In this case, CO_2 was not used. Because of that, the concentration of sodium bicarbonate (NaH_2CO_3) is reduced to 0.15 g/L. Therefore, regular DMEM (with 3.7 g/L NaH_2CO_3) is not suitable for growing ZFL cells. The medium for ZFL cells is made of high glucose DMEM (no sodium bicarbonate) supplemented with 100 units/mL penicillin, 100 μg/mL streptomycin, 15 mM HEPES, 0.15 g/L NaH_2CO_3, 50 μg/L mouse EGF, 10 mg/mL bovine insulin, and 10% FBS. Note that high glucose is used for ZFL cell culture.

Cells can be grown on microcarriers. For instance, MCF7 human breast cancer cells are first grown in a T75 flask with DMEM in a humidified atmosphere of 5% CO_2. At approximately 80% confluence, the cells are stripped with trypsin and reseeded back to the flask with 1 g sterile microcarriers. The flask is gently shaken every 15 min during the initial 2 h. After growing for 24 h, cells are trypsinized

again and reseeded back to the flash with the microcarriers. The cells are allowed to grow for 2 days after shaken every 15 min during the initial 2 h. Note that petri dishes (or non-treated dishes) are not recommended for culturing cells on microcarriers. When petri dishes are used, cells will not attach to the surface of the dishes, which might promote cells to grow on the microcarriers. However, not all of the cells anchor on the microcarriers. The suspension cells will consequently die and produce toxic metabolites. These dead cells are difficult to be separated from the microcarriers.

9.3.6 Quench and Extract Cells

One of the advantages of metabolomics using cell cultures instead of live animals is higher throughput. Although effective extrapolation to whole animal responses is ultimately required, such an approach provides obvious advantages. For example, there is no need to house and sacrifice animals, costs are significantly lower, and cells can be grown and studied rapidly. Also, human cell lines can be employed in order to avoid cross-species extrapolations.

Despite these advantages, owing to some barriers to practicality, conventional cell culture assays for macromolecules such as DNA and proteins are not amendable for cellular metabolomics. Perhaps the most serious of these barriers are related to the conventional method for quenching cell cultures. Preferably, a quenching method would instantly inactivate all cellular metabolism, without first changing the cell environment, since metabolite concentrations are sensitive to any change of the cell environment (de Koning and van Dam 1992; Villas-Boas et al. 2007). Also, a cell quenching method should be rapid, efficient, and reproducible so as to allow an unbiased measurement of cellular metabolite concentrations and to enable direct comparison of a large number of biological samples. Unfortunately, the conventional method for quenching adherent cell cultures does not meet the above requirements.

During the conventional cell quench, first, trypsin is applied to remove cells from their growth surface (Fig. 9.5a). This inevitably changes the profile of cellular metabolites since the enzyme severely alters the physiological state of cells due to its interaction with membrane proteins (Fig. 9.5c). Furthermore, there are no nutrients and supplements available in the trypsin solution, which are present in the culture medium. After the trypsinization, numerous time-consuming steps (e.g., wash and centrifugation) are required in the conventional method before cells are finally quenched or fixed. The total time for the double-wash process is 45–60 min. During this time, a considerable portion of intracellular metabolites are secreted from cells due to their small molecular size, fast turnover rates (Villas-Boas et al. 2007), and the different osmotic strength of the applied solutions (Britten and McClure 1962; Smeaton and Elliott 1967). As a result, the conventional method requires 10^7–10^8 cells in order to obtain suitable concentrations of intracellular metabolites for NMR analysis (Beloueche-Babari et al. 2010; Lane and Fan 2007;
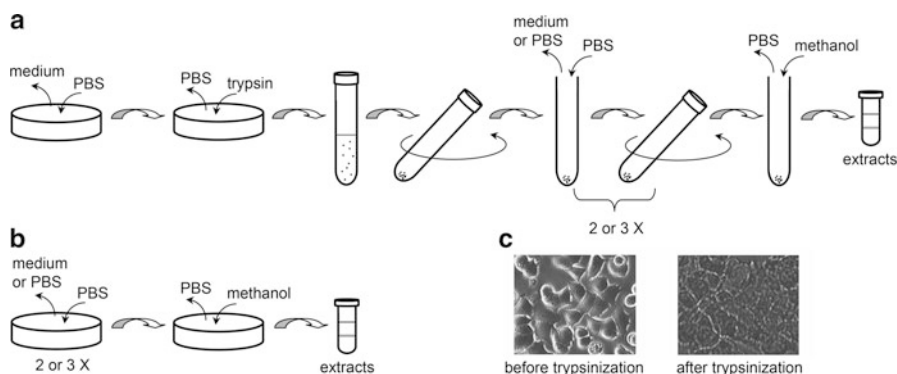
Fig. 9.5 Cell quench procedures. (**a**) In a conventional quenching method, cells are first removed from culture dish (or flask) surfaces by applying trypsin to obtain cell suspensions. Then, cells undergo a series of wash and centrifuge processes. The wash/centrifuge steps usually take approximately 45–60 min. The cells are finally quenched with methanol (or other quench solutions) and extracted. (**b**) In the direct cell quenching method, trypsin is not applied to the cells. After the culture medium is removed, cells are quickly rinsed by spraying with PBS. Residual PBS solution is removed by vacuum. The cells are then immediately quenched by methanol. The procedure takes less than 10 s. Both cells and methanol solution are collected for the cell extraction. (**c**) Images of MCF-7 cells before and after trypsinization. The images of the cells are significantly changed after trypsinization, reflecting that the trypsin severely changes the physiological state of cells due to its interaction with membrane proteins (reproduced with permission from Teng et al. (2009), Copyright © 1996 Springer)

Yang et al. 2007a, b). As a final deterrent, the conventional method is simply too time consuming to process the cell samples for metabolomics studies.

These drawbacks of the conventional method for quenching adherent cell cultures has been overcome by a direct cell quenching method that is efficient and accurate for cellular metabolomics (Teng et al. 2009). Provided in Table 9.4 is a step-by-step protocol for the direct quench (Fig. 9.6). During the direct quench, the culture medium is first removed from the 6 cm tissue culture dish (Fig. 9.5b). If extracellular metabolic profiles will also be used for the metabolomics study, 0.5 mL culture medium is collected before the medium removal. After the cells on the grown surface (culture dish surface) are flush-washed with ice cold PBS, the residual PBS is removed by vacuum. First, a 200 μL pipet tip is connected to a vacuum tubing and the vacuum line is turned on. To vacuum PBS, the dish is tilted at 45° and the residual PBS accumulated at the bottom edge of the dish is vacuumed using the tip. Cells are then quenched with 650 μL 80% methanol (HPLC grade, room temperature). In our hands, the entire quenching process (from the removal of the medium to the addition of methanol) is completed in less than 10 s, compared to approximately 45 min needed in the conventional method. Next, cells are gently detached from the culture dish using a cell lifter without crushing the cells on the dish surface. The methanol solution containing the quenched cells is then pipetted into a labeled Eppendorf safe-lock 2 mL tube in a cold rack, which contains a stainless steel bead for extraction.

Table 9.4 Cell quenching procedure for metabolomics

This procedure requires two people in order to be completed effectively and efficiently

Preparation

1. Obtain the following supplies and put them in a fume hood
    - Disposable 1 mL pipette tips
    - 1 mL pipette
    - Cell lifters
    - 100 μL pipette tips with the tips cut off by a razor
    - 80 % MeOH in a solvent dispenser
    - Waste container
    - Paper towels
    - Tube attached to lab vacuum
    - Ice bucket
    - PBS in a spray bottle
    - Paper towel rolled into a tube to support culture dish
    - Cold racks
2. Place the PBS spray bottle into the ice bucket
3. Label Eppendorf tubes (SafeLock®) with the cell line used, compound used, and sample class (e.g., "control," "low," or "high") and add a bearing (3.2 mm) to each labeled tube
4. Label microfuge tubes, if medium is to be collected
5. Place tubes into a tube rack labeled with the cell line, compound used, and date quenched

Quenching

Quencher's role

6. Note: samples should be quenched in a pseudo random order to reduce systematic error
7. Attach a cut-end 100 μL pipette tip to the vacuum hose attached to lab vaccum and turn the vacuum on
8. Take one dish from the incubator
9. If taking medium
    a. Pipette 500 μL of medium into one labeled microfuge tube
    b. Cap the tube and place it into the cold rack
10. Pour all remaining medium in the dish into the waste container and tap the dish quickly two to three times on the paper towels to remove the remaining medium
11. Holding the dish at an angle over the waste container, and spray PBS onto the dish (Fig. 9.6a)
12. Tap the dish on the paper towel to remove the remaining PBS
13. Use the pipette tip attached to the vacuum to thoroughly remove the remaining PBS from the dish (Fig. 9.6b). Vacuum only from the edges. Do not place the pipette tip directly in the middle of the dish as this will disrupt the cells
14. Dispense 650 μL of 80 % MeOH into the center of the dish using the solvent dispenser and swirl the MeOH quickly around the dish until the dish surface is fully covered by MeOH (Fig. 9.6c). Note: the quench steps (10–14) should be completed within 10 seconds
15. Angle the dish and using the flat end of the cell lifter gently scrape down the bottom of the dish to collect the cells in the pool of MeOH formed from angling the dish (Fig. 9.6d)
16. Rotate the dish 90° and continue to scrape the dish's bottom until the bottom is clear and free of cells
17. Place the dish on an angle on the paper towel tube for the sample collector to collect the cells (repeat the above steps for all samples)

Table 9.4 (continued)

Collecting quenched cells

Sample collector's role

18. Gently pick up the dish placed on the paper towel roll by the Quencher
19. Using a 1 mL pipette set on 250 µL, transfer the contents (cells and methanol) of the dish to the labeled Eppendorf tube. This step will take multiple attempts. Make sure to transfer all the cell fragments from the dish to the tube
20. Place the tube on the cold rack
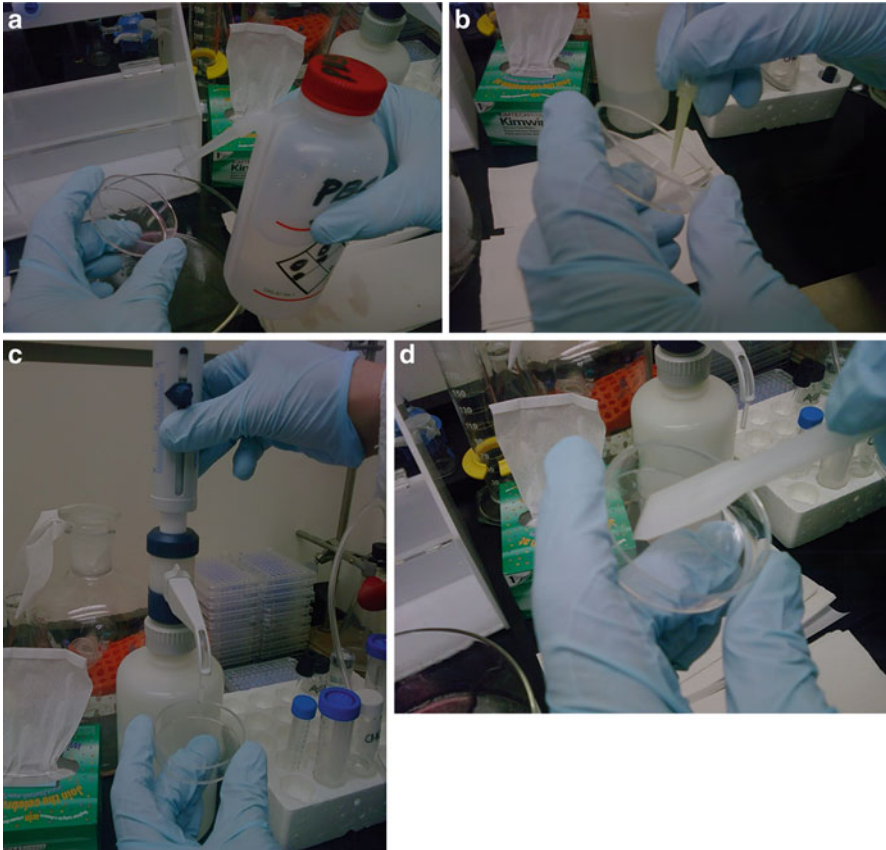21. After the entire quenching process has completed, extract the samples immediately

Fig. 9.6 Key steps of the direct cell quench method. (**a**) To flash wash the cells, hold the dish at an angle over the waste container, and spray PBS onto the dish. (**b**) Use the pipette tip attached to the vacuum to thoroughly remove the remaining PBS from the dish. (**c**) After dispensing methanol into the center of the dish using a solvent dispenser, swirl the methanol solution quickly around the dish until the dish surface is fully covered by methanol. (**d**) To collect the cells in the pool of methanol, angle the dish and gently scrape down the bottom of the dish using the flat end of the cell lifter

For an NMR measurement using a cryogenic probe, aqueous extract of 6×10^5 cells will yield a good signal-to-noise ratio with 512 scans. For a conventional NMR probe, it is recommended to grow 2×10^6 cells (in a 6 cm dish) for an NMR sample. The intracellular metabolites are extracted using a two-phase extraction procedure similar to the one described in the Sect. 9.3.4. Briefly, the cells are disrupted in tubes with stainless steel beads by a tissuelyser shaken at 15 s^{-1} for 15 min. After samples are centrifuged briefly at 4°C (1,000 $\times$ g or 3,200 RPM for 1 min), 240 μL ice cold chloroform is added into each sample. The samples are grinded again at 15 s^{-1} for 20 min. After briefly centrifuged, each sample is added with 240 μL ice cold chloroform (HPLC grade) and 220 μL ice cold dH$_2$O and grinded at 15 s^{-1} for 15 min. The last step is to centrifuge the samples at 11,000 $\times$ g (or 10,000 RPM) for 15 min at 4°C.

9.4 Practical Aspects of NMR Experiments

This section covers the practical aspects of the experiments for metabolomics studies, including the following:

1. How are the instruments calibrated and/or setup?
2. What are the routine NMR experiments for metabolomics and how are they set up?
3. What kind of information can be obtained from the individual experiment?

9.4.1 Calibration

There are a few routine 1D experiments and 2D ^{1}H/^{13}C experiments for metabolomics. Before setting the experiments, the probe should be tuned and then calibrated, including ^{1}H 90° pulse, ^{13}C decoupler 90° and decoupling pulses (for 2D ^{1}H/^{13}C experiments), and variable temperature (VT). A ^{1}H 90° pulse should be calibrated for every sample or a set of samples if samples are run in automation, while ^{13}C 90° and decoupling pulses as well as VT control are calibrated periodically. The calibration procedures are discussed in details in Chap. 4.

9.4.2 Automation

A metabolomics study usually includes a large number of samples. For a certain study, it may consist of more than 1,000 samples. It requires a large amount of time and persistence to run all the samples manually. Manual operations may also cause variations in the data. Therefore, the optimum configuration of the NMR instrument for metabolomics is to have automatic sample handler, either for tubes
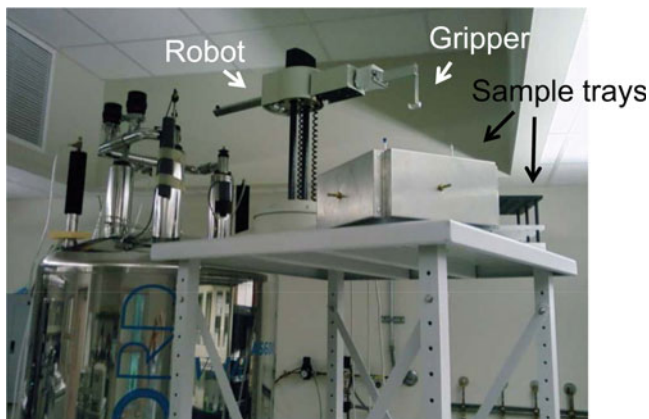
Fig. 9.7 A 100-sample robotic autosampler. The robot grips the sample holder of an NMR tube and moves it from sample trays. It returns the sample back to its original well after NMR analysis

or in flow-through mode. A sample management robot (Fig. 9.7) allows tube submission for up to 100 samples. The left rack has a home-made cooling bath to keep queuing samples at 4°C which requires an FTS precooling system. While the 100-tube robot system does a good job to run samples continuously it has several drawbacks, including that a sample has to be transferred into an NMR tube, which requires a tube washing procedure. It is time and resource consuming to wash hundreds of NMR tubes. Also, the magnetic field homogeneity (shimming) changes from sample to sample. Precaution should also be taken to see if every tube has the same sample volume. Identical sample volume before samples is transferred into NMR tubes may not always yield the identical volume in all tubes due to the sample loss during the sample transfer procedure. There are also chances that sample tubes may be broken or the robot fails to grab samples during tube ejection due to technical issues. The chance of the failures increases with the number of samples. The solution to these drawbacks is flow-through automation.

A typical flow configuration requires a flow cell in the NMR probe and a liquid sample handler to transfer samples from wells of a 96-well plate (usually HPLC glass inserts) to the NMR probe (flow cell) via polyether ether ketone (PEEK) tubing (Teng et al. 2012). The size of flow cells range from 60 to 200 μL. The smaller a flow cell volume, the more difficult the calibration. Shown in Fig. 9.8 is a Gilson® liquid sample hander (LSH) that can hold five racks. Each rack has two 96-well plates and can be cooled at 4°C. With 96-well format, tissue or cell culture samples can be extracted in much higher throughput. Biofluid samples can also be prepared in 96-well plate format. Despite the advantages of the flow NMR, there are a few challenges before flow automation can be used for NMR data collection. For example, the thin HPLC tubing for transferring sample is easy to be clogged. Inevitably, there is sample diffusion at the front end of the sample, causing dilution of sample concentration. The sample diffusion occurs inside the flow cell due to the
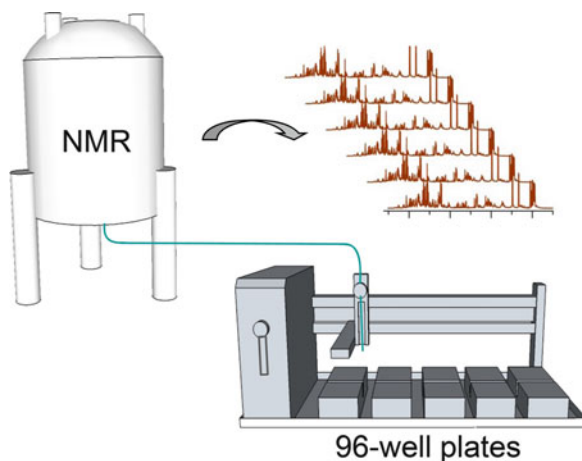
Fig. 9.8 Schematic representation of flow (or direction injection) NMR automation. The Gilson® liquid sample handler takes a sample from a specific well of a cooled 96-well plate and injects it into the probe. After NMR analysis is done, the sample is transferred back to its original vial on the 96-well plate, and the flow cell is completely washed before the next injection (reproduced with permission from Teng et al. (2012), Copyright © The Royal Society of Chemistry, 2012)

push pressure, where the contact surface is between wash solvent (^2H_2O) and sample solution, and hence dilutes sample concentration. Last, the most serious drawback to flow NMR is the well-known carryover contamination that the residual sample is retained in the flow cell during the wash process, which causes cross contamination between samples.

All of the above problems with the flow NMR have been solved recently. Sample filtration is utilized to avoid clogging the sample transfer line. The samples are passing through 0.45 μm filters in 96-well format (e.g., Millipore HTS) while spinning at $10,000 \times g$. Because metabolomics deals with small molecules, the filtration does not alter the metabolic profile observed by NMR spectroscopy. The carryover problem is overcome by a "liquid brush" and push-through procedure integrated into the wash step (Fig. 9.9). During the brush wash, the wash solvent is first pushed through, then pulled back to clean both ends of the flow cell (called delta regions) and double volume wash solvent is then pushed through. The brush step is repeated a couple of times and finally all wash solvent is pushed through the entire tubing into a waste bottle. The use of compressed air before sample injection serves two purposes: push wash solvent into waste without using ^2H_2O (meaning less cost per sample) and eliminate the diffusion since the sample will not easily diffuse into the air phase because of the thin size of the transfer tubing (inner diameter 1/16 in.).

Calibrations of the effective probe volume and volume between the injection needle to the front end of the flow cell (push solvent volume) is the first step for configuring a flow automation. The calibrations should not change as long as the tubing stays the same. The actual sample volume is about twice the size as the flow
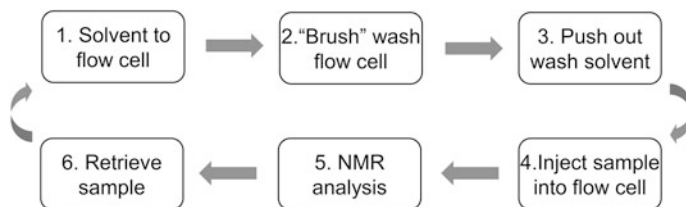
Fig. 9.9 Flow chart of push-through (PT) DI NMR automation. Before a sample in an automation queue is injected into the flow cell, it is necessary to wash the entire flow system thoroughly. The first step is to push wash solvent into the flow cell. Next, the flow cell is washed by a push–pull "brush" routine. The last step is to empty the wash solvent out of the flow line by blowing compressed air. The sample is then injected into the flow cell and the NMR data is collected. After NMR analysis is complete, the sample is retrieved and the above wash routine is repeated

cell volume to fill the space at the ends of the flow cell in order to ensure good shimming results. During the injection procedure, the sample is needed to fully fill all of the flow cell volume. Therefore, the volume for the push solvent must be properly calibrated, which equals to the volume of the tubing from the Gilson$^{®}$ sample syringe to the front end of the flow cell. To calibrate the push volume, first ^2H_2O is manually injected into the flow line to fully fill the flow cell. Then, the probe is properly tuned, the field is shimmed, and the lock level is adjusted to approximately 80%. Next, ^2H_2O with a volume 30 μL more than the volume of the flow cell (90 μL for a 60 μL flow cell) is injected into the probe pushed by H_2O using automation (0.1 μm-filtered and deionized H_2O), while monitoring the lock level change during the injection. If the lock increases to the maximum then decreases, then push volume is larger than the optimum value. If the lock level was increasing during the entire injection but it is not quite close to lock level of the manual injection, then push volume needs to be increased.

The ^2H lock level method discussed above works well for the calibration of push solvent volume. However, it will be convenient if the movement of the sample inside the flow cell can be visualized. For a cryogenic probe, the flow cell can be removed out of the probe. Using a colored sample, the calibration is very straightforward. A red food dye can be used for this purpose. The sample is prepared with a drop of red food dye in 10 mL deionized H_2O and subsequently filtered with a 0.1 μm syringe filter. The sample is injected into the flow cell by a predefined volume of push solvent (filtered H_2O) while observing the movement of the red sample inside the flow cell. A color gradient in either the outlet (front end) or the inlet end of the flow cell indicates that the volume needs to be increased or decreased, respectively. The volume of push solvent is adjusted accordingly until a uniformed color sample in the flow cell is obtained repeatedly.

The probe volume and push volume can be fine-tuned by injecting a sample consisting of 1% H_2O/99% ^2H_2O by push solvent ^2H_2O. Because the signal of H_2O is used for this calibration, the probe is tuned and the sample is shimmed again. After the ^1H 90° pulse is calibrated and gain is properly adjusted, the H_2O signal is observed

by a one-pulse experiment. Probe volume and push volume are adjusted to optimize H_2O signal. The Auto-gain setting should not be used because it changes the absolute intensity of H_2O signal and in turn affects the accuracy of the calibration. Once the probe volume (sample volume plus push solvent volume) is calibrated, it does not need be recalibrated if the length and diameter of the tubing from the injection valve to the inlet of the flow cell and the volume of the flow cell are unchanged.

9.4.3   NMR Experiments

There are a set of experiments used routinely for NMR-based metabolomics, including PRESAT, 1D NOESY, PURGE, CPMG, T_1 and T_2 measurements, COSY, TOCSY, 2D J-resolved spectroscopy, gHSQC, gHMBC, and HR MAS. PRESAT, CPMG, T_1 and T_2 measurement, COSY, TOCSY, gHSQC, and gHMBC have been discussed in previous chapters. The setup of these experiments and parameters are similar to those discussed in previous chapters with a few exceptions. Because metabolomics deals containing samples with mixtures of small molecules (endogenous metabolites), it requires high spectral resolution. Consequently, the acquisition of a 1D experiment should be sufficiently long, usually longer than 2 s (2–5 s). For a heteronuclear 2D experiment, the t_1 increments need to be as many as 1,024 to ensure a sufficient resolution required for separating ^{13}C resonances of the metabolites. Other practical considerations include that an internal standard (DSS or TSP for aqueous samples, TMS for lipophilic samples) should be used and its line width should be less than 2 Hz (usually close to 1 Hz).

9.4.3.1   PRESAT

Arguably, PRESAT is the oldest and most popular experiment for solvent suppression. It is simple to setup with fewer parameters that need to be calibrated (see Sect. 4.8.1). The key to success for this experiment is to accurately calibrate the resonance frequency of the solvent (water in most cases) to be suppressed. The saturation frequency must be set to water resonance frequency. Usually, the carrier frequency is set to the same as the saturation frequency, in which case, the center of the spectrum is at the top of the water peak. The selectivity of the saturation is achieved by a long pulse with very low pulse power. A frequently cited drawback with PRESAT is that the sequence is ineffective to suppress the hump regions of the water signal, which originated from the sample outside the probe coil. Additionally, an asymmetric water peak will inevitably deteriorate suppression efficiency of the experiment since the saturation will be only applied to a narrow bandwidth (usually less than 0.1 ppm the spectral range). Therefore, samples should be well shimmed before-hand. Another drawback with PRESAT is the unsatisfactory baseline for samples with low concentrations. All of these problems can be improved by 1D NOESY experiment.

9.4.3.2   1D NOESY

A 1D NOESY (or NOESY1D) experiment is the same as the first increment of a 2D
NOESY (Sect. 4.10.7) without an evolution time. Although NOESY is used to
obtain the distance information of spins (^{1}H, ^{13}C, ^{15}N or other 1/2 spins), 1D
NOESY is used here to achieve a better water suppression with a better spectral
baseline compared to PRESAT experiment. NOESY pulse sequence is essentially
the same as a WEFT (water eliminated Fourier transform; Patt and Sykes 1972)
sequence that suppresses water signal by taking advantage of the difference in T_1
relaxation (spin–lattice relaxation) between solute and solvent nuclei. WEFT
sequence is commonly used in paramagnetic NMR spectroscopy for fast data
acquisition to suppress signals from diamagnetic nuclei that have longer T_1 relaxa-
tion times than paramagnetic nuclei.

   Because water has a different T_1 relaxation time (usually longer) than
metabolites, 1D NOESY can be used for metabolomics applications to obtain a
better solvent suppression and flatter baseline than a PRESAT experiment. During
the predelay period, the PRESAT pulse suppresses most intensity of the water
signal. The residual water signal is further suppressed by the NOESY sequence
(Fig. 4.21). The first two 90° pulses of 1D NOESY invert all magnetizations (solute
and solvent) to the −z axis in a vector presentation. When metabolites (solutes) relax
toward the +z axis (equilibrium state) during the mixing period, the residual water
signal (mostly from the hump regions of the water peak) relaxes slower. Therefore,
the mixing time can be optimized to achieve maximal sensitivity for metabolites
and suppression for the water peak (when the residual water magnetization is near
the origin in a vector frame).

   Setup of a 1D NOESY experiment includes tuning, shimming, and calibrations
of ^{1}H 90° pulse and water resonance frequency. The data is usually collected with a
spectral window of 12 ppm, saturation delay of 2 s, mixing time 100 ms, acquisition
time of 2–3 s, saturation frequency on water resonance, 16 steady-state scans
(dummy scans), and transients of $n*32$. The saturation power should be weaker
than that in a PRESAT experiment. The drawback of 1D NOESY is the reduced
sensitivity compared to PRESAT because of the mixing time. A longer mixing time
may further improve the spectral baseline, but it will lead to further loss of the
spectral intensity. It is worth noting that the loss of spectral intensity is not equal for
the metabolites in a sample, which is mixing time-dependent, because of different
T_1 relaxation times. Therefore, if 1D NOESY is used in a metabolomics study in
which the changes in concentrations of metabolites are measured, it is necessary to
use the same mixing time for all samples, including standard samples. The NOESY
experiment works very well for samples prepared with "100%" ^{2}H$_2$O. For samples
containing relatively a large amount of water (e.g., urine samples), PURGE
sequence is one of the options.

9.4.3.3 PURGE

Presaturation utilizing relaxation gradients and echoes (PURGE; Fig. 9.10a) was introduced as an alternative approach for solvent suppression (Simpson and Brown 2005), which is as simple to use as the PRESAT but yields similar results to those by more complex sequences such as Watergate and Water-flip-back experiments. The pulse sequence consists of spin echo block sequences and gradient pulses.
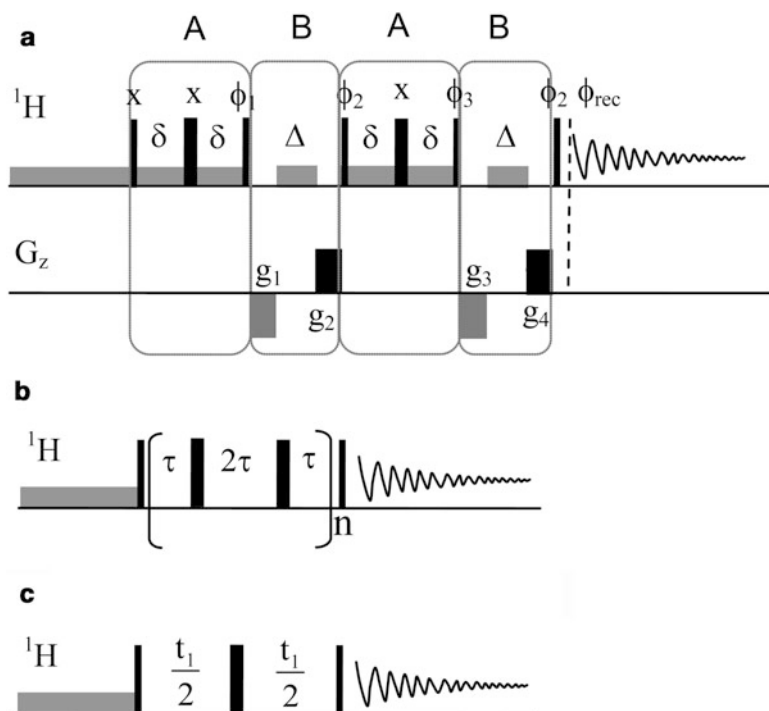

Fig. 9.10 NMR pulse sequences. ^1H represents ^1H transmitter channel, while G_z represents z gradients. For ^1H channel, narrow rectangles represent $90°$ pulses, wide rectangles are $180°$ pulses, and flat rectangles are water suppression pulses. Unless stated, ^1H pulses are applied along the x axis. (**a**) PURGE sequence (Simpson and Brown 2005). A and B represent the pulse segments. Phase $\phi_1 = x, x, -x, -x$; $\phi_2 = x, x, x, x, -x, -x, -x, -x, y, y, y, y, -y, -y, -y, -y$; $\phi_3 = x, -x, x, -x, -x, x, -x, x, y, -y, y, -y, -y, y, -y, y$; $\phi_{rec} = x, -x, -x, x, -x, x, x, -x, -y, y, y, -y, y, -y, -y, y$. Gradients G_1, G_2, G_3, and G_4 are set to -7, 30, -9, and $35 \, Gcm^{-1}$ with a duration of 1.0 ms. A typical setup of PURGE includes: calibrating ^1H $90°$ pulse length, setting both ^1H transmitter frequency and saturation frequency on water resonance, and setting saturation pulse power in the range of 40–100 Hz. Other parameters include 100 μs for δ, 200 μs for Δ, 2 s presaturation, and 2–3 s acquisition time. (**b**) CPMG sequence. The spin echo sequence is cycled n times to attain a desired T_2 relaxation delay time (or spin echo period, T). The number of cycles (n) must be an even number and is calculated according to spin echo period T. (**c**) J-resolved COSY sequence. The pulse sequence is basically a spin echo sequence, except that the spin echo delay is modulated as evolution time t_1

As its name implies, the idea of PURGE sequence is to use selective pulses during a spin echo period to tilt residual water magnetization out of a transverse plane, while the bulk magnetization of solutes is refocused by the spin echo. The 90° pulse at the end of the spin echo rotates the solute magnetization back to the z axis, whereas the residual water magnetization is brought to the transverse plane and dephased by the following z gradient pulse (G_1). Similarly, the bulk water magnetization is tilted away from the z axis by the second short saturation pulse (Δ) and again destroyed by gradient G_2. The sequence is repeated to improve the efficiency of solvent suppression.

Unlike the gradient pulses of Watergate or Water-flip-back sequence which are used to refocus solute and dephase water magnetization, the gradient pulses in PURGE are used exclusively for destroying the water magnetization that was tilted away from solute magnetization by the water selective pulse (long pulse on water resonance with the same pulse power as that of presaturation pulse). The first gradient (G_1 and G_3) after spin echo is used to dephase the magnetization of residual water, whereas the second gradient (G_2 and G_4) for dephasing bulk water magnetization. Therefore, strengths of G_1 and G_3 are much weaker than those of G_2 and G_4. Additionally, these two kinds of z gradients must be applied in different signs in order to avoid water magnetization being refocused. Each gradient pulse is optimized independently, meaning that it is not necessary to have a fixed ratio among the four gradients. Typically, the gradients G_1, G_2, G_3, and G_4 are set to -7, 30, -9, and 35 Gcm^{-1}, respectively for an aqueous sample in 90% H_2O/10% ^2H_2O. Usually, the strength of G_4 is optimized for maximum water suppression (Simpson and Brown 2005). A ^2H_2O sample only needs 20% as much as the gradient strengths used for the aqueous sample. Furthermore, the saturation periods (or saturation pulses) δ and Δ should be optimized so that the solvent magnetization is tilted about 90° out of phase to the solute magnetization.

A typical setup of PURGE includes: calibrate ^1H 90° pulse length, set both ^1H transmitter frequency and saturation frequency on water resonance, set saturation pulse power in the range of 40–100 Hz (see Sect. 4.8.1). Other parameters include 100 μs for δ, 200 μs for Δ, 2 s presaturation, and 2–3 s acquisition time. To achieve the best results of suppression, saturation pulses Δ need to be optimized first and then pulses δ.

9.4.3.4 CPMG

For plasma or serum samples, the broad signals from proteins and lipids distort spectral baseline and partially bury the signals of small metabolites. Carr–Purcell–Meiboom–Gill experiment (CPMG) makes use of the difference in T_2 relaxation between small metabolites and proteins to suppress the signals of proteins. Shown in Fig. 9.10b is a CPMG pulse sequence. The spin echo sequence is cycled n times to attain a desired T_2 relaxation delay time (or spin echo period T), where n must be an even number. During the spin echo period (T), the chemical shift is refocused and the refocused magnetization has decayed by T_2 relaxation. At the end of T period, magnetizations of all ^1H spins are scaled by the relaxation factor e^{-T/T_2}. Therefore,
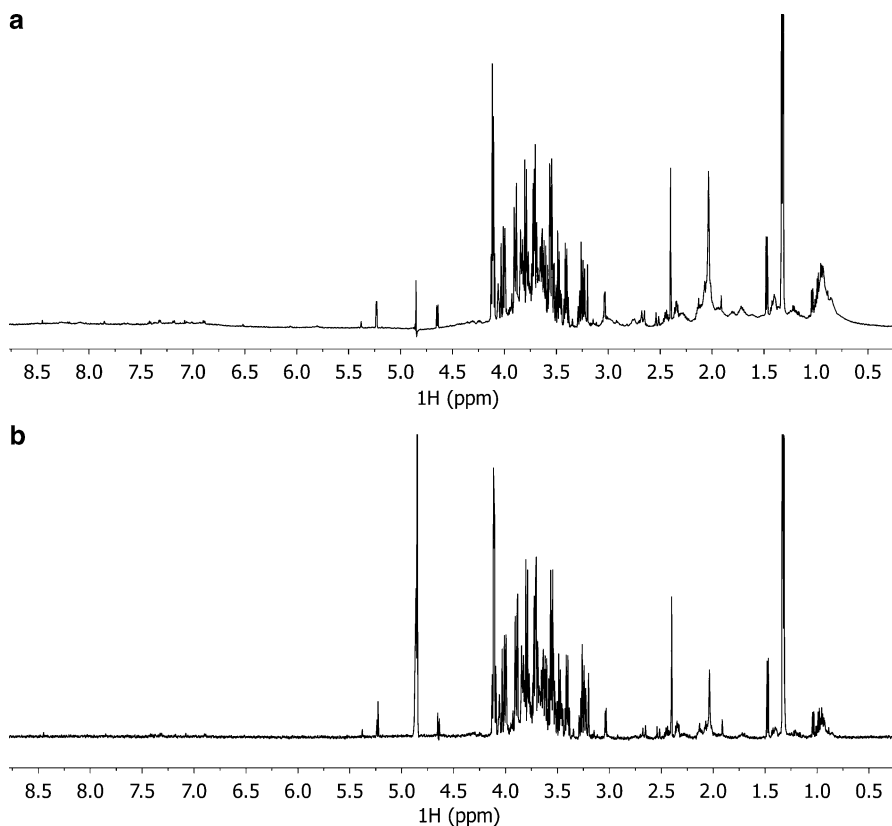
a

b

Fig. 9.11 NMR spectra of a serum (heat-inactivated fetal bovine serum). (**a**) 1D NOESY spectrum with a mixing time of 100 ms. The severely distorted baseline was caused by the broad peaks from proteins and lipids in the spectrum. (**b**) CPMG spectrum of the same sample. The data was acquired with 100 ms T_2 relaxation delay and the same number transients as in (**a**). The line shapes of peaks from small molecules and spectral baseline have significantly been improved

for the ^{1}H spins of proteins with T_2 relaxations close to the T value, the signal intensities are significantly suppressed, whereas the signals from small molecules are detected since their T_2 relaxations are much longer. Fig. 9.11 shows ^{1}H spectra of a serum sample acquired by 1D NOESY and CPMG. With a T value of 100 ms, CPMG (Fig. 9.11b) successfully suppressed the protein signals that severely buried the signals from small molecules in the PRESAT spectrum (Fig. 9.11a).

A typical setup for CPMG includes calibration of 90° and 180° ^{1}H pulses. Delay T is usually set to 100 ms or between 50 ms and 200 ms. Note that delay T is used to calculate the even number of spin echo cycles for the CPMG experiment. Delay τ is set to 0.5 ms, which can also be optimized for different samples (usually up to 1.0 ms). The 180° pulse may be fine-tuned during the setup to optimize the quality of CPMG spectrum. Parameters for presaturation are the same as in PRESAT.

9.4.3.5 J-Resolved Spectroscopy

In a J-resolved (JRES) NMR spectrum, the homonuclear spin–spin coupling is separated from ^{1}H chemical shifts and heteronuclear couplings. The [1]H–[1]H couplings appear on F1 dimension after Fourier transformations. The pulse sequence for JRES NMR (Fig. 9.10c) is basically a spin echo sequence, except that the spin echo delay is modulated as evolution time. Because of the spin echo (T_2 editing), the projection of JRES spectrum has a flat baseline. The experiment is mainly used to resolve the overlapped spin couplings. The projection of 2D JRES spectrum (45° tilted) has been used for metabolomics to improve spectral resolution (Viant 2003). The increase in the spectral resolution allows to segment the spectrum in reduced binning size (0.005 ppm compared to 0.01 ppm). It results in the reduced scatter of samples in PCA scores plots. A drawback of JRES is the loss of spectral sensitivity compared to a 1D ^{1}H experiment (e.g., PRESAT).

Setup of a JRES experiment is straightforward, including tuning, shimming, calibrations of 90° and 180° pulses, and finding water resonance. Because it is used as a 1D spectrum (1D projection), the acquisition time is set to 2 s or longer for suitable spectral resolution. The spectral width (SW_1) of F1 is set to 40 Hz while SW is the same as in PRESAT (12 ppm). Other parameters includes saturation delay of 2 s, dummy scans of 16, and t_1 increment same as 1/(2*SW) and 32 increments. The number of transients in the setup is based on the total experimental time, which should be at least double to that of a PRESAT experiment.

The data is processed with unshifted squared-sine-bell or sine-bell functions for both dimensions, linear predicted to 64 points and then zerofilled to 128 points on F1 and zero-filled to 64 k complex points on F2 before Fourier transformations. Shifted squared-sine-bell or sine-bell functions may be applied to both dimensions to increase S/N ratio of the JRES projection spectrum at the price of resolution. The processed 2D spectrum is tilted by 45° and projected along the F2 dimension. Note that JRES spectrum is displayed in magnitude mode. Therefore, spectral phasing is not needed.

9.4.3.6 High-Resolution Magic Angle Spinning

In a solution, rapid molecular tumbling averages out the anisotropic nuclear interactions such as chemical shift anisotropy, dipolar and quadrupolar couplings (see Chap. 1). In intact tissue specimens or cells, however, some of the interactions may not be averaged to zero because of slow or restricted molecular motions. The metabolites in the constrained environments have short transverse relaxation (T_2) times, resulting in reduced spectral resolution (broadened NMR line widths) and detection sensitivity (low intensity). While the *J*-coupling and CSA is negligible and ^{1}H does not have quadrupolar coupling, the ^{1}H homonuclear dipolar coupling and susceptibility anisotropy are significant in the intact tissue or cell samples, which shorten the T_2 relaxation time and cause the line broadening. Both dipolar coupling and susceptibility anisotropy have an angular dependence on the well-known ($3\cos^2\theta - 1$) term, meaning that the interaction can be averaged by

spinning the sample at the magical angle ($\theta_m = 54.74°$ or $3\cos^2\theta_m - 1 = 0$). Magical angle spinning (MAS) technique (Moestue et al. 2011) has been used to increase the spectral resolution of the intact tissue and cell samples.

In a tissue or cell sample, metabolites experience different motions. For instance, small metabolites in the cell cytosol move free and rapidly, while those that bind to macromolecules or cell membrane are restricted in mobility. Lipid molecules in the liquid crystal phase of the cell membrane have restricted diffusion motions but rapid axial rotation. For the motion-restricted metabolites or lipids, $^1H-^1H$ dipolar coupling and susceptibility anisotropy are not canceled by the molecular motion, but averaged in some extent. These reduced couplings can be averaged by slow MAS or HR MAS. Note that HR MAS can only be utilized to study the semisolid samples in which $^1H-^1H$ dipolar coupling and susceptibility anisotropy are partially averaged by rapid axial molecular motions. Large macromolecules with slow axial motion are not the subject of HR MAS. Solid-state MAS NMR is required to study those large molecules and the $^1H-^1H$ dipolar coupling can only be suppressed partially even at a spinning rate of 50 kHz or above.

For sample preparation, 5–8 mg intact tissue or 10^6–10^7 cells are placed in the bottom of a rotor. After 30 μL ice cold 100 mM PB/^2H_2O with 0.1 mM TSP is added, the cap is gently screwed in to force out excess PB. The tissue should be small enough to allow self-adjustment of balance during the initial spinning. It is important to note that air bubbles should be completely removed from the sample rotor since the presence of the air bubbles will broaden the peak line width. This can be done by gently screwing in the rotor cap with a slightly excessive amount of the buffer solution.

The general calibration of the probe is done using a liquid sample (e.g., 2 mM sucrose in a glass rotor). The calibration includes ^1H transmitter pulses and ^1H and ^{13}C decoupling pulses. The spin rate is calibrated using an oscilloscope. The first step of the experiment setup is to spin the sample. The pressure of bearing air is first increased to achieve a spin rate of approximately 300 Hz. Next, drive air flow is increased and regulated by the operation software (or manually) to the desired spin rate (2–5 kHz). The fast spin is needed to ensure that all spin sidebands of the peaks lie outside the spectral window (5 kHz for 8.5 ppm on a 600 MHz ^1H frequency). However, the fast spin may cause sample degradation. Since spectral peaks in the aromatic region (downfield) usually have intensities much lower than those in the upfield region, the sidebands of aromatic resonances may not interfere with the statistical analysis of the data. Therefore, spin rate of 2–3 kHz is commonly used for HR MAS when the data in aromatic region is not used for data analysis (spin sidebands of aliphatic resonances overlap with aromatic resonances for MAS of 3 kHz or less).

After the spin rate is regulated, basic setup of an NMR experiment for HR MAS is similar to solution NMR with the following exceptions. For a CPMG experiment, it is crucial to set the τ delays to be synchronized with the MAS spin rate. For instance, the delays are set to 500 μs when the sample is spun at 2 kHz (τ delay in second = 1/spin rate in Hz). Additionally, the sample is cooled at 4°C to prevent sample degradation. Because the sample spinner is surrounded by the bearing air, the probe temperature is regulated using refrigerated bearing air. In practice, in order to achieve 4°C sample temperature, the temperature of bearing air needs to be
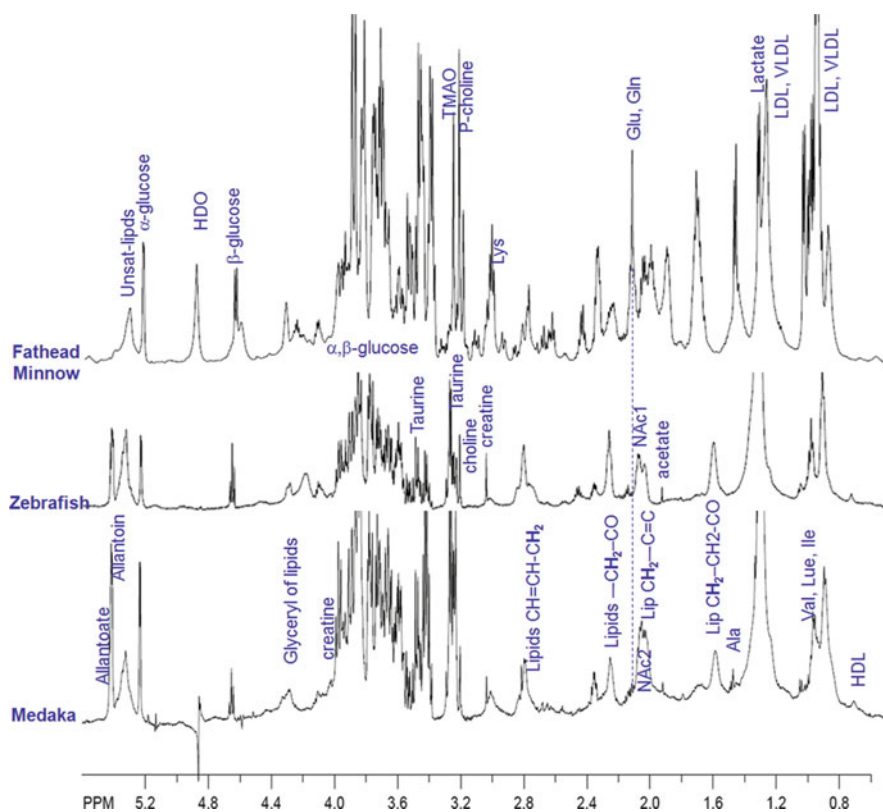
Fig. 9.12 HR MAS spectra of intact fish liver tissues using CPMG sequence. The *top* spectrum is from fathead minnow, *middle* from zebrafish, and *bottom* from medaka. Some identified metabolites are labeled above the spectra. The data were collected with approximately 5 mg tissue spun at 2 kHz

maintained at -25 to $-35°$C. Shown in Fig. 9.12 are representative MAS CPMG spectra obtained from fish liver tissues using the parameters noted above. Some of the metabolites are labeled on the spectra. Different than the tissue extracts, an HR MAS spectrum has NMR peaks from both polar and lipophilic metabolites.

9.5   Dada Analysis and Model Interpretation

In Sect. 9.2, the fundamentals of PCA have been discussed, including scaling functions and quality factors. In the current section, practical aspects of applying multivariate methods to analyze metabolomic data will be covered. Key questions to be addressed including the following:

1. What is a scores plot and how can it be used?
2. What is a loadings plot and how can it be interpreted?

3. What is PLS-DA?
4. How can the models be validated?
5. What is the difference between PLS-DA and OPLS-DA models?
6. What are the advantages and drawbacks of the methods?
7. How is the information obtained by the methods different?
8. How is an OPLS-DA model validated?

Among the results generated by multivariate approaches, four quantities are often analyzed first, which are scores and loadings plots, R^2 and Q^2. As described in Sect. 9.2, scores are the coordinates of the new dimension-reduced coordinate space obtained by the analyses using PCA, PLS-DA, or OPLS-DA, loadings are the contributions of the original variables (NMR spectral bins) to the new coordinates. R^2 indicates how well the model explains the dataset and Q^2 describes what the predictability of the model is. This section mainly discusses how the experimental data is processed; how the data can be utilized for metabolic profiling. Several routinely used modeling methods of MVA will be discussed such as principle component analysis (PCA), partial least squares discriminant analysis (or projection to latent structures discriminant analysis, PLS-DA), and orthogonal PLS-DA (OPLS-DA). Utilization of several software packages will also be discussed in detail, including Mnova (Mestrelab Research, Santiago de Compostela, Spain, www.mestrelab.com), SIMCA-P$^+$ (Umetrics, www.umetrics. com; see Table 9.5 for an example protocol), PLS Toolbox (Eigenvector Research, Wenatchee, WA, USA, www.eigenvector.com; see Table 9.6 for an example protocol), and MetaboAnalyst (www.metaboanalyst.ca; see Table 9.7 for an example protocol).

PCA and PLS-DA are the most commonly used multivariate approaches for data analysis in metabolomics. OPLS-DA is a modification of PLS-DA method by integrating orthogonal signal correction (OSC) into a PLS model. The basic idea of OPLS analysis is to separate the systematic variation in the original dataset (X) into two parts, one that is linearly related to the class information (Y variables, class labels) and the other that is unrelated (orthogonal) to Y. The components that are related to Y are called predictive, whereas the components that are unrelated to Y are called orthogonal. For an OPLS-DA model, there is one predictive component related to Y variables, and several orthogonal components with no or very little relation to Y meaning that only the first component is predictive. Usually, only one orthogonal component is calculated because for every addition of an orthogonal component, the percent of X explained is reduced.

9.5.1   NMR Dada Processing and Normalization

Before the NMR data can be utilized for the multivariate analyses, it is necessary to process the spectra and export binned spectra into a text file or an Excel® file. An example protocol for the preparation of an NMR data table using Mnova is

Table 9.5 Multivariate analysis protocol using SIMCA-P+ (Umetrics)

Load a data file

1. Open a data file
   a. Select File → New (or "New File" icon), after starting SIMCA-P+
   b. Go to the folder containing the excel file, select the file, Open, OK
   c. Select "SIMCA-P project," then "Next"
2. Transpose the data arrangement (click on Commands → Transpose)
3. Set primary Variable IDs and primary observation IDs
   a. Click on the first cell
   b. Click on primary under Variable IDs
   c. Click primary for observation IDs, NEXT, Finish

Assign Class IDs and Scale data

4. Open data worksheet (click Workset → Edit → 1)
5. Set Observation IDs
   a. Click Observations
   b. Select Primary ID for Class from Obs ID. Click Set
   c. Change Length to highlight the name for sample class
   d. Click OK
6. Scale variables
   a. Click Scale
   b. Select all variables (Ctrl+a)
   c. Select Par for Scaling type
   d. Click Set, OK

Build a PCA Model

7. Select PCA for modeling: click Analysis → Change Model Type → PCA on X-block
8. Start calculation: click on "Autofit" icon or "Calculate the first two components"
9. Generate a scores plot by clicking "Scores Scatter Plot" icon
10. Set color for classes by clicking on Color, Coloring type: by identifiers
11. Type a number for Length that only shows classes. Click OK
12. Generate a loadings scatter plot by clicking Loadings Scatter Plot icon. p1 for *X*-Axis, p2 for Series
13. Set label types, var ID (primary) for Point Label, none for Axis Label
14. Set "Save As Default Options," OK
15. Generate a loadings line plot: click Analysis → Loadings → Line plot
16. Use Num for *X*-axis, p1 for series
17. Label type: Var ID (Primary) for Axis Label, 1 for Start, 4 for Length, 90 for rotation, 100 for Display label interval. Click OK
18. Save the project: click Save icon

Build a PLS-DA or an OPLS-DA model

19. To do PLS-DA, select Analysis → Change Model Type → PLS Discriminant
20. Do steps 8–18
21. For **OPLS-DA**, Edit Excel data spreadsheet to add a row for classes at beginning of the Excel file, e.g., 111111222222333333
22. Do steps 1–2
23. Set Y variable: click the arrow of the first column, then select *Y*-Variable
24. Select the second cell of first row, then click primary under Variable IDs, and primary for observation IDs
25. Then, click Next, Yes to all for Var-1 (for variable ID), finish

(continued)

Table 9.5 (continued)

26. Answer No to All for exclusion of 0 values. Note that the 0 values are needed for loadings line plot
27. Do steps 4–6
28. Answer No to All for exclusion of 0 values. Click OK
29. Set OPLS-DA for modeling: select Analysis → Change model type → OPLS
30. Calculate first two components
31. Do step 9–1. for generating scores and loadings plots

Edit loading line plot

32. Right-click on the plot, select Plot Settings → Axis...
33. Click X axis, General, Set Sacle minimum: 950. maximum: 1,810. which gives a range of 4.75–0.49 ppm, Apply
34. X axis Title: ppm, Apply
35. Click Y axis, General, set Scale range
36. Title: PC1 (or PLS1. etc.), Apply
37. Axis aspect ratio: type in a ratio
38. Click OK
39. To save a pdf, Ctrl+p, set plot size, OK
40. Right-click on the plot, select Plot Settings → Header and Footer...
41. Right-click on the plot, select Plot Settings → Plot area...

Create Loadings plot with 3. most significant loadings

42. Create a loadings column plot: Analysis → Loadings → Column plot
43. Use p1 for Series, Label types → Axis label: Var ID; start 1. length 4. display interval 1
44. Click OK
45. Sort Ascending: right-click on the plot, select Sorting Ascending, check Sort by values in series: 1
46. Create a list: right-click on the loadings plot, Create → List
47. Save the list: right-click on the list, Save list as. Drag the save txt file into excel. Save the excel file
48. Delete all rows, except 1. rows in the beginning and the end of the file
49. Save the excel file as "top 3. loadings"
50. Select all cells in excel, Data → Sorting the smallest to largest. Note: make sure that the paired cells are sorted together (ppm and the loadings)
51. Right-click the loadings plot → Properties → Item selection
52. Select all Var ID on the selected field (right) and move them to the unselected field (left)
53. Select the Var Id according to the "top 3. loadings" list. Click OK

provided in Table 9.8. The data are first apodized with 0.3 Hz line-broadening and a multiplication of first points by 0.5. The data are then zero-filled to 64 k complex points before Fourier transformation. It is important that the spectra are properly phase- and baseline-corrected because metabolomics deals with small changes in metabolite concentrations (NMR spectral intensities). The next step involves chemical shift calibration and signal alignment to ensure that variables of all samples represent the same physical identities, meaning that the resonance (or resonances) of a metabolite has the same chemical shift values across all samples. Resonance shift occurs across samples partly due to the slight change in pH or concentration of samples. Shown in Fig. 9.13 are representative superimposed NMR spectra

Table 9.6   Multivariate analysis protocol using solo (or PLS_Toolbox; Eigenvector Research)

Load a data file

1. Open a data file
   a. Click on the Decompose (PCA) icon to open the analysis workspace, after starting Solo
   b. To import the desired .txt (Tab delimited) file, click: File → Import Data → X-block → XY…Delimited Text Files (TXT, XY), then select the desired normalized data file (must be in .txt format)
2. To make sure the data is sorted correctly for labeling, click: Edit → Calibration → X-block Data
   a. When the dialog box opens, click Row Labels. Each row should have a different label that corresponds to each sample. For each class of data (control, dose, etc.), manually change the class so that all samples within a class are matched and all different classes are differentiated (ex: Class 1, Class 2, etc.). Class 0 denotes an unknown, so do not use this class unless there is an unknown involved (certain analysis tools are unavailable for Class 0)
   b. All samples that you wish to include in the analysis (generally every sample) should have a green check in the box next to the row
   c. Click Column Labels. This section corresponds to ppm values. Right-click on the box labeled Axis Scale → Copy. Then right-click on the box that says Label → Paste. Under the label column, you should now see ppm for Name and the ppm values next to each column number. It should look the same as the Axis Scale column
   d. Again, all samples that you wish to include in the analysis (generally every sample) should have a green check in the box next to the row. Do not worry about class for Column Labels
   e. Close the dialog box. Everything has saved automatically
3. Preprocessing: On the right hand side of the screen, click 2. Preprocessing
   a. Highlight Autoscale, then click Remove to remove it from the Selected Methods
   b. From the available methods, first highlight Parato Scale, and click Add to move it to the Selected Methods
   c. Then highlight Mean Center and do the same
   d. In Selected Methods, Mean Center should be listed *above* Parato Scale (this is important). Click OK to close the dialog box

Build PCA model

4. On the right hand side of the screen, click 4. Build Model
5. When a series of values shows up on the screen, highlight the second or third row (PC2 or PC3). The % Variance Cumulative should be >50.0
6. When this is highlighted, again click 4. Build Model

View cores and loadings plots

7. When this is highlighted, again click 4. Build Model
8. Then click 6. Review Scores (the second row should still be highlighted). A graph (**SCORES PLOT**) and a small Plot Controls box will open
   a. For X:, highlight Scores on PC1 from the drop box
   b. For Y:, highlight Scores on PC2 from the drop box
   c. At the bottom, check the box next to Conf. Limits (95 %)
   d. On the top of the box, click View → Labels → Set 1
   e. The graph should now be in four quadrants with a circle around most of the data and labels on each point. Each class of data should be a different color
   f. Without closing the graph or the dialog box, click back onto the analysis workspace
9. With the second row still highlighted, click 7. Review Loadings. Another graph (**LOADINGS PLOT**) and small Plot Controls box will open
   a. For X:, highlight Loadings on PC1 from the drop box
   b. For Y:, highlight Loadings on PC2 from the drop box
   c. On the top of the box, click View → Labels → ppm

Table 9.6 (continued)

  d. The graph should now be in four quadrants with the majority of the data densely packed towards the center and labels on each point

      To view both graphs together, at the top of the box, click FigBrowser → Auto-dock User Figures. Then on the top right of the graph, select the image of a box with a single line down the middle. The graphs should now be side-by-side, and you can toggle the Plot Control boxes by clicking on the gray area of each graph

10. For developing **CONFIDENCE ELLIPSES** (Tolerance Ellipses) around the Scores Plot, click on the gray area of the Scores Plot so that its Plot Controls window is displayed

  a. Click View → Classes → Outline Class Groups → Confidence Ellipse

  b. Each class of data should now be surrounded by a corresponding confidence ellipse (solid lines). The previous overall confidence limit circle will still be present as a dashed line, unless the box is unchecked in the Plot Controls window

  c. The default confidence limit is 0.95. In order to adjust this, click View → Classes → Outline Class Groups → Set Confidence Limit. Enter the desired confidence limit and click OK

  d. Note that a confidence ellipse will not be constructed for Class 0 data, which are unknowns (this is the reason we generally do not use Class 0)

11. For developing line graphs for the **LOADINGS OF INDIVIDUAL VARIABLES** on the Loadings Plot, click on the gray area of the Loadings Plot so that its Plot Controls window is displayed

  a. Click Data in the middle section

  b. Then highlight the point you would like to analyze by creating a box around it

  c. A line graph will then pop up, with Sample in the X-range and the loading values for the selected variables in the Y-range

  d. Click View → Labels → Set 1. Each point should now be labeled according to which sample it correlates to

12. For developing **PPM BAR CHARTS FROM LOADINGS**, click on the gray area of the Loadings Plot so that its Plot Controls window is displayed

  a. Click PlotGUI → Duplicate Figure. The same graph will pop up in a new window

  b. In Plot Controls, for X:, select ppm from the dropdown box

  c. In Plot Controls, for Y:, select either Loadings on PC1 or Loadings on PC2, depending on which graph you wish to analyze

  d. There should now be a line graph with ppm on the x-axis and Loadings on the y-axis

  e. In order to create a bar graph, click View → Settings… Under Plot Settings, there should be a box that says plot type. Select bar from the dropdown box. Click OK

obtained from 21 intracellular extracts. Several peaks are shifted in the data (Fig. 9.13a) and aligned well after spectral alignment process (Fig. 9.13b).

Resonance region of water or any other solvent is excluded from the spectra before spectral binning. The binning is a process in which a spectrum is segmented into a number of regions (called bins) by integrating spectral peak area over small ppm ranges (bin size), usually 0.005 ppm or smaller. Integration of a binned segment can be summed to the high limit of the bin or the center point of the bin, either of which produces the same results, provided all data of the dataset are treated in the same way. Binning is used to reduce the effect of minor peak shift and decrease the number of variables used for the statistical analysis. For a spectrum with 10 ppm spectral width and 32 k data points, spectral binning with 0.005 ppm bin size reduces the original 32 k variables to 2,000 bins.

Table 9.7 Multivariate analysis protocol using web-based MetaboAnalyst (http://www.metaboanalyst.ca)

Upload a data file

1. Go to www.metaboanalyst.ca
2. Click on "click here to start"
3. Make sure your dataset is in the proper format (.csv format)
   a. Open the Excel normalized file you wish to analyze
   b. Where it says ppm in the Excel file, write Sample
   c. Add a row beneath the sample labels. In the first Column, write Label. Then under each sample label, put the sample's class. Make all control samples 1, low dose 2, and high dose 3, etc
   d. Save the file as .csv
4. Upload your data
   a. Under Data type, select spectral bins
   b. Under Format, select Samples in columns (unpaired)
   c. Click Browse, and locate your .csv data file
   d. Click Submit

Preprocessing

5. When the Data Integrity Check screen comes up, click Skip if you are sure your data is in the correct format
6. When Data Filtering comes up click none and then "Submit"
7. Normalize your data
   i. For Row-wise normalization, select Normalization by sum
   ii. For Column-wise normalization, select Pareto Scaling
   iii. Click Process
   iv. Click Next when the normalization figure is shown

Multivariate analysis

8. Select an analysis method
   a. To create a PCA scores plot
      i. Under Multivariate Analysis, click Principle Component Analysis (PCA)
      ii. Click on the tab that says 2D scores Plot
      iii. A scores plot should be displayed with PC1 plotted against PC2 with 95 % confidence ellipses around the classes
   b. To create a PLSDA scores plot
      i. On the left hand menu select PLSDA (directly under the PCA button)
      ii. Click on the tab that says 2D scores Plot
      iii. A scores plot should be displayed with Comp #1 and Comp #2 with 95 % confidence region

Download results

9. In the far left menu click the "Download" button. Click "Download.zip" and save the results to the appropriate folder

---

Metabolomics uses multivariate analyses (PCA, PLS-DA, etc.) to extract biological information from the changes in metabolite concentrations relevant to biological functions. In order to reduce the errors in the multivariate analyses due to variations in sample concentrations or variable sample dilutions, spectral normalization is applied to the binned spectra. The spectral normalization may not be an ideal method for the data normalization, but it reduces contribution from concentration

Table 9.8 Protocol for spectral alignment and binning using Mnova (Mestrelab)

Load NMR data files

1. Start Mnova
2. Drag all data files one by one in sequential order into Mnova window to load fid files
3. Double check the file order
4. Select all files by clicking on the first data on the pages menu and Ctrl+a. Make sure that all files are selected during following steps (steps 5–12)

Process NMR data

5. Setup apodization: click Processing → Apodization and set 0.3 Hz for line broadening and 0.5 for first point. Click OK
6. Correct spectral phases
   a. Click phase correction icon and click on Biggest
   b. Hold left mouse button (LMB) and move it up or down to adjust 0-order phase (PH0) and RMB to adjust first-order phase (PH1). PH0 should be between 40 and 50° and PH1 ≈ 0. Or type in a value for PH0 if it is known
   c. Close the phase window
7. Correct spectral baseline
   a. Click on the arrow next to baseline correction icon
   b. Select Baseline Correction
   c. Select Bernstein Polynomial Fit, and 3 for polynomial order. Click OK

Superimpose all spectra

8. Superimpose all spectra: Select all spectra, then Tools → Superimpose spectra
9. Right-click on the superimposed spectra and left-click the superimposed spectra on the pages window to select the spectra

Calibrate chemical shift

10. Zoom the DSS peak, then click "fit to height"
11. Calibrate 0.00 ppm for DSS peak
    a. Click on TMS icon
    b. Click the cursor on the top of the DSS peak
    c. Type 0 for New Shift, select Auto Tuning, and press enter key

Save the superimposed spectra

12. Save the processed data: click File → Save As and type the sample name, select mnova for file type, change to the data file folder and click Save. Note; the file size should be >10 MB
13. Delete all spectra except the superimposed one and save as new_file.mnova

Align spectra

14. Zoom a region
15. Align peaks
    a. Select "Add region graphically" and click and hold LMB to select a region
    b. Repeat the above step for all peaks
    c. Preview the peak alignment by clicking Preview. Click again to go back to "Align Spectra" window
    d. Accept the alignment (click OK)

Remove water peak and bin the spectra

16. Exclude water resonance: click View → Cuts → Manual Cuts. Enter ppm range (4.7–5.0). Click OK
17. Bin the spectra: click Processing → Binning, uncheck Full Spectrum and type 0.5 (From), press Tab key, type 10.0 (To), press Tab key, type 0.005 for width of each integral region. Click OK

Table 9.8 (continued)

Normalize and save the spectra

18. Set the total area of individual spectrum to unity: click Processing → Normalize, select total area and value 1.0. Click OK

19. To save the binned spectra into a text file, click File → Save As, select ASCII Text file (*.txt) for file type, use sample name for text file name and click Save. Note that the file size should be ~522 KB

20. Close the data file **without** saving changes. Note Saving changes will overwrite the raw data

Edit the txt file

21. Run excel, and drag the txt file to excel to open the txt file

22. Insert first row and type in sample IDs. Column A should be labeled as ppm. Save the file as xlsx with a new filename

23. Add (or replace with) zeros for the rows containing water

24. Sort the ppm value from largest to smallest: click on the first ppm value (0.5), and click home → sort and filter → largest to smallest

25. Check sum of columns. They should be 1.0. Scroll to the last cell of column B and click on the end blank cell

26. Click Σ and Enter key

27. Save the file as txt file

Process J-resolved COSY data

1. Drag a JRES file into Mnova. The data will be processed automatically and 45° tilted

2. Zero-fill the spectral size to 16 k

3. Rephase the spectrum if necessary. Click on F2

4. Click on Phase Correction and adjust zero-order and first-order phase according to step 6

5. Click F1 and do the phase correction in the same way as step 3

6. Right-click on the spectrum and select Properties

7. Click on 2D Spectrum, and select horizontal Trace and click OK

8. Right-click on the projection and then select Setup

9. Use sum for projection type and click Extract icon. The projection will be extracted to a 1D spectrum

changes unrelated to the biological functions. The last step in the generation of the data table is to export binned and normalized NMR spectra into a text file (or an Excel® file).

9.5.2 Analysis of Metabolomic Data

Once the data table is generated, MVA can be utilized to analyze the data. Routinely used analysis methods include PCA, PLS-DA, and OPLS-DA. After the data table is loaded into an analysis software package (such as SIMCA-P, PLS-Toolbox, and web-based MetaboAnalyst), the dataset is preprocessed with scaling, followed by mean-centering. PCA is first applied to provide an overview on the dataset.

As an unsupervised pattern recognition method, PCA is the first analysis to be conducted for a metabolomics study. Herein, a study on the responses of human
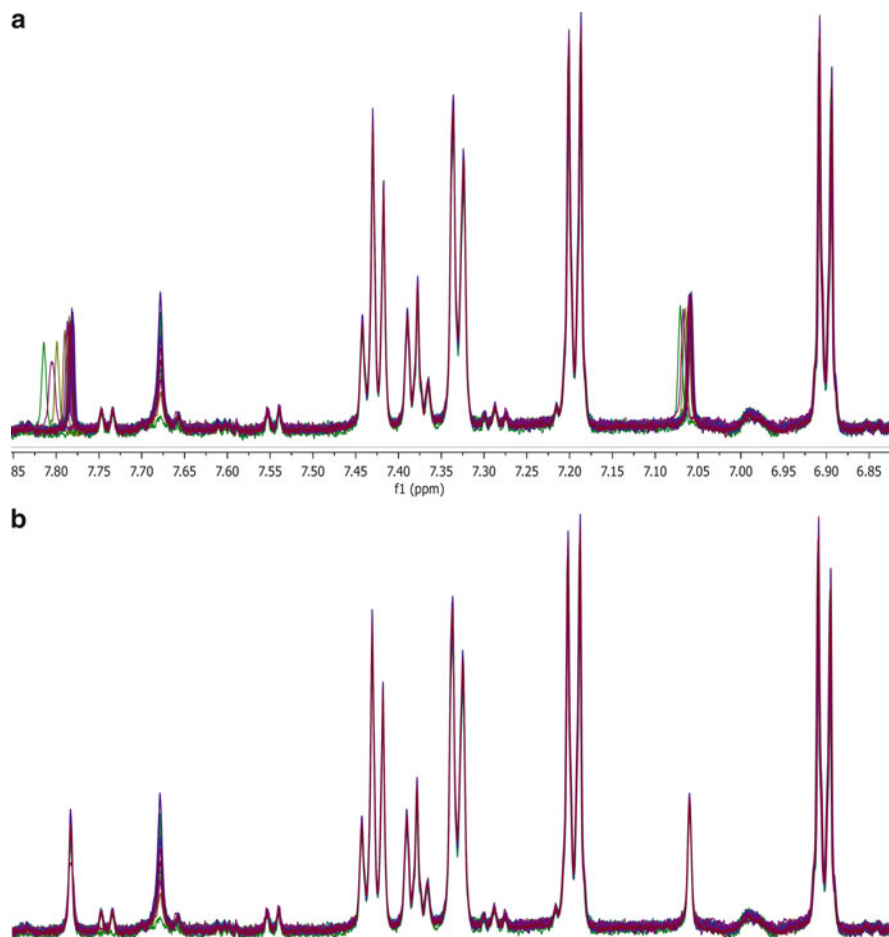
Fig. 9.13 Superimposed spectra of 21cell extracts. (**a**) Peak shifts in the superimposed spectra are clearly visible, which can be problematic during multivariate analyses. (**b**) The spectral alignment was able to correct the peak shift problem (the peaks were aligned with Mnova spectral align routine)

brain cancer cells (cell line LN229) to the treatment of anticancer drug paclitaxel (or Taxol™; PTX) is used as an example to explain the MVA process. Paclitaxel is a mitotic inhibitor used in cancer chemotherapy. In the study, the cells were treated with 1 nM (low dose) and 10 nM (high dose) PTX for 24 h. After 0.4 mL of the culture medium was taken, the treated and untreated (control, treated only with solvent vehicle) cells (7 dishes per class) were quenched and extracted, according to the methods described in Sect. 9.3. The NMR data were processed with spectral alignment, resonance exclusion (spectral region cut), binning, and normalization. Shown in Fig. 9.14 are the superposed spectra before and after resonance exclusion. The intense peaks from water, methanol, and dimethylamine (DMA) were removed before binning.
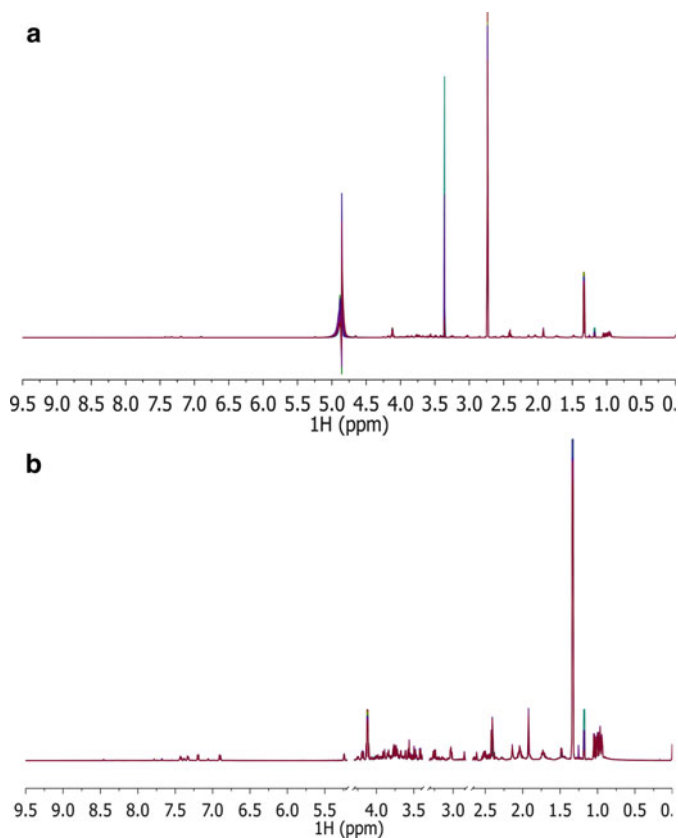
Fig. 9.14 Normalized superimposed spectra of 21cell medium samples from LN229 cells treated with PTX for 24 h as discussed in the text. (**a**) Before and (**b**) after the intense peaks from water, methanol, and dimethylamine (DMA) were removed prior to binning. The removal of the intense peaks clearly improved relative intensities of metabolite resonances

After the data file was loaded into SIMCA-P, (Table 9.5) preprocessed with Pareto scaling, and mean-centering, the PCA model was calculated. Shown in Fig. 9.15 is the summary of R^2 and Q^2 of the first three components of the models for cell culture medium and intracellular polar extracts. Both models have acceptable R^2 and Q^2 values. Next, a scatter two-PC scores plot was generated for the PCA models. In both models (cell culture medium and intracellular extracts), there are clearly class separations between control and high-dose sample classes, and low dose samples spread in between the two classes (Fig. 9.16a, b). The scores can also be visualized by plotting single-component scores vs. sample numbers (Fig. 9.16c, d). For the model of medium samples, most control samples have positive PC1 scores and the high-dose samples have negative scores with a few exceptions. For cellular extracts, however, the segregation is mostly in PC2 as shown in both 2D and 1D scores plots. This means that the variation among sample
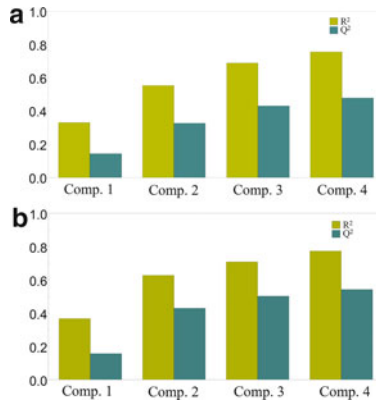
**Fig. 9.15** Overview plots highlight the quality and predictability of two PCA models derived from human brain cancer cells (LN229) treated with different dose levels of paclitaxel (PTX). R^2 describes the quality of the model and how well the model explains the dataset, while Q^2 measures predictability of the model, that is how well the model can predict classifications of observations (samples). The models are generated with NMR data collected (**a**) for intracellular extracts and (**b**) for culture media. The PCA model for culture media has a better quality (higher R^2) and predictability (high Q^2)
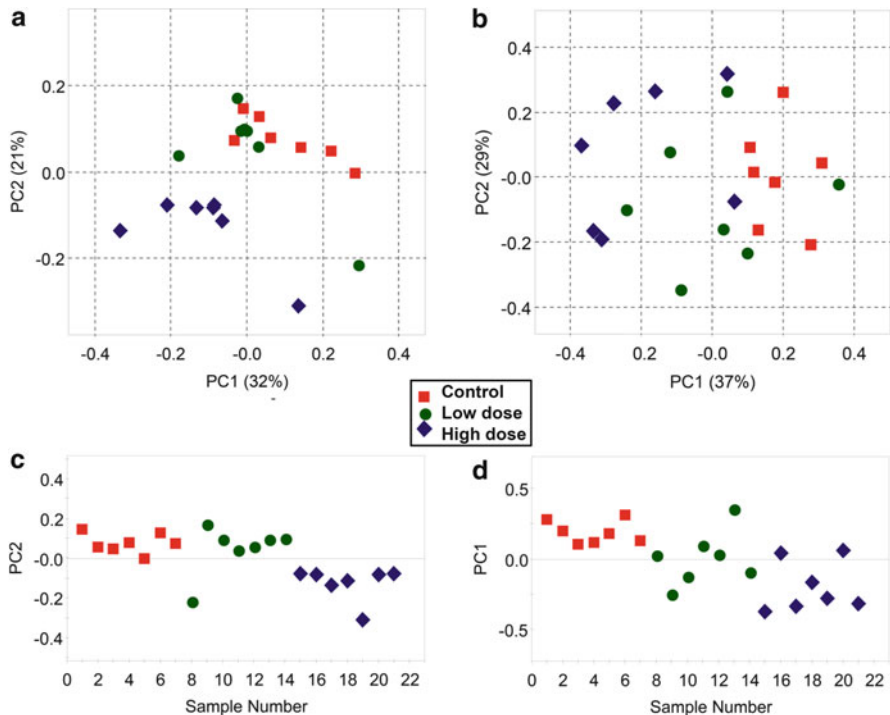
Fig. 9.16 Scores plots of two PCA models for human brain cancer cells (LN229) treated with different dose levels of paclitaxel (PTX). One- and two-component PCA scores plots are generated for intracellular extracts (**a, c**) and culture media (**b, d**), respectively. Both models clearly show class separations between control and high-dose (10 nM) sample classes, and scores of the low dose (1 nM) samples spread in between the two classes
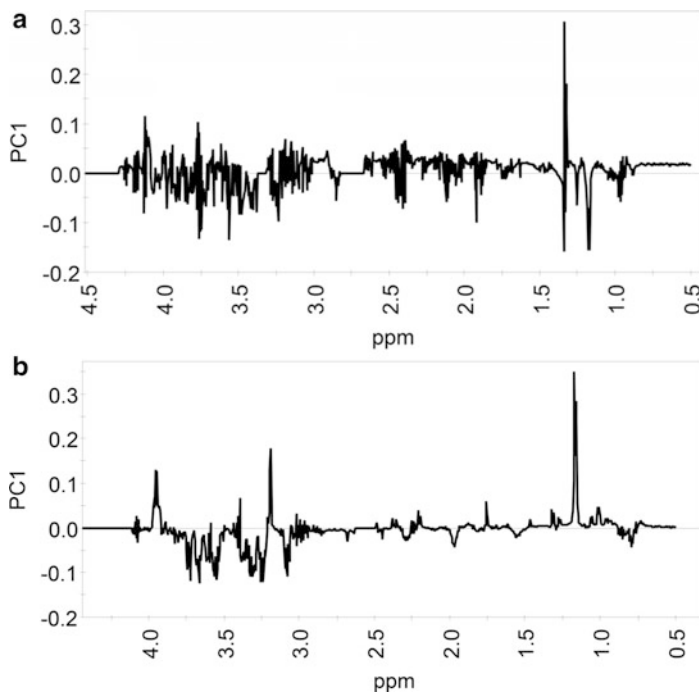
Fig. 9.17 Loadings plots for the two PCA models (summarized in Fig. 9.15). PC2 is used for the loadings plot of cell extracts in (**a**), while PC1 is used for that of culture media in (**b**), because the class separation is primarily along the respective component (see Fig. 9.16a, b). For both plots, the positive loadings indicate that the NMR spectral intensities are higher in controls than high-dose samples, or vice versa

classes of medium samples is larger than those of intracellular extracts since PC1 captures the most variation in the dataset.

Now, the question in one's mind is what metabolites contribute to the class discrimination or which variables vary most between the control and dosed classes. Loadings plots are used to address these questions. Shown in Fig. 9.17 are the 1D loadings plots of the two models. For cellular extracts, the positive PC2 loadings indicate that the NMR spectral intensities are higher in controls than high-dose samples, or vice versa. For medium samples, the positive PC1 loadings represent that the NMR spectral intensities are higher in control samples since most controls are located in the positive PC1 region of the scores plots (Fig. 9.16). It is important to note that, generally, a loadings plot cannot be used to determine the metabolite concentration changes among three classes. However, since in this case, the low-dose treated class is located in the middle region between controls and high-dose treated samples in the PCA scores plots and near the origin of PC1 or PC2, the changes in the loadings plots primarily reflect those between control and high-dose groups.

In the next steps, supervised multivariate analyses (PLS-DA and OPLS-DA) are used to maximize the class separations by taking into account the class information.
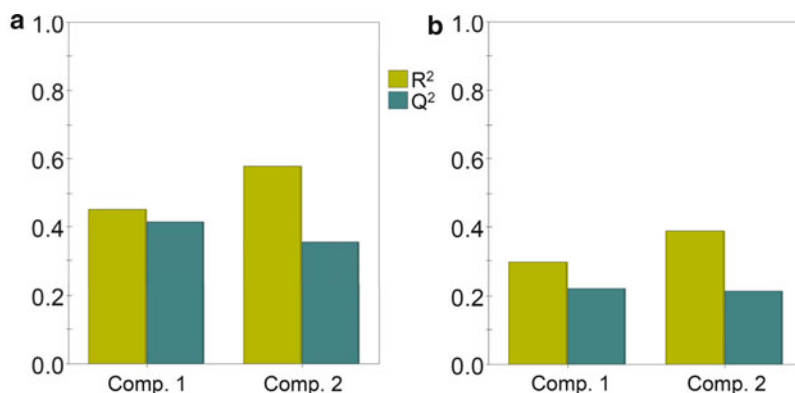
Fig. 9.18 Overview of the quality and predictability of PLS-DA models built using the ^{1}H NMR spectra obtained for human brain cancer cells (LN229) treated with different dose levels of PTX. R^2 and Q^2 measure the quality and predictability of a PLS-DA model, respectively. The PLS-DA models are generated (**a**) for intracellular extracts and (**b**) for culture media. The PLS-DA model for intracellular extracts has a better quality and predictability as indicated by much higher values of R^2 and Q^2. Lower R^2 and Q^2 of the second component than the first one in both models indicate that the first component captures the most class discrimination

In order for a supervised model to be statistically meaningful, the model must be validated and Q^2 values are important indicators in model validation. Using the same set of data preprocessing parameters, PLS-DA models are generated for both datasets (Fig. 9.18). Lower values of R^2 and Q^2 of culture medium model than those of cell extract model indicate that the PLS-DA model for medium is weaker than that for cellular extracts. This is understandable because the culture medium contains nutrients some of which are also cellular metabolites. The high contractions of the nutrients produce large background noise for metabolomic analysis, which is contained in the original dataset. In order for a supervised model to be statistically significant, the model must be validated by means of cross validation and Q^2 values are important indicators in model validation. During a validation, a fraction (usually 10–15%) of the class labels (*Y* variables) are randomly permuted among classes and the model is rebuilt using the permuted class labels. The process is repeated for a few 100 times. The original model is a valid one, if all rebuilt models have Q^2 less than the original model and a majority of rebuilt models have R^2 less than the original model.

The results of validation for the PLS-DA models are shown in Fig. 9.19, which include fitted regression lines for R^2 (triangles) and Q^2 (squares) values of the rebuilt models. Empirically, for a good model, the intercepts of the regression lines should be less than 0.4 for R^2 and less than 0.05 for Q^2 (Eriksson et al. 2006), meaning that rebuilt models with permuted class labels are much worse than the original one. Therefore, the validation results indicate that both models are validated PLS-DA, although the one for medium has smaller R^2 and Q^2 compared to those for cell extracts.

Because the class discrimination in the PLS-DA scores plots (Fig. 9.20) are primarily along PLS1, the 1D PLS1 loadings plots (Fig. 9.21) are used to explore
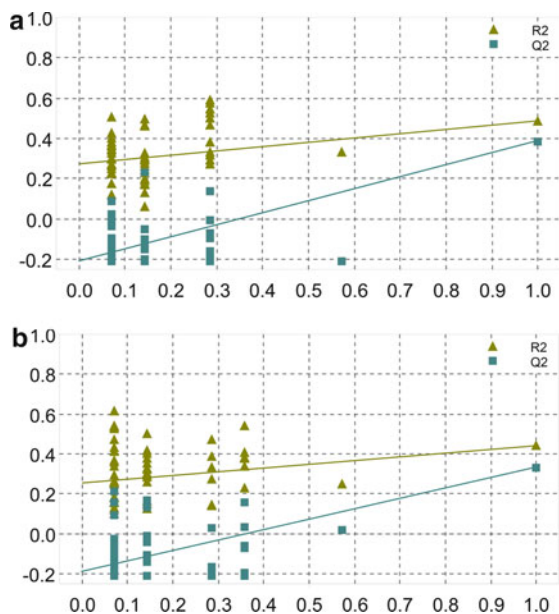
Fig. 9.19 Validation plots of the PLS-DA models (summarized in Fig. 9.18) for (**a**) intracellular extracts and (**b**) culture media. The scale of the y axis represents values of R^2 or Q^2. The x axis reflects the extent of the permutations with the scale of 1.0 representing the case that no class label is permuted and 0.0 meaning that all class labels are permuted. Both models were validated with 50 permutations. The results indicated that either model is valid because none of the rebuilt models had Q^2 values higher than the original model and a few rebuilt models had higher R^2 values with smaller Q^2, which indicated that these rebuilt models were over-fitted
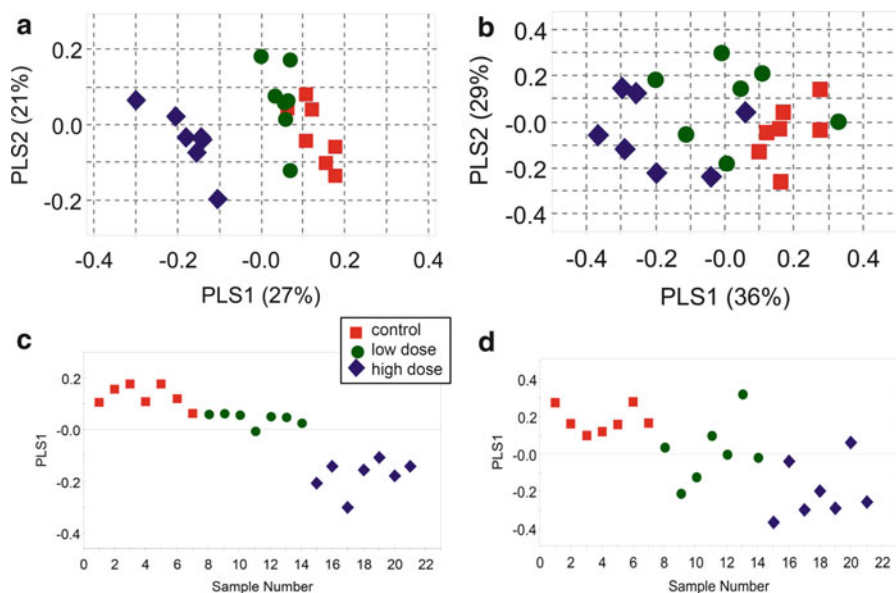
Fig. 9.20 Scores plots of two PLS-DA models for human brain cancer cells (LN229) treated with different doses of PTX. Two- and one-component PLS-DA scores plots are generated for intracellular extracts (**a**, **c**) and culture media (**b**, **d**), respectively. Both models clearly show improvement in class separations compared to the PCA models (Fig. 9.16)
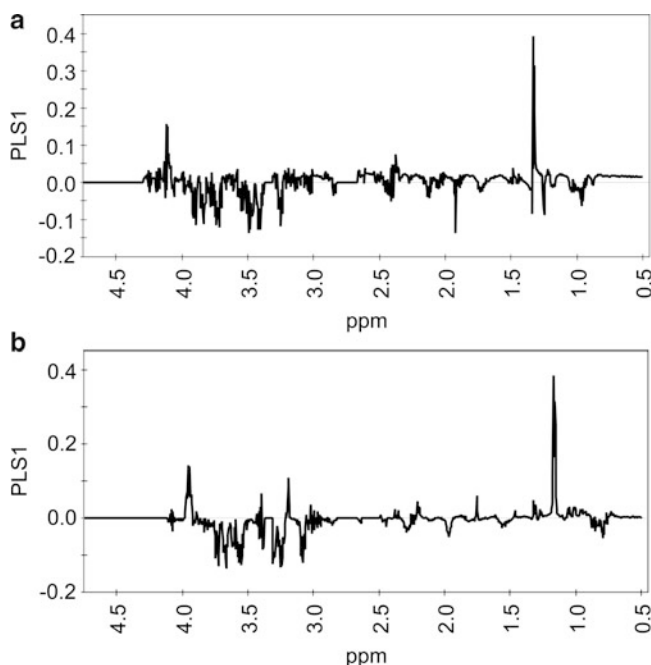
Fig. 9.21 Loadings plots for the two PLS-DA models (summarized in Fig. 9.18) for (**a**) intracellular extracts and (**b**) culture media. The loadings plots show better baseline, less noise, and better peak resolution, which clearly demonstrates the improvement of PLS-DA over PCA. For both plots, the positive PLS1 loadings represent that the NMR spectral intensities are higher in control samples or vice versa, since all scores of controls are located in the positive PLS1region (Fig. 9.20)

the identification of cellular responses in terms of the concentration changes induced by the treatment of PTX. Compared to those of PCA models, the loadings plots of PLS-DA have a better baseline, less noise, and better peak resolution, which clearly demonstrates the improvement of PLS-DA over PCA (provided a PLS-DA model is validated). Some of the identified metabolite resonances are labeled on the loadings plots.

To further refine the model and identify the chemical shifts (X variables) of metabolites related to the class separation (Y variables), OPLS-DA is utilized. OPLS-DA maximizes the percentage of X variables (NMR bins) explained by the first component. R^2 and Q^2 values for the two OPLS-DA models (cell culture medium and cell extracts) are summarized in Fig. 9.22. Both R^2 and Q^2 have been significantly improved in OPLS-DA models compared to those in PLS-DA models. It is important to note that any outlier should be removed before OPLS-DA is conducted. The scores plot of the OPLS-DA model shows improved class separation for the cell extracts, compared to the PLS-DA model and the separations among the three classes are along the first component of the OPLS-DA model (Fig. 9.23). The loadings plots generated from the models are shown in Fig. 9.24,
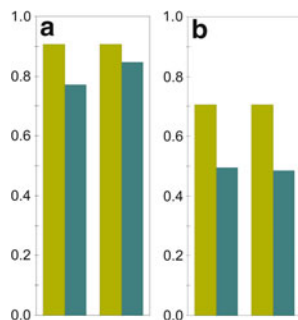
Fig. 9.22 Overview plots of two OPLS-DA models built using ^{1}H NMR data derived from human brain cancer cells (LN229) treated with different concentrations of PTX. R^2 and Q^2 measure the quality and predictability of an OPLS-DA model, respectively. The relatively high values of R^2 and Q^2 of (**a**) intracellular extracts compared to (**b**) the model for culture media indicated that the OPLS-DA model of intracellular extracts had much better quality and predictability than the model for media, provided the model is a valid one. Nonetheless, both models were notably improved over PCA and PLS-DA models
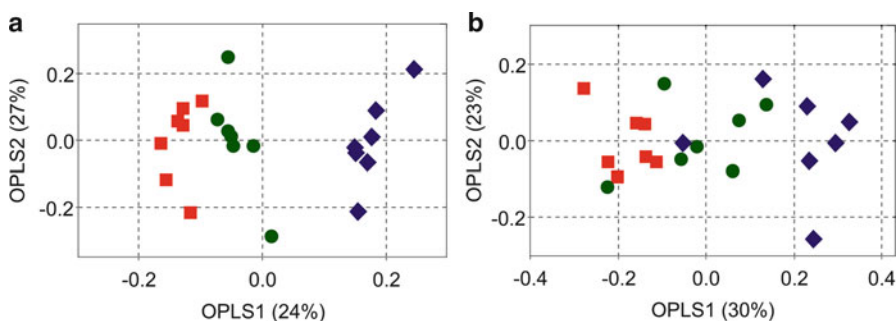

Fig. 9.23 Two-component scores plots of OPLS-DA models for human brain cancer cells (LN229) treated with different dose levels of PTX. (**a**) The separations among the classes of cell extracts are mainly along OPLS1 and the variations within a class spread along OPLS2. (**b**) The class discrimination between control and high-dose groups is largely on OPLS1, except for one outlier of the high-dose group, while the scores of the low dose group distribute between the control and high-dose groups

which are almost identical to those obtained by PLS-DA for cell extracts and medium (Fig. 9.21). OPLS-DA models are validated using the following sevenfold validation procedure slightly different than the procedure used for PLS-DA. First, the dataset is randomly partitioned into 7 subsets of samples, each of which contains 1 sample per class. Then, 6 subsets of samples (approximately 86% of the original data) are used as a training set (6 per class) to build an OPS-DA model and the remaining subset is used as a test set to validate the model. The model is generated and validated seven times (sevenfold) with a different test set (each of the 7 subset is used once as the test set). The results (scores and loadings) from the 7 models are then averaged to obtain

Fig. 9.24 Loadings plots of
OPLS-DA models for (**a**) cell
extracts and (**b**) culture media
of human brain cancer cells
(LN229) treated with
different dose levels of PTX.
For both plots, the negative
OPLS1 loadings represent
that the NMR spectral
intensities are higher in
control samples or vice versa,
since all scores of controls are
located in the negative
OPLS1 region
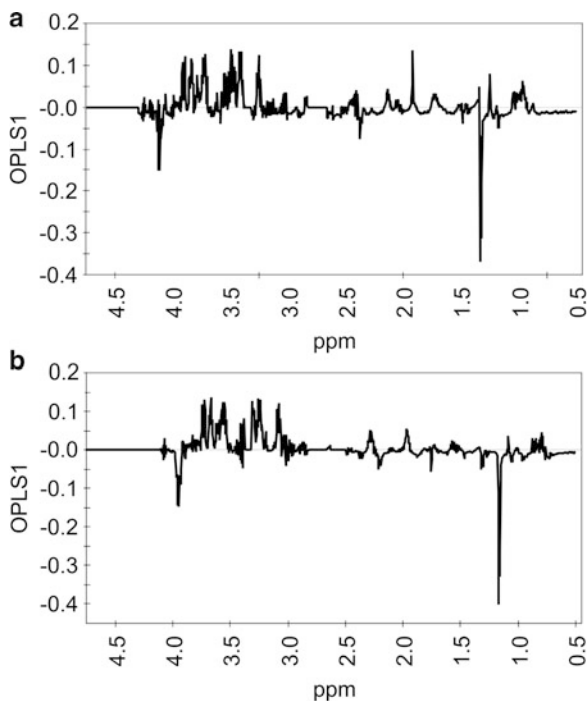
the final model. The models obtained with the training set have correctly predicted
the remaining sample in all classes during the validation (Fig. 9.25). All test samples
are located in the OPLS1 range of their own sample classes.

The next step of the model interpretation is to assign the NMR resonances that
contribute to class separation in the models. Because samples contain a mixture of
metabolites, one should pay extra attention to resonance degeneracy. The ^{1}H
resonances of metabolites can be assigned by analyzing their ^{1}H–^{1}H spin system
coupling patterns and comparing their chemical shifts with reported values (Fan 1996;
Lane and Fan 2007; Lindon et al. 1999; Nicholson and Foxall 1995; Teng et al. 2009;
Yang et al. 2007a, b), standard compounds (Ulrich et al. 2008; Wishart et al. 2007),
and Chenomx database entries (Edmonton, Alberta, Canada). Some resonances
observed in the aliphatic region (lower frequency than that of water) can readily be
assigned to metabolites such as alanine (Ala), valine (Val), leucine (Leu), isoleucine
(Ile), lactate (Lac), glycine (Gly), and aspartate (Asp) based on their characteristic
chemical shifts. Assignments of tryptophan (Trp), tyrosine (Tyr), phenylalanine
(Phe), and histidine (His) can be done by starting at the aromatic resonances and
then extended to side chain protons. NADP$^+$ is usually assigned by the spin system
formed near 8.41, 8.81, 9.12, and 9.32 ppm in the TOCSY spectrum. ATP/ADP/AMP
(AXP) and GTP/GDP/GMP (GXP) can be assigned based on their readily distin-
guishable ^{1}H chemical shifts, ^{1}H–^{1}H and ^{1}H–^{13}C coupling patterns. Singlet
resonances such as choline (Cho), Phosphatidylcholine (PCh), creatine (Cr),
trimethylamine *N*-oxide (TMAO), Gly, formate (For), Formarate (Frt), pyruvate

Fig. 9.25 Y-predicted scatter plots of OPLS-DA models validated with 86 % of the original data used as a training set (6 per class) and the remaining 14 % (1 per class) used as a test set to validate the model. All test samples are correctly predicted with each locating within the region of its corresponding classes. The models were generated for (**a**) cell extracts and (**b**) culture media of human brain cancer cells (LN229) treated with different dose levels of PTX
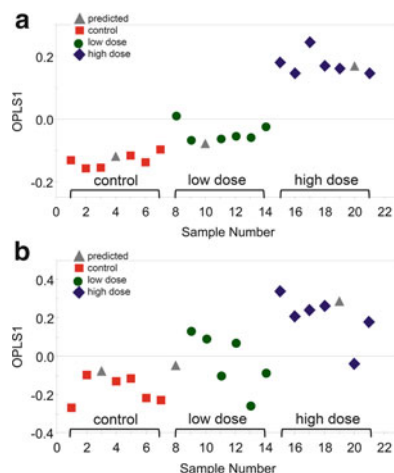
(Pyr), and malonate (Mlo) are often assigned by comparing the resonances with published values. An HSQC spectrum is needed to verify the assigned resonances. Figure 9.26 shows typical PRESAT and HSQC spectra of intracellular extracts. An example protocol to identify metabolites using Chenomx is provided in Table 9.9.

9.6   Metabolomics of Biofluids

Biofluids are mostly used in metabolomics applications (see Fig. 9.1). Since it is not feasible to cover all aspects of the applications of biofluid metabolomics, this section intends to provide a brief introduction on how the metabolomic research is conducted using a metabolomics study of cerebral infarction as a case study (Jung et al. 2011). The key questions to be addressed in this section include the following:

1. What are the metabolic profiles of human urine and plasma observed by NMR spectroscopy?
2. How is multivariate statistical analysis applied to study the changes of urine and plasma profiles?
3. How are the results of the metabolomics study interpreted?
4. Can the results of the biofluid metabolomics be linked to specific metabolic pathways?

Cerebral infarction is an ischemic type of stroke, which occurs when blood flow to a part of the brain stops due to complex health factors such as cardiovascular disease, high cholesterol, and/or diabetes. Because it is very difficult to diagnose and prognose a stroke using traditional methods, NMR-based metabolomics was applied to study metabolic profiles of urine and plasma samples from healthy individuals and patients with cerebral infarction. By applying MVA to the NMR data, metabolic indicators were identified for early diagnosis of cerebral infarction.
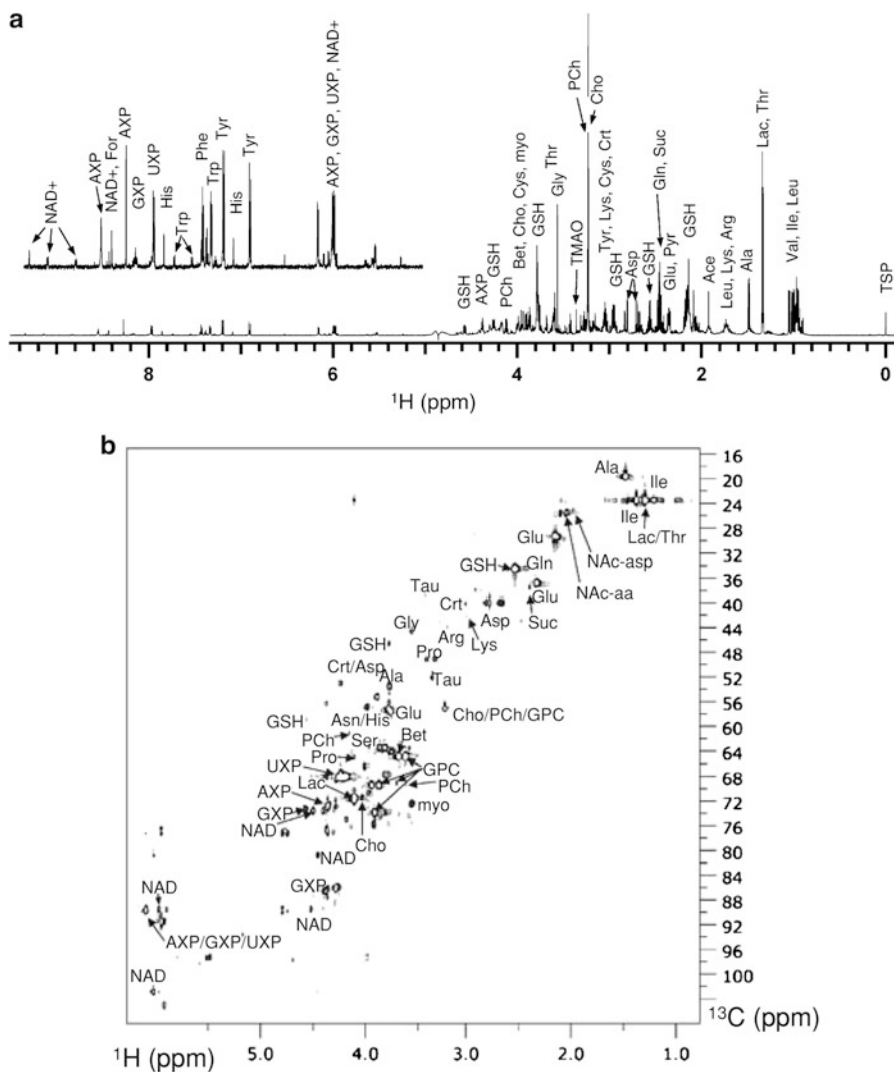
Fig. 9.26 Representative 600-MHz ^{1}H-NMR spectra of a polar cellular extract. (**a**) 1D PRESAT spectrum. (**b**) ^{1}H-^{13}C HSQC. Some of the assigned resonances are labeled in both spectra (reproduced with permission from Teng et al. (2009), Copyright © Springer, 2009)

A total of 101 plasma (47 healthy individuals and 54 patients) and 54 urine (27 for each group) samples were used in the study. For 1D NOESY experiments, 200 μL of plasma and 400 μL of urine were used with a phosphate buffer. After spectral alignment, each spectrum was segmented into 0.005 ppm for plasma samples and 0.003 ppm bins for urine samples and normalized to the total spectral area. Representative ^{1}H NMR spectra of plasma and urine samples are shown in Fig. 9.27. By comparing the spectra of healthy individuals to those of patients,

Table 9.9 Protocol for profiling metabolites using Chenomx profiler (Chenomx, Inc.)

1. Start Chenomx Profiler by clicking on Chenomx Profiler shortcut icon

Load data

2. Load a raw 1D NMR data file: Select "File" → Open, and select a .fid file from Open Spectrum window and click Open

3. Input parameters in the Data import window for chemical shift (CS) reference, concentration (mM), pH, etc. Then, click OK

Process data

4. If the spectrum needs to reprocess, send the data to Chenomx Processor: click File → Send to Processor

  (a) Zoom in the reference range by dragging the LMB on the spectral range. To zoom out, drag the LMB on the small spectral overview window (lower right window)

  (b) Move the red triangle to adjust CS reference ppm

  (c) To redo the phase correction, click "Clear last" twice. Click toolbar Processing → Phasing. Sliding the bars for manual phase correction, or select auto with the fine adjustment. Click Accept

  (d) Baseline correction

     (i) Clicking Processing → Baseline Correction

     (ii) Click on Auto Spline (or Auto Linear)

     (iii) Move blue points and/or add more points (click on the spectrum) if necessary to achieve a flat baseline

     (iv) Click on Accept

  (e) Switch to Profiler by clicking "Send to Profiler" icon

Profile metabolites (identify compounds and determine concentrations)

5. Click on a metabolite from the library, then click a CS value. Move the arrows to fit the peak(s). Repeat this step all assigned peaks.

6. Zoom in a range and right-click on an unassigned peak → Filter for compounds

7. Select a compound from the list. Move the blue triangle on x-axis to align CS position and on y-axis to align intensity. Alternatively, the adjustment can be done by dragging the top of the calculated peak to overlap the spectral peak

8. Click on the CS value (if any) under the metabolite name to select another resonance of the metabolite and repeat step (6–7)

9. Click save icon and give a name to the profile

10. Perform similar analysis for all peaks

11. Uncheck Chenomx library compound set (only profiled compound set is checked)

Export analysis results

12. Click File → Export → Profiled Data (.txt)

13. Answer all question accordingly, and use Transposed for export format

14. Click Finish to export. The text file is saved in the same folder as the NMR data file

significant changes in metabolic concentrations can be visualized, including plasma very low density lipoprotein (VLDL), low density lipoprotein (LDL) lactate and formate, and urine citrate, DMA, creatinine, and hippurate. PCA and OPLS-DA were used to further investigate metabolic changes between healthy individuals and the patients.

Since PCA scores plots of both plasma and urine models showed slight differentiations between the two classes, OPLS-DA was conducted to minimize the
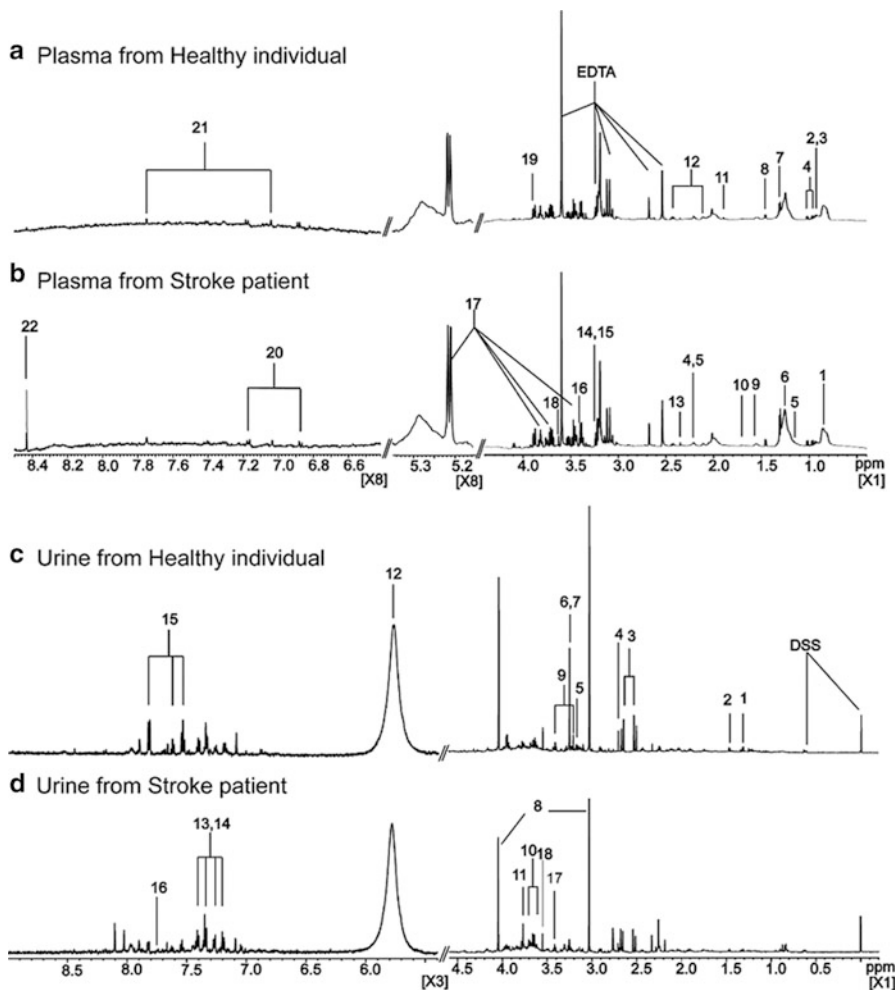
Fig. 9.27 ^{1}H-NMR spectra of plasma (**a**, **b**) and urine (**c**, **d**) of healthy individuals (**a**, **c**) and stroke patients (**b**, **d**). Labels for plasma samples: 1, very low-density lipoprotein (VLDL)/low-density lipoprotein (LDL) CH$_3$; 2, leucine; 3, isoleucine; 4, valine; 5, 3-hydroxybutylate; 6, VLDL/LDL (CH$_2$)$_n$; 7, lactate; 8, alanine; 9, lipid CH$_2$CH$_2$CO; 10, lipid CH$_2$CH$_2$C_C; 11, acetate; 12, glutamine; 13, pyruvate; 14, trimethylamine-*N*-oxide; 15, betaine; 16, taurine; 17, glucose; 18, methanol; 19, glycolate; 20, 4-hydroxyphenylacetate; 21,τ-methylhistidine; 22, formate. Labels for urine samples: 1, lactate; 2, alanine; 3, citrate; 4, DMA; 5, *O*-acetylcarnitine; 6, trimethylamine-*N*-oxide; 7, betaine; 8, creatinine; 9, carnitine; 10, mannitol; 11, glycolate; 12, urea; 13, phenylacetyglycine; 14, phenylalanine; 15, hippurate; 16, salicylurate; 17, taurine; 18, glycine (reproduced with permission from Jung et al. (2011), Copyright © American Heart Association, Inc., 2011)
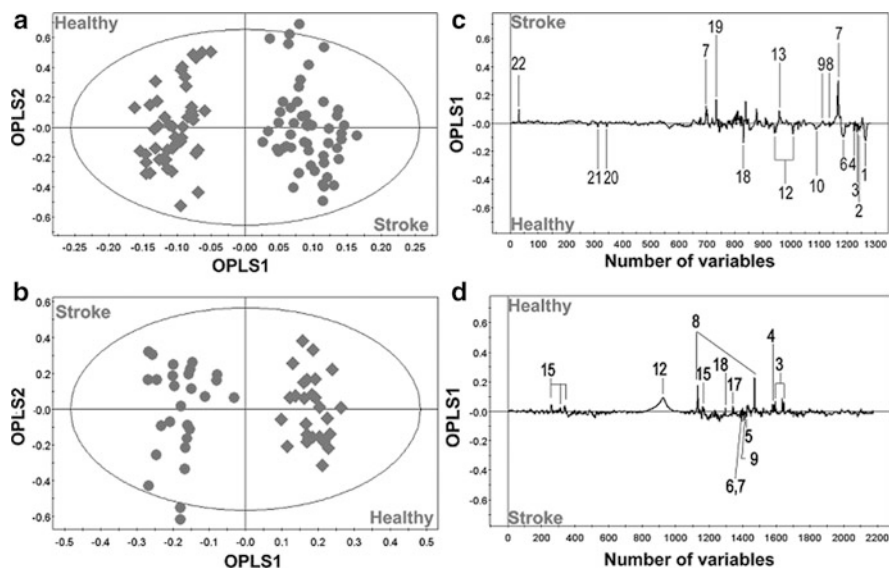
Fig. 9.28 OPLS-DA scores (**a**) and loadings plots (**c**) obtained from the ^{1}H-NMR spectra of plasma samples from healthy individuals ($n = 47$) and stroke patients ($n = 54$). OPLS-DA scores (**b**) and loadings (**d**) plots obtained from the spectra of urine samples from healthy individuals ($n = 27$) and stroke patients ($n = 27$). See Fig. 9.27 for the metabolite labels (reproduced with permission from Jung et al. (2011), Copyright © American Heart Association, Inc., 2011)

contribution of the variability within the classes to the models so as to focus on the variability between the two classes. The scores plots of plasma (Fig. 9.28a) and urine data (Fig. 9.28b) show significant separations between healthy individuals and patients. In addition, the high R^2 and Q^2 values for both plasma and urine ($R^2 > 0.9$, $Q^2 > 0.6$) data indicate that the models have good quality and predictability. The models were validated with 80% samples as a training set and 20% as a test set. The validation results for plasma model are shown in Fig. 9.29. All test data were correctly predicted by the models.

OPLS-DA loadings plots of plasma and urine models were used to identify the metabolites that contribute to the differentiations between the two classes in the scores plots (Fig. 9.28c, d). The loadings plot of plasma model reveals that compared to healthy individuals, plasma of the stroke patients have elevated levels of several metabolites including lactate, alanine, lipid -CH$_2$CH$_2$CO, pyruvate, glycolate, and formate, along with decreased levels of VLDL/LDL, leucine, isoleucine, valine, unsaturated lipid, glutamine, methanol, 4-hydroxyphynellactate, and τ-methylhistidine. The urine loadings plot indicated that urine metabolic profiles of the stroke patients were characterized by increased levels of O-acetylcarnitine, TMAO, betaine, and carnitine, and decreased levels of citrate, DMA, creatinine, taurine, glycine, and hippurate. To further investigate the significance of the metabolite changes, Student's *t*-test was applied to the NMR bins of the metabolites
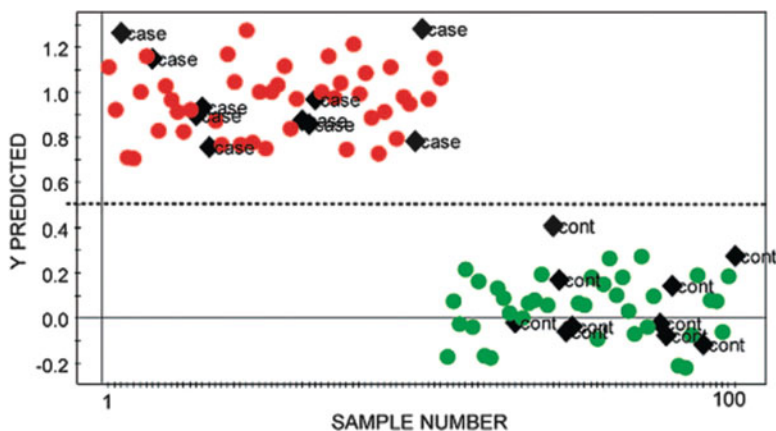
Fig. 9.29 Validation plot of OPLS-DA model validated with 80 % (44 stroke, 37 healthy) of the data as the training set and the remaining 20 % [10 stroke (case) and 10 healthy (control)] as the prediction set (reproduced with permission from Jung et al. (2011), Copyright © American Heart Association, Inc., 2011)

(see Table 9.10 for an example protocol of Student's *t*-test application). The changes with calculated *p* value of 0.05 or less is usually considered as statistically significant.

Shown in Fig. 9.30 is the summary of metabolic pathways involving the metabolites of the stroke patients with statistically significant changes compared to healthy individuals. The pathway map is used to explain the pathogenesis of a stroke from the metabolic biology's point of view based on the results of the multivariate statistical analyses. As the map indicates, the metabolites with altered levels in stroke patients are primarily involved in three pathways: anaerobic glycolysis, folic acid pathway, and hyperhomocysteinemia.

Anaerobic glycolysis. It is well known that under hypoxia, cells rely more on glycolysis to generate energy than on the tricarboxylic acid cycle (TCA cycle), which converts glucose into pyruvate to form the high-energy compound ATP (adenosine triphosphate) and reducing agent NADH (reduced nicotinamide adenine dinucleotide). Under the hypoxia condition, a majority of pyruvate does not enter the TCA cycle and instead is oxidized to form lactate. The observed higher levels of lactate and pyruvate in the patient's plasma indicate the enhanced anaerobic glycolysis of plasma glucose to meet the energy consumption because of limited availability of oxygen in the patient's body. Additionally, a decreased citrate level is an indicator that a limited amount of pyruvate entered the TCA cycle.

Folic acid deficiency. In this study, the increased level of formate coupled with the decreased levels of glycine and hippurate was related to folic acid deficiency, which has been considered as a risk factor for stroke (Howard et al. 2002). One of the folic acid metabolic activities is to form tetrahydrofolate (THF) that is further converted

Table 9.10 Protocol for generating *t*-test different spectrum and plotting using Igor (WaveMetrix)

Generate text file of t-test different spectrum

 1. Read an excel file for PCA analysis
 2. In cell P2, write a formula for average difference = SUM(I2:O2)/7-SUM(B2:H2)/7 (note: 7 controls and 7 exposed samples)
 3. In cell Q2, write *t*-test formula: =TTEST(I2:O2,B2:H2,1,1) (one-tail and paired)
 4. In cell R2, write IF formula: =IF(Q2 < 0.05,P2,0) [this means that if the value in cell Q2 is less than 0.05, then copy the value of cell P2 to R2. If same or greater, write 0 in R2]
 5. Change #DIV/0! to 0 (by copying 0 and then pasting 0 to those cells)
 6. Copy column 1 to a new file
 7. Copy last column and paste the values to the new file and save the file as *.txt (tab delimited)
 8. Repeat steps 2–7 for a different dose class of samples

Plot the text file using Igor

 9. Run Igor
10. Click File → Open File → Notebook and brow the directory where the text file is stored
11. Click Data → Load wave → general text (not delimited text)
12. Select the text file and click open
13. Type in name for the first column, e.g., wave1 and check double precision, overwrite, make table
14. Click on the second name field and type in a name, wave2 click Load
15. Click Windows → New graph. Select wave2 for Y Wave (left axis), wave1 for X Wave (bottom axis)
16. Click do it
17. Click Graph → Modify trace appearance and change color, then click Do it
18. Right-click on the graph and select axis properties
19. Click Axis Range
20. Click wave2, then select bottom (up left-corner)
21. Click Swap. The spectrum should swap *X*-axis. If not, click Swap a few more times until it swaps
22. Click wave2 and select left for axis
23. Click Manual Min and increase the range a little bit
24. Click Do it
25. Click File → Save graphics and give a name for the graph. Make sure it is in JPEG format and set resolution to higher than 600
26. If there are nonzeros in the cutoff ranges (water and HEPES), copy zeros to the rows in the text file and reload wave file (step 9)
27. To zoom the spectrum, left-click and drag the mouse arrow on the spectrum and adjust the select box. Right-click inside the box and select Expand
28. To reduce the spectrum, select Shrink in the box
29. You can also change the axis range in Axis properties menu
30. To remove labels and axis line(s), click Axis submenu and change color to white
31. To change labels, click Axis label submenu and type text
32. To add ticks, bottom axis, Auto Ticks, Approximately 10, Minor ticks in 5 steps
33. To save the experiment, click File → Save experiment as and give a name
34. To load the graph in an experiment, open the experiment and then click Windows → graph macro and select the macro
35. To save the graph into a pdf, use 0.1 size for line: Graph → Modify Trace Appearance

---

(continued)

Table 9.10 (continued)

36. For multiple plots in same file
    (a) To add plots onto an existing graph, Click Graph → Append Traces to Graph. Select the data you would like to add, then click Dot It. Select all of the datasets you would like to add. The X Wave should never change; it is always the ppm values
    (b) The plots will appear on top of the other graph
    (c) Right-click on the new graph and click Modify Trace Appearance. All of the datasets will appear in the box under Trace. Highlight the one you would like to modify, and change the color to black
    (d) For the new plots, check the box that says Offset. A dialog box will open. Change the value of Y offset so that the two graphs are not touching. Click Ok to close the dialog box, then click Do It to apply the changes
    (e) The new graph should now be offset from the old one and in black
    (f) Continue to offset each new dataset by the same intervals, making sure that none of them are touching
    (g) To add labels, click Graph → Add Annotation. In the area that says Annotation, type the name of the dataset you would like to label. Click Frame, and in the dropdown box next to Annotation Frame, select None. Then click Do It
    (h) A text box should appear somewhere on the graph. Drag the text to the plot that you would like to label
    (i) Label each plot separately

Notes

1. If there are gaps in the plot
   a. The cut regions of solvents in the original spectra will cause gaps in the t-test filtered difference spectra.
   b. To correct this, go into the text or excel files, and insert zeros
   c. When you reconstruct the plot, there should now be a continuous line
2. If the plot is inverted from what it should be
   a. Check the Mnova file for the dataset
   b. If all of the graphs are upside down, this is the cause
   c. Highlight all of the samples, and click on the phase correction icon. Click the box that shows +180
   d. The graphs should now all be flipped over and upright
   e. Reprocess the data to get new normalized data and *t*-test
   f. Reconstruct the plotting in Igor to get a corrected difference plot
3. If the plot shows an abnormal amount of dispersion (lines zigzagging above and below the baseline, looking like positive and negative readings for the same points)
   a. Make sure all of the samples have had their phase adjusted properly
   b. The likely cause is miscalibration of the TSP points
   c. Check peak alignment using Mnova
   d. Reprocess the data to get new normalized data and *t*-test
   e. Reconstruct the plotting in Igor to get a corrected difference plot
4. To change the title of a graph
   a. Press Ctrl+y
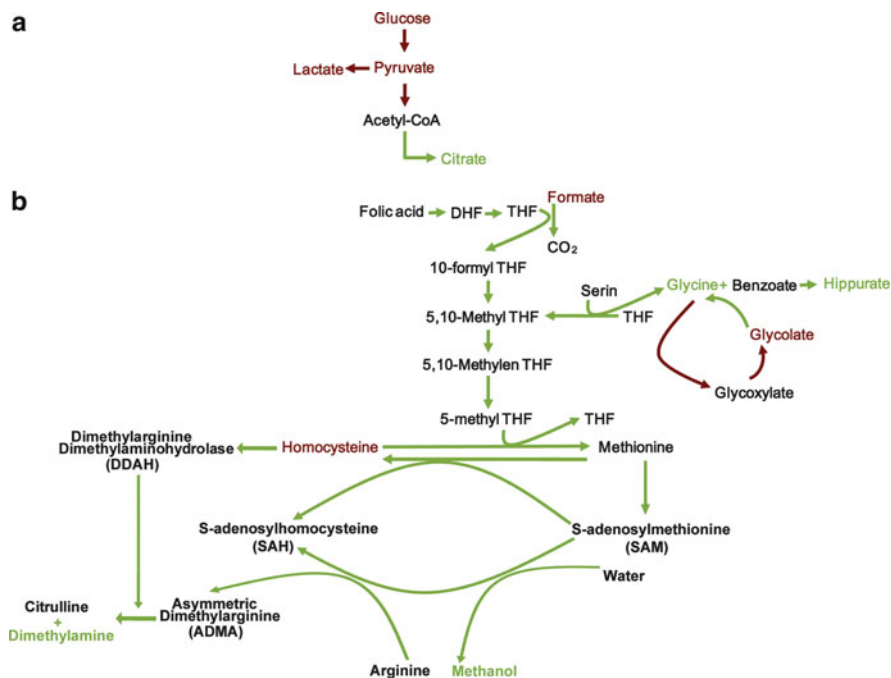   b. A window will open where you can edit the title of the graph

Fig. 9.30 Summary of the metabolic pathways including some of the detected metabolites. *Red* indicates metabolites with increased level due to cerebral infarction. *Black* indicates metabolites not found or not identified. *Green* indicates metabolites with decreased levels. The anaerobic glycolysis pathway was particularly enhanced after cerebral infarction (**a**). Many metabolites were detected because of folic acid deficiency (**b**) (reproduced with permission from Jung et al. (2011), Copyright © American Heart Association, Inc., 2011)

to 10-formyl-THF via the oxidization of formate in the folate pathway. The elevated level of formate may be an indicator of folic acid deficiency. Additionally, glycine is biosynthesized from serine through the biodegradation of folate, which is metabolically converted to hippurate via the conjugation with benzoate in the liver (Gatley and Sherratt 1977). Decreased levels of glycine and hippurate further support the explanation of a folic acid deficiency.

Hyperhomocysteinemia. Hyperhomocysteinemia is a high level of homocysteine in the blood, as a result of folic acid deficiency (Miller et al. 1994) and has been linked to an increased risk of stroke (Homocysteine Studies Collaboration 2002; Howard et al. 2002). A high level of homocysteine causes oxidative injury to vascular endothelial cells and damages the production of nitric oxide in the endothelium, which is a strong vascular relaxing factor (Dardik et al. 2000). In addition, hyperhomocysteinemia increases platelet adhesion to endothelial cells, which may obstruct blood vessels. The observed increased level of homocysteine in the patients is consistent with the previous finding. *S*-adenosylmethionine is metabolically converted to methanol, which is formed via the homocysteine biodegradation (Axelrod and Daly 1965). Therefore, decreased methanol in stroke patients may

also be related to hyperhomocysteinemia that leads to limited availability of methionine and *S*-adenosylmethionine.

In summary, by applying metabolomics, the study of stroke patient's NMR metabolic profiles has identified a set of metabolites as metabolic indicators for early diagnosis of cerebral infarction, which are involved in three metabolic pathways: anaerobic glycolysis, folic acid pathway, and hyperhomocysteinemia.

9.7 Cellular Metabolomics

In this section, a case study is discussed to illustrate how a cell-based metabolomics study is conducted. The experiments were designed to study the cellular responses of different types of human breast cancer cells (MCF-7 and MDA-MB-231) to the treatment of a synthetic estrogen, 17α-Ethynylestradiol (EE2). MCF-7 cells possess estrogen receptors (ER), whereas MDA-MB-231 cells are ER-negative. ERα (one of two different forms of the estrogen receptors) is an important predictive and prognostic marker in human breast cancer, being expressed in over 60% of cases. Binding of estrogens to ERα leads to transcriptional activation of ER target genes that are important in cell proliferation, angiogenesis, and survival of breast cancer cells (Green and Carroll 2007). EE2 was the first orally active synthetic steroidal estrogen in contraceptive pills, and is now commonly used in the estrogen–progestin combination preparations. EE2 is hormonally effective by activating the estrogen receptor.

9.7.1 Experiments

9.7.1.1 Cell Treatment and Quench

In this section, examples will be described on:

1. How are adherent human cells cultured for metabolomics?
2. How are cells seeded and treated with a stressor?
3. How are the treated cells quenched for metabolomics studies?
4. What is a suitable cell quench method for metabolomic studies?

Human breast cancer cell lines MCF7 and MDA-MB-231 were cultured in a humidified atmosphere of 5% CO_2 using low glucose DMEM supplemented with 10% (v/v) heat-inactivated FBS, 100 units/mL penicillin, 100 μg/mL streptomycin, and 25 mM HEPES buffer at 37°C. It is worth noting that high glucose medium is not suitable for culturing cancer cells because low glucose medium has a glucose concentration to mimic the physiological concentration of glucose in solid tumor. In addition, a high glucose level can cause interference during metabolomic data analysis because the intense glucose peaks will dominate the contribution to the statistical modeling. The cells need to be thawed on ice, if they are stored in a liquid

nitrogen dewar or a $-80°C$ freezer. The medium for cell storage should be removed immediately after cells are thawed because 5% DMSO in the solution is toxic to the cells, which is used as a cryoprotectant to protect the cells from damage caused by ice formation when frozen. The cells were initially seeded in a T-25 tissue culture flask with 5 mL DMEM described above and cultured at $37°C$ and 5% CO_2 for 2 weeks before they were expanded to four T-75 tissue culture flasks with 12 mL DMEM in each flask. The medium was changed twice weekly.

When they reached 80% confluence, the cells were harvested and seeded onto 21 dishes (6 cm) for 24 h treatment with approximately 3×10^5 cells in 3 mL DMEM per dish without HEPES buffer and 21 dishes for 48 h treatment with approximately 2×10^5 cells per dish. For each treatment, there were 3 dose levels, including 0 µg/L (controls), 5 µg/L (low dose), and 10 µg/L (high dose) and 7 replicates per class (per dose level). After seeded, the cells were allowed to be incubated for 2 h to be able to attach on the culture dish surface before the treatment, which increases the reproducibility of the experiments. For each dish, a 3 µL dose solution was added, which were 0, 0.5, or 1.0 µg/L EE2 for control, low dose, or high-dose samples, respectively.

When the treatment was done (at the end of 24 or 48 h), the cells were quenched immediately to instantly stop all enzymatic reactions using a direct quench method (Sect. 9.3.6, Teng et al. 2009). Briefly, after the culture medium was removed from the culture dish, cells were quickly washed twice with ice cold PBS (pH 7.4). The residual PBS was removed by vacuum (Fig. 9.6a). Cells were then quenched using 0.8 mL HPLC grade methanol (Fig. 9.6b). Next, the cells were gently detached from the culture dish using a cell lifter (Fig. 9.6c). The methanol solution containing the quenched cells was pipetted into a 2 mL Safe-lock® centrifuge tube containing a 3.2 mm stainless steel bead for extraction.

9.7.1.2   Extraction of Intracellular Metabolites

Intracellular metabolites were extracted using a dual phase extraction procedure (Sect. 9.3.4, Ekman et al. 2008; Viant 2007). First, 127.5 µL ice cold DI water was added to each tube. The tubes were placed in a tissuelyser (e.g., a Qiagen tissuelyser) and shaken for 10 min at 15 Hz. The second step is to add 300 µL ice cold chloroform, then shake cell samples again for 20 min at 15 Hz. Next, 300 µL ice cold chloroform and 300 µL ice cold DI water were added followed by shaking the samples for another 10 min at 15 Hz. The samples were centrifuged at $1,000 \times g$ at $4°C$ for 10 min. The solutions were separated into an upper methanol/water phase containing polar metabolites and a lower chloroform phase containing lipophilic metabolites. Proteins and other biological macromolecules are precipitated by the addition of methanol and chloroform, and entrapped in the middle layer between the aqueous and organic phase. The upper and lower phases were collected as described in Sect. 9.3.6.

9.7.1.3 NMR Sample Preparation

The solvents were removed under vacuum at room temperature using a vacuum concentrator (or speedVac; 4 h for chloroform samples and 6 h for aqueous samples). The samples were reconstituted in 230 μL 0.1 M phosphate buffer (pH 7.4) with 20 μM TSP, vortexed for 30 s and centrifuged at 10,000 $\times$ g for 10 min. The last step of the sample preparation was to transfer the sample into 3 mm NMR tubes using a gel loading tip for each sample. Careful inspection of each NMR tube should be done for air bubbles. If air bubbles are formed inside the tube, storing the tube at 4°C will help remove the air bubbles. If air bubbles still exist inside the tube after a few hours in a refrigerator, vacuum can be used to remove the air bubbles by placing the tubes in a 1 L bottle with a screw cap through which vacuum is applied.

9.7.1.4 NMR Experiments

The NMR data for all samples were acquired using 1D PRESAT experiment. For experiment setup, ^{1}H channel of the probe was tuned by minimizing the power reflection of the transmitter. The shimming for the first sample was done by gradient shimming of Z_1–Z_4. Since the diameter of a 3 mm tube is relatively small, changes in higher order of Z shims (Z_5 and Z_6) will not influence spectral line width noticeably. Next, transverse shims were optimized by adjusting $x, zx, y, zy, xy,$ and x^2–y^2. After shimming was done, ^{2}H lock frequency was set on ^{2}H$_2$O resonance, followed by adjusting lock power, gain and phase. RF calibration includes the calibrations of 90° pulse width, transmitter frequency, and saturation frequency that were set on water resonance.

   The dead delay (the delay between last RF pulse and start of acquisition) were adjusted so as to eliminate error in the first-order phase correction—the first-order phase correction equals to zero. This step makes it easier to phase the whole set of the 1D spectra during data processing (see Sect. 9.7.2). The last step was to set experimental parameters, including spectral width of 12 ppm, acquisition time of 2 s (or longer), saturation duration of 2 s, dummy scans of 8, number of transients, and receiver gain. Before the acquisition of each sample, autoshimming must be performed by autogradient shimming of Z_1–Z_4. Autolock should not be used during any stage of the experiment because it will change lock frequency, resulting in poor water suppression. The data was stored every 16 transients. As a measure of spectral quality, the line width of DSS (or TSP) in a 1D ^{1}H NMR spectrum acquired on a 600 MHz NMR spectrometer is usually less than 1.0 Hz.

9.7.2   NMR Data Processing

After the experiments are done, the data can be transferred to an off-line computer for data processing. There are a variety of NMR software packages that can be used for metabolomics, including Mnova, Spinworks, ACD 1D NMR, Nuts, MestreC, etc. Because of its useful features such as peak alignment, spectral normalization and spectral superimposement, Mnova is used for all data processing discussed here. All data for the treatment study are loaded into Mnova. Line broadening of 0.3 Hz is applied to all spectra. The first-order phase constant is set to zero, followed by adjusting zero-order phase correction. The phases of all data are usually identical if the data is collected in automation for the whole set of samples. The baseline of the spectra is corrected using Bernstein polynomial fit with third polynomial order. All spectra are then superimposed together, after chemical shifts are referenced to internal standard DSS (or TSP). The residual water peak is then removed from the superimposed spectrum and chemical shift reference is inspected again.

To reduce the size of the data points that are variables for the statistical modeling, the spectra are binned into fewer data points before submitted for statistical analysis. Spectra with a range of 0.50–10.00 ppm are segmented into 0.005 ppm wide bins. In order to remove the unwanted systematic errors among the samples, spectral normalization is frequently used before the NMR data is used for statistical modeling. Therefore, the total area of the binned spectra is normalized to 1. The binned normalized superimposed spectrum is then saved as a text file and imported into Microsoft® Excel®. The sample labels are inserted into the first row of the text file. The resulted text file contains chemical shift values in the first column of the data file, which are variables, while the rest of the columns are the data for samples that are observations for PCA or other statistical analyses.

9.7.3   Principal Component Analysis

As discussed in Sect. 9.2, PCA is used to reduce dimensionality of the data for finding the variance among the observations (or samples). In this case, PCA decreases dimensionality from >2,000 variables to two principal components (PC1 and PC2). There are several statistical analysis software packages available such as SIMCA-P (Umetrics), PLS-Toolbox (or Solo, a stand-alone version of PLS-Toolbox; Eigenvector Research), MetoboAnalyst (web-based freeware; http://www.metaboanalyst.ca), and R (a freeware for statistical computing and graphics; http://www.r-project.org/).

Regardless of the statistical software used for metabolomics studies, it is necessary to preprocess the binned NMR data before the analysis is preformed. After all bins are mean-centered and Parato-scaled, principal components analysis (PCA) was conducted using all binned spectra. All samples of MCF-7 ($n = 6$) and MDA-MB-231 cells ($n = 7$) were observed to fall inside of the Hotelling's T2 ellipse at
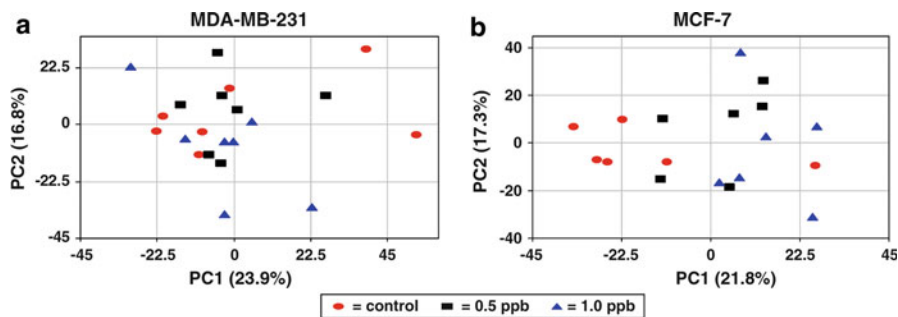
Fig. 9.31 Two-component scores plots from PCA models built with ^{1}H NMR spectra of the extracts of the aqueous extracts of (**a**) estrogen receptor (ER)-negative human breast cancer cells (MDA-MB-231), and (**b**) ER-positive MCF-7 cells following the treatment of 17a-ethynylestradiol (EE2) at the level of 0 (control), 0.5 ppb and 1.0 ppb for 48 h (reproduced with permission from Teng et al. (2009), Copyright © Springer, 2009)

the 95% confidence interval in scores plots of the first two components, meaning that there were no outliers. Loadings plots were used to identify the metabolites responsible for the separation in scores plots.

Shown in Fig. 9.31a is a two-component (PC1/PC2) scores plot of a PCA model built for the MDA-MB-231 cell line treated with EE2, which shows that there is no discernable separation in score values among the three groups, where group (or class) is defined by treatment level (control cells, cells treated with 0.5 ppb EE2, and those treated with 1.0 ppb EE2). The results indicate that MDA-MB-231 cells do not have significant response to the treatment of EE2. This can be explained by the fact that MDA-MB-231 cells are ER negative, and thus are not influenced much by the presence of estrogen EE2. The same type of plot is displayed for the MCF-7 cells treated with EE2 (Fig. 9.31b). In the MCF-7 scores plot, the classes are well separated primarily along the PC1 that is the component describing the maximum variation in the dataset. Scores for the 1.0 ppb treated class are clustered to the right region of the origin, while those for the control class are clustered to the left, with the exception of one outlier in the control class. Scores for the low-dose treated class (0.5 ppb) distribute mostly between the control and the high-dose treated classes. Thus, PCA (a widely used and unsupervised multivariate method) is able to capture the dose-dependent changes in metabolite profiles associated with the treatment for the case of ER-positive cells (MCF-7), which is a widely used and unsupervised multivariate method. In contrast, the treatment with EE2 does not induce significant metabolic changes in ER-negative cells (MDA-MB-231). These results demonstrate that cell-based metabolomics is a valuable tool to study the cellular responses to the presence of external stressors.

To further investigate the effect of EE2 on metabolic profiles in these two cell lines, univariate analysis can be conducted using the normalized and binned spectral spreadsheet (but not scaled) (Ekman et al. 2008). First, an "average class spectrum" is obtained by averaging the binned spectra across all class members ($n = 6$ for MCF-7 or $n = 7$ for MDA-MB-231). Next, a difference spectrum is generated by
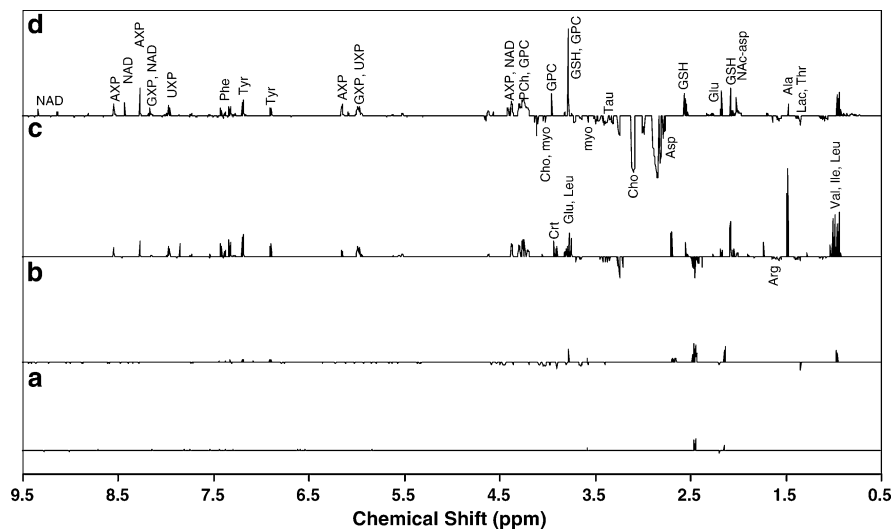
Fig. 9.32 The "*t*-test filtered difference spectra" generated from ^{1}H NMR spectra of cell extracts after the treatment with 17α-ethynylestradiol (EE2): (**a**) MDA-MB-231, 0.5 ppb, (**b**) MDA-MB-231, 1.0 ppb, (**c**) MCF-7, 0.5 ppb, and (**d**) MCF-7, 1.0 ppb (ppb = μg/L). Positive peaks correspond to metabolites that increase upon treatment, whereas negative peaks are from metabolites that decrease upon treatment (reproduced with permission from Teng et al. (2009), Copyright © Springer, 2009)

subtracting the average class spectrum of the control class from that of each treated class. Then, a Student's *t*-test is applied to each bin to test the statistical significance of each averaged bin using a *P*-value <0.05 where *P* is the higher probability of the Student's paired *t*-test. If the *P*-value is less than 0.05, the average for the treated class is considered to differ significantly from that of the relevant control class. If not, the bin value for the difference spectrum is replaced with a zero. This results in a "*t*-test filtered difference spectrum" for each treatment class. Positive peaks represent metabolites that increase in concentration (with statistical significance) upon treatment, whereas negative peaks are from metabolites that decrease in concentration.

Shown in Fig. 9.32 are four *t*-test filtered difference spectra of MDA-MB-231 and MCF-7 cells treated with EE2. Because all four difference spectra are displayed with the same intensity scale, the overall changes of the metabolic profile for each treatment class can be evaluated by comparing the number and intensity of peaks in each spectrum. Being consistent with PCA results, these spectra confirm that the treatment of EE2 induces much larger changes in overall metabolite profiles of MCF-7 cells (Fig. 9.32c, d) than those of MDA-MB-231 cells (Fig. 9.32a, b). Additionally, it is evident that the effect of EE2 on MCF-7 cells is dose-dependent, by comparing the difference spectrum from the low-dose treatment (Fig. 9.32c) with that of the high-dose treatment (Fig. 9.32d). It is worth noting that many metabolites with low-frequency resonances (i.e., lower ppm values) have an

increase in concentration in the EE2-treated MCF-7 cells (Fig. 9.32c, d) compared to the controls, including valine (Val), leucine (Leu), isoleucine (Ile), alanine (Ala), threonine (Thr), glutamate (Glu), *N*-acetyl aspartate (NAc-asp), Glutathione (GSH), and PCh, while decreases were observed for myoinositol (myo), choline (Cho), taurine (Tau), and glutamine (Gln). In the high-frequency region of the water resonance (i.e., higher ppm values), the treated MCF-7 cells show increases in concentrations of nicotinamide adenine dinucleotide (NAD^+), phenylalanine (Phe), Tryptophan (Trp), tyrosine (Tyr), histidine (His), and the nucleotides Guanosine-phosphate (GXP, where X is M, D, or T) and adenosine phosphate (AXP). There is very minor change in the metabolic profiles of the treated MDA-MB-231 cells compared to the treated MCF-7 cells.

9.7.4   Partial Least Squares for Discriminant Analysis

Partial Least Squares for Discriminant Analysis (PLS-DA) is a regression extension of PCA, which builds a linear regression model using class information to discriminate the classes in a new set of components. Because it uses the class information to achieve the utmost separation, PLS-DA is a supervised statistical analysis, and hence models generated by PLS-DA must be undergone a validation process.

Shown in Fig. 9.33a, b are the PLS-DA scores plots for the first two components of MDA-MB-231 and MCF-7 cells, respectively. Similar to its PCA scores plot, MDA-MB-231 cells do not show any clear separations between the classes. Because PCA determined one of MCF-7 control samples as an outlier, the outlier was excluded in the PLS-DA. The class separation is improved slightly in PLS-DA, compared to that in PCA excluding C5. The validation results with 50 permutations are shown in Fig. 9.34, with R^2 and Q^2 intercepts of 0.19 and -0.27, respectively. All 50 permutations have generated models with smaller Q^2 than those of the original model, indicating that overall PLS-DA of MCF-7 samples yields a valid model.

In the 1920s, Warburg et al. discovered that one of the hallmarks of cancer is the increase of glucose uptake by cancer cells to meet the energy demand for the rapid cell growth (Warburg et al. 1924; Warburg 1956), which can cause the increases in intracellular concentrations of energy-containing molecules such as ATP, GTP, and NDA^+. Because of the lack of estrogen receptors, the treatment of EE2 should not cause any impact on the intracellular metabolic profile of the MDA-MB-231 cells. This is clearly shown by the *t*-test filtered difference spectra of treated MDA-MB-231 cells (Fig. 9.32a, b). The changes in the metabolic profiles of treated MDA-MB-231 cells are negligible, although there were few changes in the high-dose treatment (Fig. 9.32b). In the ER-positive MCF-7 cells, however, the treatment of EE2 significantly increases the levels of ATP, GTP, and NAD^+ (Fig. 9.32c, d), indicating that the EE2 treatment may promote cell proliferation in the MCF-7 cell line. Furthermore, the increase in concentration of several key amino acids in the treated MCF-7 cells indicates that the treatment of EE2 may also increase the availability of the amino acids for protein synthesis in MCF-7 cells. The rapid growth of cancer cells requires not only energy but also proteins in addition to lipids and fatty acids.
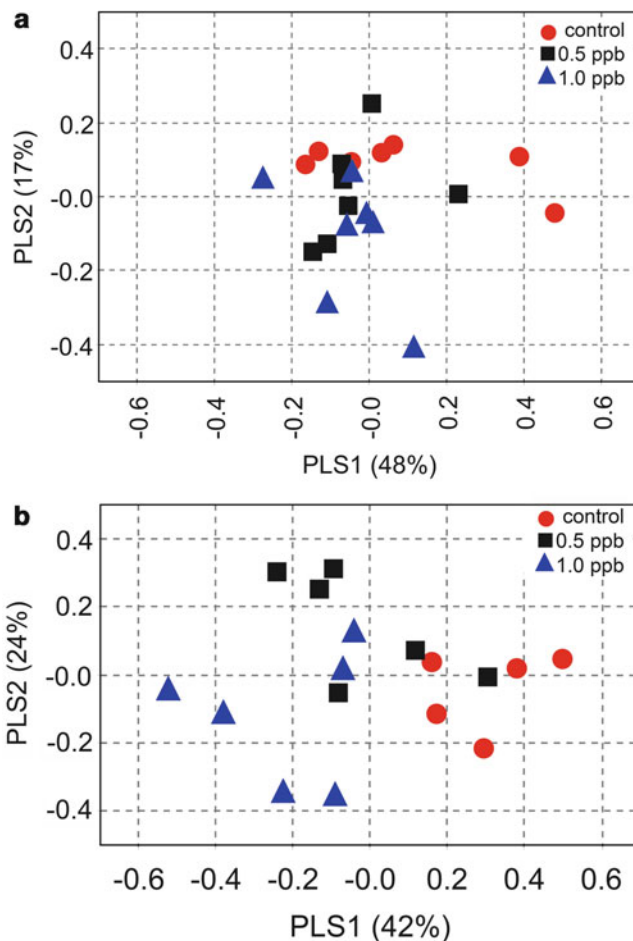
Fig. 9.33 Two-component scores plot from a PLS-DA model built with the same data as the PCA model (Fig. 9.31) from ^{1}H NMR spectra of the aqueous extracts of (**a**) MDA-MD-231 and (**b**) MCF-7, excluding the control outlier. The distribution of the scores is very similar to that of PCA model

9.8   Metabolomics of Live Cell

In mammalian cells, glucose and glutamine, as major nutrients in plasma, account for most carbon and nitrogen metabolism. According to the Warburg effect (Warburg et al. 1924; Warburg 1956), the metabolic hallmark of tumor cells is the rapid glucose consumption and lactate production to meet the requirements of rapid synthesis of biomacromolecules including nucleotides, proteins, and lipids. Dynamic ^{13}C NMR has been used in the real-time observation of metabolic fluxes in glucose and glutamine metabolism of live cells. The most comprehensive
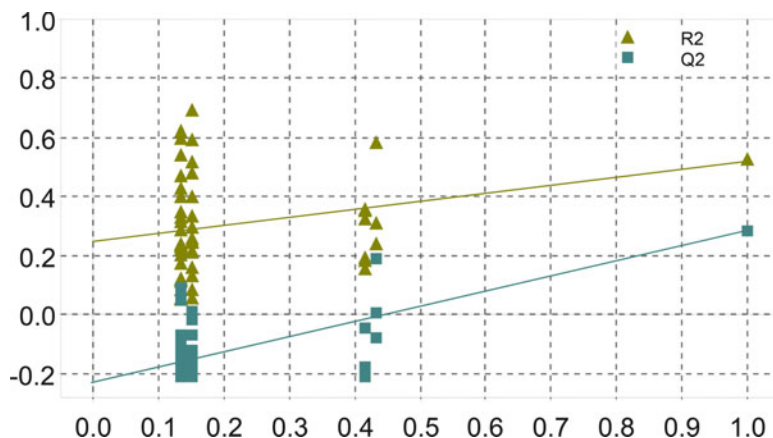
Fig. 9.34 Validation plot of the PLS-DA model generated from polar extracts of MCF-7 after the treatment with EE2 for 48 h. The scale of the y axis represents values of R^2 or Q^2. The x axis reflects the extent of the permutations with 1.0 representing the case that no class label is permuted and 0.0 meaning that all class labels are permuted. The model was validated with 50 permutations. The results indicated that the PLS-DA model is valid because none of the rebuilt models had Q^2 values higher than the original model and a few rebuilt models had higher R^2 values with smaller Q^2, which indicated that these rebuilt models are over-fitted

research on the real-time metabolic profiling of live cells is probably the recent work done by DeBerardinis et al. (2007). The study provides the insights on how glucose and glutamine are consumed during cancer cell growth. While glucose is rapidly consumed by the tumor cells for high energy demand and lipid production, the rapid glutamine metabolism is used to support both the production of reductive power (nicotinamide adenine dinucleotide phosphate, NADPH) and anaplerosis of the TCA cycle. Anaplerosis is a process of restoring the cellular metabolic intermediate pools.

As much of the study is not easily carried out, here the attempt is only to outline some of the results of the live cell study. Shown in Fig. 9.35 is a typical instrument setup for dynamic NMR experiments of live cells. Cells grown in microcarriers are loaded in an NMR tube and maintained by circulating the culture medium using a peristaltic pump. After standard setup and calibration is done (tuning, shimming, and 90° pulse calibration), culture medium is switched to the medium containing ^{13}C-labeled substances (^{13}C glucose and/or ^{13}C glutamine). To study the glucose metabolism, the live SF188 glioblastoma cells were maintained inside the NMR magnet by feeding culture medium with constant oxygenation, while a series of 1D NOE-enhanced ^{13}C NMR experiments were acquired at the rate of 18 min/experiment, the cells were then given a bolus of medium with 10 mM $1,6\text{-}^{13}C_2$-glucose. After 2.5 h, 4 mM $1,6\text{-}^{13}C_2$-glucose medium was continuously flowed through the NMR tube.

The real-time ^{13}C NMR spectra (Fig. 9.36) observed in the $1,6\text{-}^{13}C_2$-glucose perfusion experiment have shown that $3\text{-}^{13}C$ lactate is the most abundant metabolite

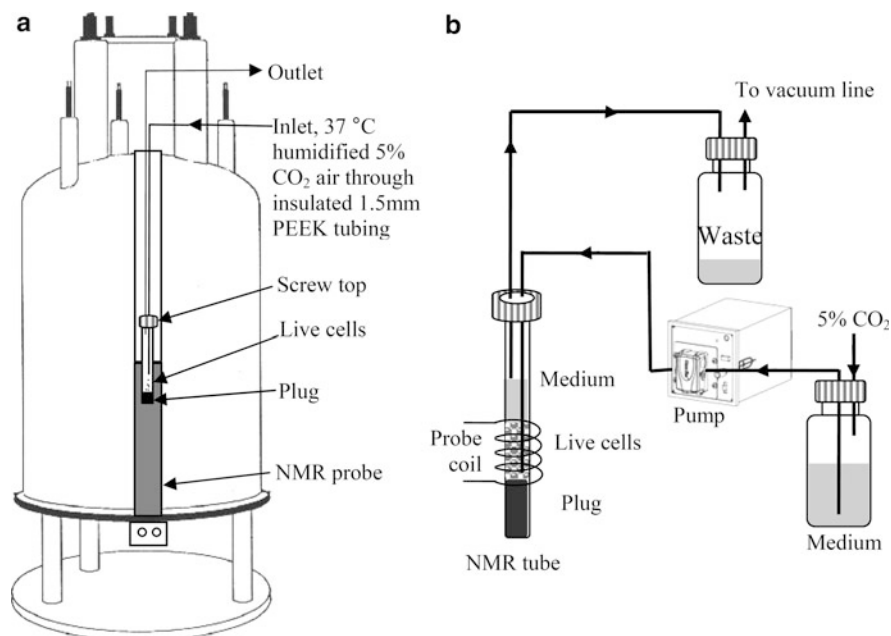**Fig. 9.35** Schematic drawing of NMR setup for live cell experiments. (**a**) The culture medium containing ^{13}C labeled nutrients is constantly flowing through the NMR tube inside the magnet. (**b**) The medium is pumped to the NMR tube containing live cells from the medium reservoir where 5 % CO_2 is maintained. The 5 % CO_2 gas can be adjusted for hypoxia or nomoxia condition

produced from the labeled glucose, which was predominantly extracellular. This suggests that the cancer cells consume glucose to meet the energy demands predominantly through rapid glycolysis producing lactate as the end product, which is consistent with the Warburg Effect. The activities of the TCA cycle were also observed by the real-time experiments, as evidenced by the increased intensity of 4-^{13}C-glutamate (Fig. 9.36b) as a function of perfusion time. The formation of 4-^{13}C-glutamate from 1,6-^{13}C$_2$-glucose is an indicative of TCA cataplerotic reactions. As shown in Fig. 9.36a, 4-^{13}C-glutamate is formed by TCA cycle intermediate α-ketoglutarate (α-KG). The cells also consumed glucose for lipid synthesis through citrate efflux of the TCA cycle (Fig. 9.36a). The graduate increase in intensity of fatty acid ^{13}C signals $(CH_2)_n$ suggests that acetyl-CoA formed by the TCA cycle was used for lipid biosynthesis by the sequential addition of acetyl-CoA. The study determined that approximately 60% of the lipogenic acetyl-CoA pool comes from glucose.

Progressive growth of tumor cells requires rapid lipid synthesis for cell proliferation. To fulfill this need, it is necessary for cells to possess two supporting pathways for maintaining sufficient reductive power (NADPH as an electron donor for lipid synthesis) and to restore oxaloacetate for the continuation of the TCA cycle during citrate export (for the production of lipogenic precursor acetyl-CoA).

Fig. 9.36 Glucose
metabolism in proliferating
glioblastoma cells (SF188).
(**a**) The ^{13}C carbons (filled
circles) of [1,6-^{13}C$_2$]glucose
distribute into various
metabolites in several major
pathways, including
glycolysis, anaerobic
pyruvate metabolism, TCA
cycle, and fatty acid
synthesis. (**b**) Stacked spectra
from a real-time [1,6-^{13}C$_2$]
glucose perfusion experiment
in SF188 cells. (**c**) An
extended spectrum from time
6 h shows the enrichment of
metabolic intermediates, after
the subtraction of background
signals from the time 0-h
spectrum (**d**) Concentration
plots of ^{13}C labeled glucose,
lactate, and alanine as the
function of the experimental
time. (reproduced with
permission from
DeBerardinis et al. (2007),
Copyright © The National
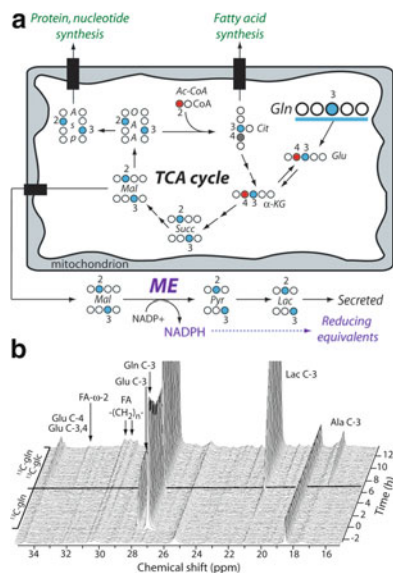Academy of Sciences of the
USA. 2007)
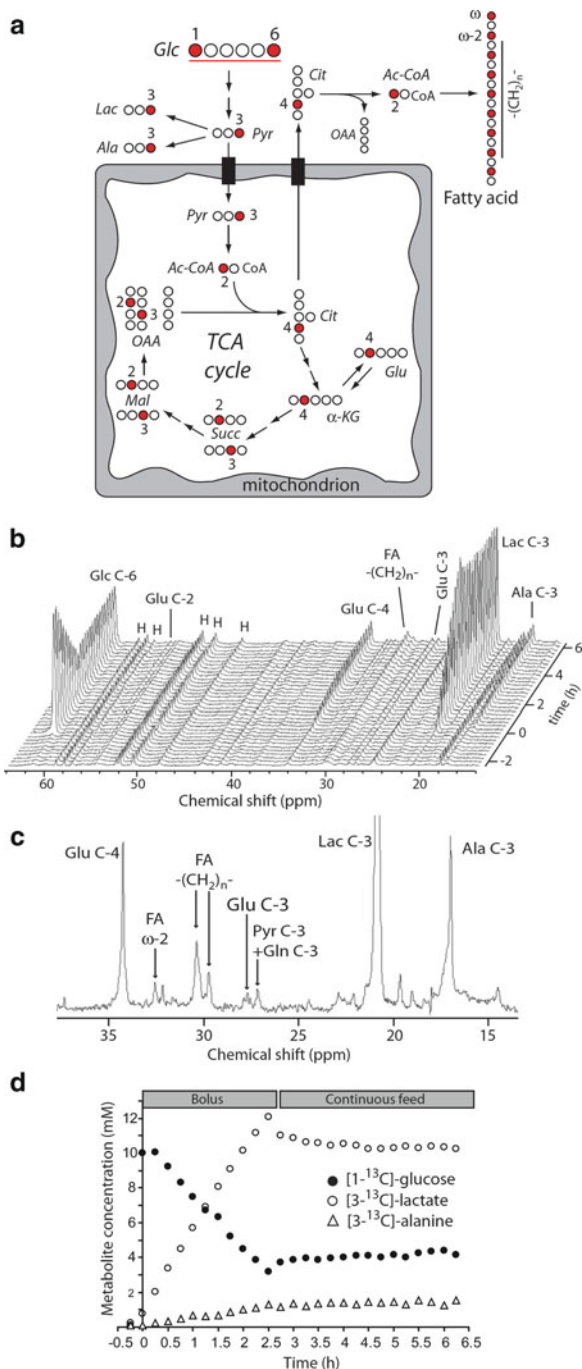
Fig. 9.37 Glutamine metabolism in proliferating glioblastoma cells (SF188). (**a**) The labeling pathway of ^{13}C from [3-^{13}C]glutamine (*blue*; at the *top right* of the mitochondrion) highlights glutaminolysis that produces NADPH by converting glutamine-derived carbon to lactate using malic enzyme (ME), [2-^{13}C]Ac-CoA (*red*) comes from [1,6-^{13}C$_2$]glucose. (**b**) Stacked spectra acquired during the two-stage perfusion experiment (reproduced with permission from DeBerardinis et al. (2007), Copyright © The National Academy of Sciences of the USA. 2007)

Similar to glucose, glutamine is metabolized by proliferating cells using a variety of pathways to support bioenergetics and biosynthesis. It was proposed that the high rate of glutamine metabolism can meet the both needs to support the lipid synthesis. A two-stage perfusion was used for investigating the activation of pathways involved in the glutamine metabolism. In the first stage, cells received medium containing 4 mM 3-^{13}C glutamine and 10 mM unlabeled glucose, while in the second stage, the medium contained both 4 mM 3-^{13}C glutamine and 10 mM 1,6-^{13}C$_2$ glucose (Fig. 9.37b). Glutamine provided a secondary source of carbon for lipid synthesis. It was determined by the real-time measurement of the (^{13}CH$_2$)$_n$ signal that the rate of ^{13}C labeling in –(CH$_2$)$_n$– during stage 2 was triple the rate in stage 1, consistent with glucose's role as the major lipogenic precursor. The study showed that glutamine contributed to approximately 25% of total lipid carbon, while glucose accounted for 50%.

During stage 1, labeled 3-^{13}C glutamate was the most abundant ^{13}C metabolite produced rapidly from 3-^{13}C glutamine (Fig. 9.37b). The immediate appearance of ^{13}C labeled aspartate in C-2 and C-3 after 3-^{13}C glutamate was formed indicated that anaplerotic oxaloacetate was derived from the glutamine pool, which suppressed pyruvate carboxylation despite substantial mitochondrial pyruvate metabolism. This depicts that some of the glutamine pool entered the TCA cycle for protein production (through Asp synthesis) and lipid synthesis (formation of Acetyl-CoA).

The other important finding of the glutamine experiment is that the cancer cells also utilize glutamine as a secondary energy source by glutaminolysis. Similar to the utilization of glucose, the cells metabolize glutamine to lactate, rather than the complete oxidation through the TCA cycle. The observed formation of ^{13}C labeled lactate at C-2 and C-3 is likely the indication of oxidation of labeled malate at C-2 and C-3 derived from the glutamine. More importantly, NADPH is produced by the malate oxidation. It was proven that this conversion of glutamine to lactate (glutaminolysis) is rapid enough to produce sufficient NADPH for fatty acid synthesis.

The study also uncovered that glutamine catabolism was accompanied by secretion of alanine and ammonia in a way that most of the amino groups from glutamine were released from the cells rather than used by other molecules. The results demonstrate that cancer cells possess a rapid glutamine consumption, which exceeds the nitrogen demand of nucleotide synthesis or maintenance of nonessential amino acid pools. Instead, the cells use glutamine metabolism as a carbon source in the same manner as to use glucose-derived carbon and TCA cycle intermediates as biosynthetic precursors.

In summary, live cell metabolomic profiling is a powerful tool to study, in real time, the important metabolic pathways such as glucose metabolism and glutamine metabolism. It can bridge the gap between high-resolution metabolite profiling of cell or tissue extracts and the low-resolution MRS studies in animal models and patients. The results of real-time metabolic profiling can provide insights on the metabolic fluxes of cancer cells.

9.9   Metabolomics Applied to Cancer Research

In recent years, metabolomics has been successfully applied in cancer research to study cancer biochemical networks and metabolic pathways. For example, metabolomics has been used to differentiate different cancer cell lines (Donato et al. 2008; Ramanathan et al. 2005), to monitor metabolic processes in cancer cells during apoptosis, to detect and prognose cancers, including liver (Yang et al. 2007a, b), leukemia (MacIntyre et al. 2010), colon (Piotto et al. 2009; Ritchie et al. 2010), breast (Oakman et al. 2011), ovarian (Odunsi et al. 2005), and prostate cancer (Raina et al. 2009). Metabolic profiles of cancerous and noncancerous tissues have been studied to characterize cell growth and death, specific tumor types, and pathologic states of tumors (Griffin and Kauppinen 2007; Griffin and Shockcor 2004; Serkova et al. 2007; Serkova and Niemann 2006). In this section, two in vivo case studies are used as examples to describe how NMR-based metabolomics can be applied in cancer research. The first example involves a quantitative characterization of the metabolic profile of a silibinin-treated mouse tumor model to assess the chemopreventive efficacy of silibinin. The second example describes the application of metabolomics to predict overall survival time of patients with metastatic colorectal cancer (CRC).

9.9.1   Silibinin Anticancer Efficacy

Silibinin (also known as silybin) has long been recognized to have inhibitory effects on tumor growth, progression, and migration (Bhatia et al. 1999; Raina et al. 2007; Hogan et al. 2007). Although the mechanism of the tumor inhibition was studied (Raina et al. 2008), the lack of information on silibinin-induced metabolic perturbations of the tumor tissue makes it difficult to assess its chemopreventive efficacy. To quantitatively characterize the metabolic profile of a silibinin-treated mouse tumor model, an NMR-based metabolomics study was conducted (Raina et al. 2009). The study aimed to specifically identify a set of biomarkers that can sensitively detect metabolic changes in the tumor tissue upon treatment with silibinin.

Four prostate tumor tissue samples were used for a control or a silibinin-treated group. The silibinin treatment was started at week 4 for 20 weeks. Approximately 0.1 g of frozen tissues were extracted and both aqueous (polar) and lipophilic (nonpolar) fractions were dried at room temperature using a speedVac. For NMR analysis, aqueous samples were reconstituted in 450 μL ^2H_2O and lipophilic samples in 600 μL deuterated chloroform/methanol mixture (2:1, v/v) and PRESAT or one-pulse experiments were conducted for aqueous or lipophilic samples, respectively. For each tissue, a total of 38 metabolites were assigned and their concentrations were calculated based on the concentration of an external TSP reference standard (0.5 mM for polar and 1.2 mM for nonpolar extracts) and subsequently normalized to the unit (wet) weight of individual tissues. It was the concentrations of the metabolites (plus 4 concentration ratios) that were used as variables for MVA, not NMR spectral bins.

A PCA model was built using the 42 concentration variables. The scores plot of the PCA model (Fig. 9.38a) clearly shows that the treated group is well separated from the control group. The shattered loadings plot of PCA model (Fig. 9.38b) shows the contributions of individual metabolites to the group discrimination described by the model. By comparing the average concentrations of treated and control group, a total of 14 biomarkers contributed to the class discrimination, including silibinin-induced elevated levels of citrate, glucose, hydrophilic choline, glycerophosphocholine (GPC), polyols, and glutathione (reduced form; GSH), as well as significant decrease in levels of lactate, cholesterol, alanine, lipophilic choline (Cho), and phosphatidylcholine (PtdCho). One of the hallmarks of prostate cancer is a decrease in citrate levels due to the increased consumption of citrate in the TCA cycle along with abnormal secretion functions and zinc metabolism (Costello and Franklin 2000, 2005). In normal prostate cells, citrate is not a functional intermediate of the TCA cycle, but an end product due to the high concentration of zinc that is involved in inhibition of the conversion of citrate into isocitrate. On the other hand, the decreased amount of zinc in prostate tumors leads to the formation of isocitrate from citrate, causing increased activities of the TCA cycle and decreased levels of citrate to meet the increased energy demand and fatty acid synthesis (Costello and Franklin 2000). Additionally, the decrease in cholesterol indicated that the increased citrate was not utilized by cholesterogenesis. Its ability to restore
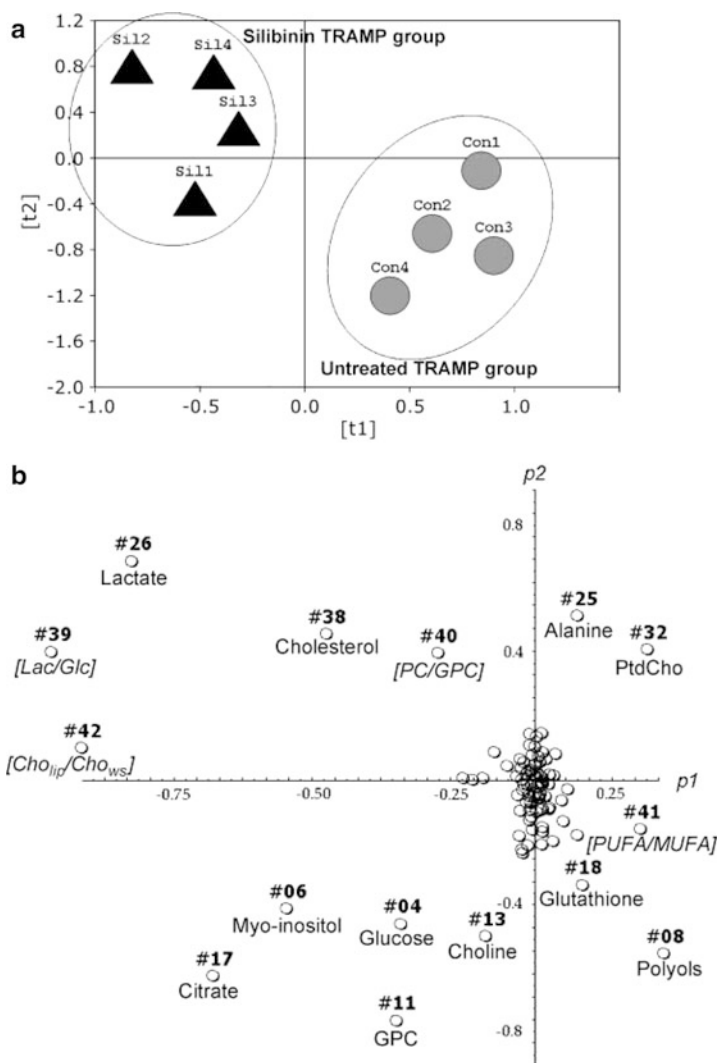
Fig. 9.38 Two-component scores (**a**) and loadings (**b**) plots of a PCA model generated from concentrations of metabolites. The concentrations of water-soluble and lipid metabolites were calculated from ^{1}H-NMR spectra of prostate tissue extracts. The scores plot clearly showed class separations between control and silibinin-treated groups. Axis labels t_1 and t_2 in (**a**) are PC1 and PC2, respectively (reproduced with permission from Raina et al. (2009), Copyright © American Association for Cancer Research, 2009)

the citrate accumulation and secretion of the prostate tissue indicates that silibinin possesses a remarkable chemopreventive efficacy.

Silibinin induced a significant decrease in PtdCho level (a major membrane phospholipid) and lipophilic fraction of choline (a precursor of PtdCho) compared to the untreated controls, which are intermediates of choline phospholipid

metabolism. The decrease in levels of both metabolites has been observed in the malignant tissues (Glunde and Serkova 2006; Glunde et al. 2006). Other bioindicators of silibinin chemopreventive efficacy includes the significant increase in concentrations of polyols (decreased during prostate tumor progression) and glutathione (an antioxidant), meaning that silibinin exhibited inhibitory effect on progression of prostate cancer.

In summary, the concentrations of metabolites in tissue extracts of a mouse prostate tumor model were used to generate a PCA model that showed the silibinin-induced metabolic changes of the treated tumors compared to the untreated controls. The metabolic changes related to silibinin chemopreventive efficacy against prostate cancer include the restoration of normal citrate concentrations, decrease in membrane phospholipid synthesis, and increase in polyols and antioxidants.

9.9.2   Metabolomic Profiling of Colon Cancer

Colorectal cancer (CRC, or colon cancer) is the third leading cause of cancer-related deaths in the United States. Most of CRC symptoms may not be noticeable until the progression stage in which many patients have metastases and the tumor has spread to other organs. Therefore, early detection of CRC, especially for patients with metastatic CRC (mCRC), can significantly improve the treatment and survival outcomes. An NMR-based metabolomics study (Bertini et al. 2012) was conducted to determine whether there is a significant difference in the metabolomic profiles, which can be used to predict overall survival (OS) time.

Serum samples from 139 healthy individuals and 153 patients with mCRC (resistant to 5-FU, oxaliplatin, and irinotecan) were used to determine the metabolic signature of mCRC. The samples were collected and allowed to clot at room temperature for 30–120 min. For NMR analysis, serum was mixed with an equal volume of 70 mM PB containing 6.15 mM NaN_3 and 6.64 mM TSP at pH 7.4. CMPG spectra were acquired for metabolite profiling and J-resolved (JRES) spectra were used to identify the metabolites. The residual water resonance of 4.2–5.0 ppm was excluded and metabolite resonances were aligned. The spectral region of 0.5–8.5 ppm was then segmented into 0.02 ppm bins and normalized over the total spectral intensity of the region.

The whole dataset (all binned spectra) was split into a training set with $n = 141$ (96 healthy individuals and 45 patients) for generating a PLS model, and a validation set with $n = 151$ (43 healthy individuals and 108 patients) for assessing the prediction accuracy of the model. The PLS model was generated and subsequently validated by a tenfold cross validation. First, all data of the training set were partitioned into 10 subsets of samples, of which 9 subsets were used to produce a PLS model and the remaining subset was used as a test set to validate the model. The model generation and cross validation were repeated nine more times (tenfold) using each of the 10 subsets as the test set. The random partition was repeated 100 times, resulting in a total of 1,000 random subsamplings (tenfold cross validation is repeated 100 times). The results (scores and loadings)
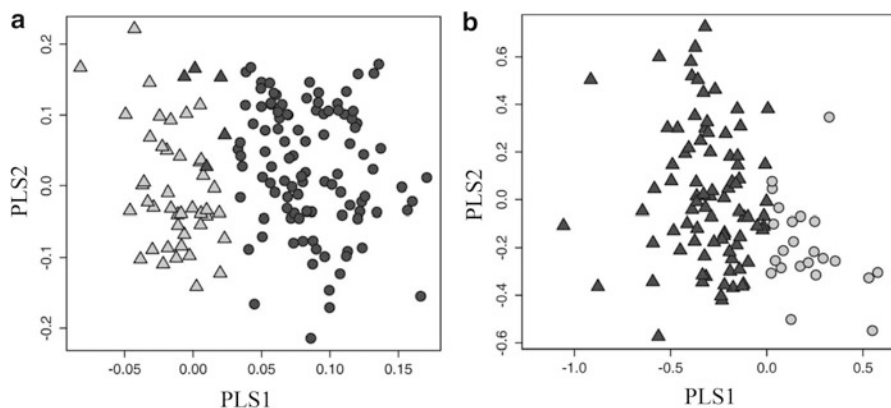
Fig. 9.39 Two-component PLS scores plots. (**a**) The score plot for the healthy subjects (*triangles*) and patients with mCRC (*circles*) of the validation set based on the PLS model built on the NMR spectra of the training set (see text for definition of training and validation set). *Triangles* are predicted as healthy subjects and circles as mCRC. (**b**) The score plot for the long OS group (*triangles*) and short OS group (*circles*) of the validation set based on the model built on the NMR spectra of the training set (reproduced with permission Bertini et al. (2012), Copyright © American Association for Cancer Research, 2012)

for the training set were then averaged to obtain the final model. A validated PLS model was then used to predict classifications of samples using the validation set. The results are shown in Fig. 9.39a. The model clearly shows a good separation along PLS1 (the first component) between the healthy and patient group of the validation set, meaning that the spectral patterns (or metabolite profiles) of serum from the patients with mCRC have characteristic features compared to those of healthy individuals. Note that the validation set in this study has a different meaning than the commonly defined test set, which was not used to generate or cross validate the PLS models.

One of the main goals of the study was to determine whether NMR metabolic profiles of serum from mCRC patients can be related to OS time. To test the prognostic classification using serum metabolic profiles, a model was generated using 20 samples from mCRC patients of the training set, of which 10 patients had the shortest OS and the remaining 10 had the longest OS. After the model was validated with cross validation (1,000 times of permutations), the model was used to predict OS group classification of mCRC patients in the validation set. The results (the scores plot, Fig. 9.39b) showed two slightly overlapped clusters of long and short OS group. The model assigned 85 of 108 mCRC patients in the validation set to the long OS class and 23 patients to the short OS class, with mean OS values of 12.3 and 5.2 months, respectively. This suggested that the model has reasonable accuracy for predicting OS time of mCRC patients.

Characterization of metabolites related to the mCRC serum profiles was conducted using the loadings plot generated with all sets of healthy individuals and patients (training set and validation set). After the NMR resonances in the
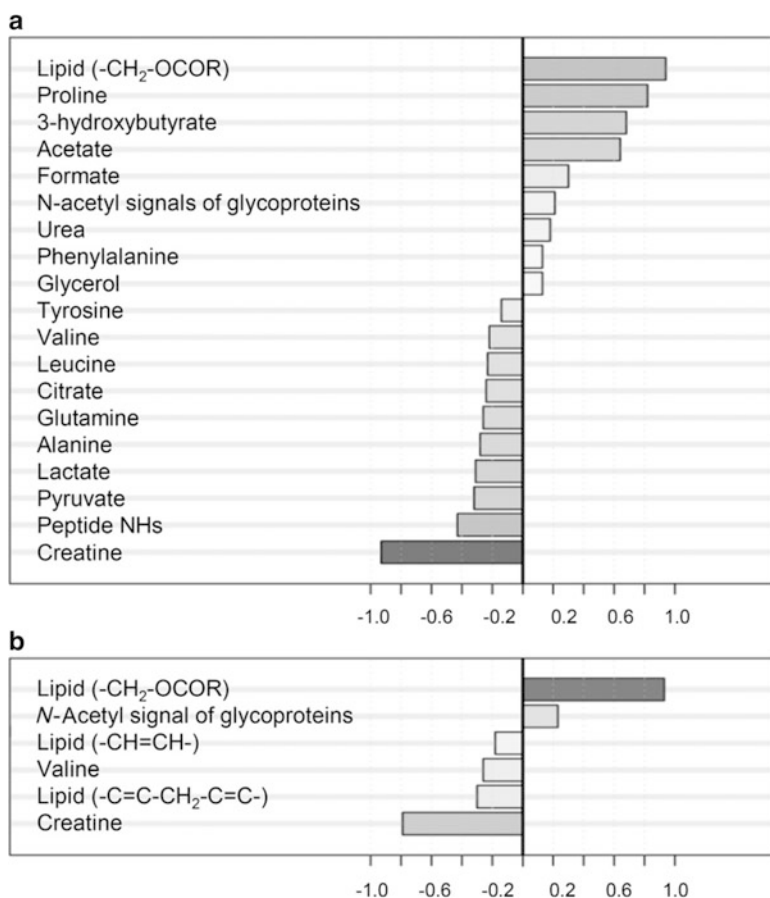
Fig. 9.40 Values of ln(FC) obtained (**a**) by comparing healthy subjects to patients with mCRC, and (**b**) by comparing patients with mCRC and long OS (>24 months) to short OS (<3 months). FC represents fold change. (**a**) ln(FC) with a positive value indicates an increased serum concentration in the patients with mCRC, whereas a negative value means a decreased serum concentration in the patients with mCRC. (**b**) ln(FC) with a positive value indicates an increased serum concentration present in the patients with mCRC and short survival (<3 months), whereas a negative value means a decreased serum concentration in the patients with mCRC and short survival (<3 months) (reproduced with permission Bertini et al. (2012), Copyright © American Association for Cancer Research, 2012)

loadings plot were identified, the concentration changes of the metabolites were estimated by comparing the average spectral integrates of the metabolites in serum samples of patients with those of healthy individuals. The analysis showed that mCRC patients are characterized by lower serum levels of creatine, peptide NHs, pyruvate, lactate, alanine, glutamine, citrate, leucine, valine, and tyrosine, and higher levels of lipid ($-CH_2-OCOR$), proline, 3-hydroxybutyrate, acetate, formate, N-acetyl groups of glycoproteins, phenylalanine, and glycerol (Fig. 9.40a).

Comparison of the average spectra of serum samples from patients with an OS shorter than 3 months to those longer than 24 months revealed that patients with a short OS have lower serum levels of creatine, valine, and lipid ($-C=C-CH_2-C=C-$ and $-CH=CH-$) and higher levels of lipid ($-CH_2-OCOR$) and N-acetyl groups of glycoproteins (Fig. 9.40b). It was suggested that elevated levels of lipid $-CH_2-OCOR$ and N-acetyl groups of glycoproteins reflected a nonspecific inflammatory response and the effect is more severe for the patients with shorter OS (Torri et al. 1999). These results were consistent with the previous finding that the OS is primarily governed by the state of the local adaptive immune response (Pages et al. 2009; Galon et al. 2006).

Lactate, pyruvate, alanine, and glutamine are glucose precursors of gluconeogenesis—a metabolic pathway that generates glucose from non-carbohydrate carbon substrates. Decreased levels of the glucose precursors in the patients, coupled with no significant change in glucose concentration, could be an indicator that an increased gluconeogenesis caused an increased hepatic uptake of the glucose precursors (Leij-Halfwerk et al. 2000; Shearer et al. 2010; Holroyde et al. 1975). The increased energy demand for gluconeogenesis was fulfilled via fatty acid oxidation as evidenced by the decreased level of fatty acids along with an accumulation of 3-hydroxybutyrate—a product of fatty acid oxidation (Tiziani et al. 2009).

In summary, metabolomic profiles of serum samples from mCRC patients revealed an inflammatory state and altered energy metabolism (increased gluconeogenesis and fatty acid oxidation), compared to healthy individuals. NMR-based metabolomics can be applied to predict OS time of mCRC patients with a sufficient accuracy.

References

Axelrod J, Daly J (1965) Pituitary gland: enzymic formation of methanol from S-adenosyl-methionine. Science 150:892–893

Baxter MJ (1995) Standardization and transformation in principal component analysis, with applications to archaeometry. Appl Statist 44:513–527

Beloueche-Babari M, Chung YL, Al-Saffar NM, Falck-Miniotis M, Leach MO (2010) Metabolic assessment of the action of targeted cancer therapeutics using magnetic resonance spectroscopy. BrJ Canc 102:1–7

Bernini P, Bertini I, Luchinat C, Nincheri P, Staderini S, Turano P (2011) Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. J Biomol NMR 49:231–243

Bertini I, Cacciatore S, Jensen BV, Schou JV, Johansen JS, Kruhøffer M, Luchinat C, Nielsen D, Turano P (2012) Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. Cancer Res 72:356–364

Bhatia N, Zhao J, Wolf DM, Agarwal R (1999) Inhibition of human carcinoma cell growth and DNA synthesis by silibinin, an active constituent of milk thistle: comparison with silymarin. Cancer Lett 147:77–84

Britten RJ, McClure FT (1962) The amino acid pool in Escherichia coli. Bacteriol Rev 26:292–335

Costello LC, Franklin RB (2005) 'Why do tumour cells glycolyse?': from glycolysis through citrate to lipogenesis. Mol Cell Biochem 280:1–8

Costello LC, Franklin RB (2000) The intermediary metabolism of the prostate: a key to under-standing the pathogenesis and progression of prostate malignancy. Oncology 59:269–282

Dardik R, Varon D, Tamarin I, Zivelin A, Salomon O, Shenkman B, Savion N (2000) Homocys-teine and oxidized low density lipoprotein enhance platelet adhesion to endothelial cells under flow conditions: distinct mechanisms of thrombogenic modulation. Thromb Haemost 83:338–344

Davis JM (ed) (2002) Basic cell culture, practical approach series, 2nd edn. Oxford University Press, USA

DeBerardinis RJ, Mancuso A, Daikhin E, Nissim I, Yudkoff M, Wehrli S, Thompson CB (2007) Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis. PNAS 104:19345–19350

de Koning W, van Dam K (1992) A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. Anal Biochem 204:118–123

Deprez S, Sweatman BC, Connor SC, Haselden JN, Waterfield CJ (2002) 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 30:1297–1310

Donato MT, Lahoz A, Castell JV, Gómez-Lechón MJ (2008) Cell lines: a tool for in vitro drug metabolism studies. Curr Drug Metab 9:1–11

Ekman DR, Teng Q, Villeneuve DL et al (2008) Investigating compensation and recovery of fathead minnow (Pimephales promelas) exposed to 17alpha-ethynylestradiol with metabolite profiling. Environ Sci Technol 42:4188–4194

Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2006) Multi- and megavariate data analysis. Umetrics Academy, Sweden

Fan TWM (1996) Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures. Prog Nucl Magn Reson Spectr 28:161–219

Fiehn O (2001) Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp Funct Genom 2:155–168

Freshney RI (2005) Culture of animal cells: a manual of basic technique, 5th edn. Wiley-Liss, Hoboken, NJ

Freshney RI (2010) Culture of animal cells: a manual of basic technique and specialized applications, 6th edn. Wiley-Blackwell, Hoboken, NJ

Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C et al (2006) Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. Science 313:1960–1964

Gatley SJ, Sherratt HS (1977) The synthesis of hippurate from benzoate and glycine by rat liver mitochondria. submitochondrial localization and kinetics. Biochem J 166:39–47

Gates SC, Sweeley CC (1978) Quantitative metabolic profiling based on gas chromatography. Clin Chem 24:1663–1673

Glunde K, Serkova NJ (2006) Therapeutic targets and biomarkers identified in cancer choline phospholipid metabolism. Pharmacogenomics 7:1109–1123

Glunde K, Jacobs MA, Bhujwalla ZM (2006) Choline metabolism in cancer: implications for diagnosis and therapy. Expert Rev Mol Diagn 6:821–829

Goodacre R (2005) Metabolomics – the way forward. Metabolomics 1:1–2

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22:245–252

Green KA, Carroll JS (2007) Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. Nat Rev Cancer 7:713–722

Griffin JL, Shockcor JP (2004) Metabolic profiles of cancer cells. Nat Rev Cancer 4:551–561

Griffin JL, Kauppinen RA (2007) Tumour metabolomics in animal models of human cancer. J Proteome Res 6:498–505

Hogan FS, Krishnegowda NK, Mikhailova M, Kahlenberg MS (2007) Flavonoid, silibinin, inhibits proliferation and promotes cell-cycle arrest of human colon cancer. J Surgical Res 143:58–65

Holroyde CP, Gabuzda TG, Putnam RC, Paul P, Reichard GA (1975) Altered glucose metabolism in metastatic carcinoma. Cancer Res 35:3710–3714

Homocysteine Studies Collaboration (2002) Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. JAMA 288:2015–2022

Horning EC, Horning MG (1971) Metabolic profiles: gas-phase methods for analysis of metabolites. Clin Chem 17:802–809

Hotelling H (1931) The generalization of student's ratio. Ann Math Stat 2:360–378

Howard VJ, Sides EG, Newman GC, Cohen SN, Howard G, Malinow MR, Toole JF et al (2002) Changes in plasma homocyst(e)ine in the acute phase after stroke. Stroke 33:473–478

Jackson JE (1991) A user's guide to principal components. Wiley, New York

Jung JY, Lee HS, Kang DG, Kim NS, Cha MH, Bang OS, Ryu DH, Hwang GS (2011) 1H-NMR-based metabolomics study of cerebral infarction. Stroke 42:1282–1288

Keenan MR, Kotula PG (2004) Accounting for poisson noise in the multivariate analysis of ToF-SIMS spectrum images. Surf Int Anal 36:203–212

Khoo SHG, Al-Rubeai M (2007) Metabolomics. Al-Rubeai M, Fussenegger M (eds) Cell engineering vol. 5. system biology. Springer

Kurien BT, Everds NE, Scofield RH (2004) Experimental animal urine collection: a review. Lab Anim. 38:333–361

Lane AN, Fan TWM (2007) Quantification and identification of isotopomer distributions of metabolites in crude cell extracts using 1H TOCSY. Metabolomics 3:79–86

Leij-Halfwerk S, Dagnelie PC, van Den Berg JWO, Wattimena JDL, Hordijk-Luijk CH, Wilson JHP (2000) Weight loss and elevated gluconeogenesis from alanine in lung cancer patients. Am J Clin Nutr 71:583–589

Lindon JC, Nicholson JK, Everett JR (1999) NMR spectroscopy of biofluids. Ann Rev NMR Spectr 38:1–88

MacIntyre DA, Jimenez B, Lewintre EJ, Martin CR, Schafer H, Ballesteros CG et al (2010) Serum metabolome analysis by 1H-NMR reveals differences between chronic lymphocytic leukaemia molecular subgroups. Leukemia 24:788–797

Miller JW, Nadeau MR, Smith D, Selhub J (1994) Vitamin B-6 deficiency vs folate deficiency: comparison of responses to methionine loading in rats. Am J Clin Nutr 59:1033–1039

Moestue S, Sitter B, Bathen TF, Tessem MB, Gribbestad IS (2011) HR MAS MR spectroscopy in metabolic characterization of cancer. Curr Top Med Chem 11:2–26

Nicholson JK, Foxall PJ, Spraul M, Farrant RD, Lindon JC (1995) 750 MHz 1H and 1H-13C NMR spectroscopy of human blood plasma. Anal Chem 67:793–811

Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29:1181–1189

Oakman C, Tenori L, Biganzoli L, Santarpia L, Cappadona S, Luchinat C, Di Leo A (2011) Uncovering the metabolomic fingerprint of breast cancer. Int J Biochem Cell Biol 43:1010–1020

Odunsi K, Wollman RM, Ambrosone CB, Hutson A, Mccann SE, Tammela J et al (2005) Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. Int J Cancer 113:782–788

Oliver SG, Winson MK, Kell DB, Baganz B (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16:373–378

Pages F, Kirilovsky A, Mlecnik B, Asslaber M, Tosolini M, Bindea G et al (2009) In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. J Clin Oncol 27:5944–5951

Patt SL, Sykes BD (1972) Water eliminated Fourier transform NMR spectroscopy. Chem Phys 56:3182

Pearson K (1901) (1901) On lines and planes of closest fit to systems of points in space. Phil Mag 2:559–572

Piotto M, Moussallieh FM, Dillmann B, Imperiale A, Neuville A, Brigand C et al (2009) Metabolic characterization of primary human colorectal cancers using high resolution magic angle spinning 1H magnetic resonance spectroscopy. Metabolomics 5:292–301

Raina K, Blouin MJ, Singh RP et al (2007) Dietary feeding of silibinin inhibits prostate tumor growth and progression in transgenic adenocarcinoma of the mouse prostate model. Cancer Res 67:11083–11091

Raina K, Serkova NJ, Agarwal R (2009) Silibinin feeding alters the metabolic profile in TRAMP prostatic tumors: 1H-NMRS–based metabolomics study. Cancer Res 69:3731–3735

Raina K, Rajamanickam S, Singh RP et al (2008) Stage-specific inhibitory effects and associated mechanisms of silibinin on tumor progression and metastasis in transgenic adenocarcinoma of the mouse prostate model. Cancer Res 68:6822–6830

Ramanathan A, Wang C, Schreiber SL (2005) Perturbational profiling of a cell-line model of tumorigenesis by using metabolic measurements. PNAS 102:5992–5997

Rencher AC (2002) Methods of multivariate analysis. Wiley, New York

Ritchie SA, Ahiahonu PWK, Jayasinghe D, Heath D, Liu J, Lu YS et al (2010) Reduced levels of hydroxylated, polyunsaturated ultra long-chain fatty acids in the serum of colorectal cancer patients: implications for early screening and detection. BMC Med 8:13–32

Serkova NJ, Spratlin JL, Eckhardt SG (2007) NMR-based metabolomics: translational application and treatment of cancer. Curr Opin Mol Ther 9:572–585

Serkova NJ, Niemann CU (2006) Pattern recognition and biomarker validation using quantitative 1H-NMR-based metabolomics. Expert Rev Mol Diagn 6:717–731

Shearer JD, Buzby GP, Mullen JL, Miller E, Caldweell MD (1984) Alteration in pyruvate metabolism in the liver of tumor-bearing rats. Cancer Res 44:4443–4446

Simpson AJ, Brown SA (2005) Purge NMR: effective and easy solvent suppression. J Magn Reson 175:340–346

Smeaton JR, Elliott WH (1967) Selective release of ribonuclease-inhibitor from Bacillus subtilis cells by cold shock treatment. Biochem Biophys Res Commun 26:75–81

Teng Q, Ekman DR, Huang W, Collette TW (2012) Push-through direct injection NMR: an optimized automation method applied to metabolomics. Analyst 137:2226–2232

Teng Q, Huang W, Collette TW, Ekman DR, Tan C (2009) A direct cell quenching method for cell-culture based metabolomics. Metabolomics 5:199–208

Tiziani S, Lopes V, Gunther UL (2009) Early stage diagnosis of oral cancer using 1H NMR-based metabolomics. Neoplasia 11:269–276

Torri GM, Torri J, Gulian JM, Vion-Dury J, Viout P, Cozzone PJ (1999) Magnetic resonance spectroscopy of serum and acute-phase proteins revisited: a multiparametric statistical analysis of metabolite variations in inflammatory, infectious and miscellaneous diseases. Clin Chim Acta 279:77–96

Ulrich EL, Akutsu H, Doreleijers JF et al (2008) BioMagResBank. Nucleic Acids Res 36: D402–408

van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. Genomics 7:142–156

Van der Greef J, Smilde AK (2005) Symbiosis of chemometrics and metabolomics: past, present, and future. J Chemomet 19:376–386

Viant M (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Commun 310:943–948

Viant MR (2007) Revealing the metabolome of animal tissues using 1H nuclear magnetic resonance spectroscopy. In: Methods in molecular biology, Weckwerth W (ed). Clifton, NJ, Humana Press, 358:229–246

Villas-Boas SG, Nielsen J, Smedsgaard J et al (2007) Metabolome Analysis: An introduction. Wiley, Hoboken, NJ

Warburg O, Posener K, Negelei E (1924/1930) Ueber den Stoffwechsel der Tumoren. Biochem Z (German) 152:319–344 (Reprinted in Warburg O. On metabolism of tumors. Constable, London)

Warburg O (1956) On respiratory impairment in cancer cells. Science 124:269–270

Williams R (1956) Biochemical Individuality, the basics for the genetotrophic concept. Univ of Texas Press, Austin

Wishart DS, Tzur D, Knox C et al (2007) HMDB: the human metabolome database. Nucleic Acids Res 35:D521–D526

Wold S, Ruhe A, Wold H, Dunn WJ III (1984) The collinearity problem in linear regression. the partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 5:735–743

Wold S, Antti H, Lindgren F, Ohman J (1998) Orthogonal signal correction of near-infrared spectra. Chemometr Intell Lab Syst 44:175–185

Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometr Intell Lab Syst 2:37–52

Yang C, Richardson AD, Smith JW, Osterman A (2007a) Comparative metabolomics of breast cancer. Pac Symp Biocomput 12:181–192

Yang YX, Li CL, Nie X, Feng X, Chen W, Yue Y et al (2007b) Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning 1H NMR spectroscopy in conjunction with multivariate data analysis. J Proteome Res 6:2605–2614

Kurien BT, Everds NE, Scofield RH (2004) Experimental animal urine collection: a review. Lab Anim 38:333–361