UNIVERSITY OF NAPOLI - FEDERICO II

FACULTY OF MATHEMATICAL, PHYSICAL
AND NATURAL SCIENCES

PhD THESIS

to obtain the title of

PhD IN CHEMISTRY

DEVELOPMENT AND APPLICATION OF ADVANCED NMR TECHNIQUES FOR BIOMARKERS IDENTIFICATION

Defended by

DEBORA PARIS

NOVEMBER 2006 - DECEMBER 2009

Tutors: Prof. Rosa Lanzetta
"Federico II" University
Napoli, Italy

Dr. Andrea Motta
Institute of Biomolecular Chemistry
(ICB-CNR), Pozzuoli (NA), Italy

Coordinator: Prof. Aldo Vitagliano
"Federico II" University
Napoli, Italy

Contents

Abbreviations

Ala	Alanine
APF	Alpha-fetoprotein
BMI	Body Mass Index
CAKE	Monte *CA*rlo pea*K* volume *E*stimation
CIR	Cirrhosis
COPD	Chronic Obstructive Pulmonary Disease
EBC	Exhaled Breath Condensate
FEV	Forced Expiratory Volume
FVC	Forced Vital Capacity
Gln	Glutamine
Glu	Glutamate
Gly	Glycine
GOLD	Global Initiative for Chronic Obstructive Lung Disease
GPC	Glycerophosphocholine
HCC	Human Hepatocellular Carcinoma
HCV	Hepatitis C Virus
HMQC	Heteronuclear Multiple Quantum Correlation
HS	Healthy Subject
HSQC	Heteronuclear Single Quantum Correlation
Leu	Leucine
MET-CRC	Metastasis from ColoRectal Carcinoma
MRI	Magnetic Resonance Imaging
MVA	MultiVariate data Analysis
NMR	Nuclear Magnetic Resonance
NOESY	Nuclear Overhauser Effect SpectroscopY
NT	Non Tumoral or Normal Tissue
O2PLS-DA	Orthogonal Projection to Latent Structures Discriminant Analysis
PBS	Phosphate Buffer Saline
PC	Phosphocholine
PC$_S$	Principal Components
PCA	Principal Component Analysis
PCR	Principle Component Regression
PE	Phosphoryl-ethanolamine
PLS	Projection to Latent Structures
PLS-DA	Projection to Latent Structures Discriminant Analysis
RF	Radio Frequency
SOFAST	Selective Optimized Flip Angle Short Transient
SUS-plot	Shared and Unique Structure plot
Thr	Threonine
TMAO	Trimethylamine-N-oxide
TOCSY	TOtal Correlation SpectroscopY
TSP	trimethylsilylpropionate

Original publications

This thesis is based on the following publications and submitted papers:

1. D. Paris, D. Melck, M. Stocchero, O. D'Apolito, R. Calemma, G. Castello, F. Izzo, G. Palmieri, G. Corso, A. Motta. Monitoring liver alteration during hepatic tumorigenesis by NMR profiling and pattern recognition. Submitted to Metabolomics.

2. G. de Laurentiis, D. Paris, D. Melck, M. Maniscalco, S. Marsico, G. Corso, A. Motta and M. Sofia. Metabonomic analysis of exhaled breath condensate in adults by Nuclear Magnetic Resonance spectroscopy. Eur Respir J 2008; 32: 1-9.

3. R. Romano, D. Paris, F. Acernese, F. Barone, A. Motta, Fractional volume integration in two-dimensional NMR spectra: CAKE, a Monte Carlo approach. J Magn Res 192 (2008) 294-301.

4. A. Motta, D. Paris, G. Andreotti, D. Melck. Monitoring real-time metabolism of living cells by fast two-dimensional NMR spectroscopy. Submitted to Analitical Chemistry.

Introduction

The large amount of data derived from genomics and proteomics, aiming at elucidating biochemical mechanism, has often revealed the complexity of cellular regulation. Therefore, metabolic studies are increasingly contributing to gene function analysis, and an increased interest in metabolites as biomarkers for disease progression or response to natural or external intervention is also growing.

Nuclear Magnetic Resonance (NMR) spectroscopy has emerged as a key tool for understanding metabolic processes in living systems. Recently, a new approach to elucidate metabolism and its mechanisms has been put forward. It is metabonomics: an analysis based on a minimum number of assumptions on the biochemical processes that occur in a living system, mainly investigated by advanced spectroscopic techniques including mass spectrometry and NMR spectroscopy.

Metabonomics is formally defined as "the quantitative measurement of the multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modification" [1]. It has been coined to describe the combined application of spectroscopy and multivariate statistical approaches to investigate of the multicomponent composition of biofluids, cells and tissues. In particular, NMR-based metabonomics has proven to be particularly suited for the rapid analysis of complex biological samples. Indeed, the so generated NMR spectral results yield a unique metabolic fingerprint for each complex biological mixture. According, if the status of a given organism changes, such as in a disease state or following exposure to a drug, the unique metabolic fingerprint or signature reflects this change, thus supplying relevant biochemical indications.

Multivariate statistical methods provide an expert means of analyzing and maximizing information recovery from complex NMR spectral data. Detailed inspection of NMR spectra and integration of individual peaks can give valuable information on dominant biochemical changes. However, subtle variation in spectra may be overlooked and it is difficult to envisage general effects as a function of both dose and time in a large cohort of samples with biological variability. Pattern recognition methods can be used to map the NMR spectra into a lower dimensional space (than that implied by the number of points in the digital representation of the NMR spectrum) such that any clustering of the samples based on similarities of biochemical profiles can easily be determined and the biochemical basis elucidated.

The development of new spectroscopic tools for high thoughput analysis of selected biochemical pathways is crucial for metabolome investigations. The purpose of the present thesis is to explore the recent NMR improvements by applying and developing new metabolomic strategies for biomarkers discovery, including NMR data handling, peaks quantification and fast data acquisition.

In the first chapter, a general overview of the multivariate data analysis and pattern recognition methods is given. In particular, we highlighted the advantages of using those tools to NMR data for biomarkers investigations. The most common regression methods (Principal Components Analysis and Projection to Latent Structures) and plot visualization (scatters scores plots and loadings plots) are described to supply the reader with the basic statistical tools for a better understanding of the application the biological issues reported in the last section. NMR and regression techniques were applied to different patient classes to discriminate a) hepatic tissues and b) exhaled breath condensates belonging to patients with different pathological states.

In the second chapter we describe a new integration method developed for two-dimensional NMR spectra quantification. Indeed, one-dimensional NMR spectra are often too complex for interpretation and metabolite identification as most of the signals overlap heavily. By introducing an additional dimension, peaks are spread and spectra are simplified. Quantitative information from multidimensional NMR experiments can be obtained by peak volume integration. The standard procedure (selection of a region around the chosen peak and addition of all values) is often biased by poor peak definition because of peak overlap. In this chapter we reported a simple method, called CAKE, for volume integration of moderately to strongly overlapping peaks. Starting from the peak line shapes in two-dimensional NMR, we describe how the CAKE routine was constructed using the Monte Carlo Hit-or-Miss techniques and some simple mathematical relationships.

The third chapter is a general introduction to fast NMR two-dimensional spectroscopy. In particular, we describe the details of the so-called SO-FAST-HMQC pulse sequence [2, 3] we would like to apply to investigate in cell metabolism. The SOFAST-HMQC sequence was created and designed by Shanda and Brutscher and co-workers for proteins as it is based upon very short experimental recycle delays, which, of course, must rely on short T_1 relaxations time. At a first sight, this is an evident drawback since metabolites are often characterized by T_1 relaxations time longer than those of proteins. However, as detailed in Chapter 6, we have applied the SO-FAST experiment to the diatom *T. rotula* cells obtaining, to the best of our knowledge, the first application of fast NMR spectroscopy to ^{1}H-^{15}N metabolic profiling directly on living cells.

The fourth chapter reports the metabolic characterization of: a) the progressive liver alterations during tumorigenesis and b) the exhaled breath condensate of patients with airway diseases. We describe the multivariate data analysis and pattern recognition methods starting from NMR spectra of liver tissues extracts and exhaled breath condensates. a) Samples were collected and grouped in four classes: hepatocellular carcinoma (HCC) developed on hepatitis C cirrhosis (CIR), the cirrhotic adjacent HCC tissue, liver metastasis from colorectal carcinoma (MET-CRC), and the related adjacent "normal" tissue considered as control. The results indicate that the lactate/glucose ratio is able to characterize and distinguish the analyzed subsets of hepatic samples. In particular, we identified a statistical model that could be used to distinguish hepatic metastasis and human hepatocarcinoma from a "normal" (healthy) hepatic tissue. b) Exhaled breath condensates (EBC) and paired salivas were collected from healthy subjects, laryngectomized and chronic obstructive pulmonary disease (COPD) patients. The results showed that all NMR saliva spectra were significantly different from corresponding EBC samples, which assessed no saliva contamination in EBC samples. Indeed, EBC taken from condensers washed with recommended procedures invariably showed spectra perturbed by disinfectant. By carefully choosing non-contaminated spectra regions, each EBC sample clustered with corresponding samples of the same group, while presenting intergroup qualitative and quantitative signal differences.

The fifth chapter is dedicated to the simulations and the experimental tests of the CAKE integration method. In particular, we tested CAKE integration efficacy on simulated peaks in different overlapping conditions and signal-to-nose ratios. Furthermore, since experimental two-dimensional peak shapes are close to elliptic, we tested CAKE on a simulated ellipse of known volume at different eccentricity degrees. Finally, we used CAKE on experimental NMR data by making use of a sample containing two tripeptides at known concentrations. Peak volume estimations obtained with CAKE comparison with standard methods indicated that CAKE obtains un umbiased volume estimation.

In the sixth chapter, the application of the SO-FAST-HMQC experiment to ^{15}N-labeled *Thalassiosira rotula* diatoms is described. We demonstrate the effective applicability of SO-FAST experiments to cells, collecting spectra in 10-15 s of acquisition time. Our results, definitively show the applicability of SO-FAST experiments for fast metabolic data acquisition thus providing an instantaneous of the metabolic pathways going on in a well-defined physiological state, therefore avoiding the measurement of an "average" metabolism, obtainable with acquisition time of hours.

Chapter 1

NMR analysis and pattern recognition methods

Contents

1.1 Introduction

Metabonomics and metabolomics based on Nuclear Magnetic Resonance (NMR) spectroscopy are nowadays widely used for toxicological assessment, biomarker discovery, and studies on toxic mechanisms. The metabonomic approach, (defined as the quantitative measurement of the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification) was originally developed to assist interpretation in NMR-based toxicological studies. However, in recent years there has been a convergence with metabolomics and other metabolic profiling approaches developed in plant biology, with much wider coverage of the biomedical and environmental fields. Specifically, metabonomics involves the combination of spectroscopic techniques with statistical and mathematical tools to elucidate dominant patterns and trends directly correlated with time-related metabolic fluctuations within spectral data sets, usually derived from biofluids or tissue samples. Temporal multivariate metabolic signatures can be used to discover biomarkers of toxic effect, as general toxicity screening aids, or to provide novel mechanistic information. This approach is complementary to proteomics and genomics and is applicable to a wide range of problems, including disease diagnosis, evaluation of xenobiotic toxicity, functional genomics, and nutritional studies. The use of biological fluids as a source of whole organism metabolic information enhances the use of this approach in minimally invasive longitudinal studies.

In this chapter, the main features of the statistical tools for such investigation are exposed. As described in Chapter 4, we applied the "pattern recognition analysis" to metabonomic characterization of: a) liver alterations during hepatic tumorigenesis and b) exhaled breath condensates (EBC) from patiens with airway diseases. Tissue samples associated with four different liver pathological states collected from surgical excisions and EBC obtained by cooling exhaled air from spontaneous breathing, were analyzed by ^{1}H NMR spectroscopy coupled with multivariate data analysis (MVA). Metabolic profiles were analyzed and clustering analysis readily separated and classified the tissues and the exhaled breath condensates according to the relative pathological conditions.

1.2 Pattern recognition methods for biomarker investigations

The use of chemometric methods to analyze complex spectral data sets was perhaps the most important development in the practical application of metabonomics, and has defined the development and progression of the field ever since. Early pattern recognition studies on NMR data employed a reductionist approach preselecting the metabolite signals of interest. However, NMR spectra yield a unique metabolic fingerprint for each biofluid, sample which consists of thousands overlapping resonances, is obviously of limited use. If the status of a given organism changes, such as in a diseased state or following exposure to a drug, the unique metabolic fingerprint or signature reflects this change [1, 4].

Multivariate statistical methods provide a robust tool for analyzing and maximizing information recovery from complex NMR data sets. Detailed inspection of NMR spectra and integration of individual peaks can give valuable information on dominant biochemical changes; however, subtle spectral variation may be overlooked, and it is difficult to envisage general effects as a function of both dose and time in a large cohort of samples with biological variability. Pattern recognition methods can be used to map the NMR spectra into a representative lower dimensional space such that any clustering of the samples based on similarities of biochemical profiles can be determined and the biochemical basis of the pattern elucidated.

As described in the next section, the first step in metabonomics is spectra classification according to peak patterns. The second one relies upon identification of spectral features responsible for the classification (according to physiological or pathological status), and this can be achieved via both supervised and unsupervised pattern recognition techniques.

1.3 Multivariate data analysis techniques

MVA efficiently extracts useful information from data generated *via* chemical or physical measurements. Indeed, most scientific data generating systems are multivariate, *i.e.* any particular phenomenon we would like to study in detail usually depends on several factors (variables). For instance, the health status of a human individual depends on many elements, including genes, social status, eating habits, stress, environment etc. Consequently, it is often necessary to simultaneously sample several variables to fully describe the system.

A panoply of multivariate data analysis techniques exists, and the choice depends on the answer one wants to obtain. A large part of the method is concerned with simply "looking" at the data, characterizing then by useful summaries and displaying the intrinsic data structures visually by suitable plots. Therefore, it is important to formulate the analytical problem in such a way that the goal is clear and the data are in a form suited for reaching this goal. Usually, spectral data are preprocessed, which typically involves Fourier transformation, calibration of the chemical shift scale with respect to an internal reference standard, and phase and baseline corrected. For multivariate modeling, NMR spectra are often divided into vertical regions (along the chemical shift axis), and their areas summed to provide an integral so that the intensities of peaks in such defined spectral regions can be extracted; such a process is known as bucketing. As a consequence, a data matrix is obtained, which consists of rows that represent observations/samples, and columns that represent variables as the spectral. From this matrix format, data are suitable for MVA that can be used for a number of distinct, different purpose: data description (explorative data structure modeling), discrimination and classification, regression and prediction. So, more simply, we can describe MVA as composed by two main methods: multivariate classification (pattern recognition) and multivariate regression techniques [5, 6, 7, 8].

The pattern-recognition techniques deal with the separation of data groups. Such clustering ability, even for large set of measurements, gives the possibility to derive a quantitative data model in order to discriminate among different groups of data. Multivariate classification can be divided into two categories: unsupervised and supervised procedures. In an unsupervised pattern recognition, no *a priori* knowledge of the training set samples is required, *i.e.* the class membership of the training samples. Hence, samples will be grouped into a number of classes with certain communalities without initial qualification of the samples and their class assignment. Therefore, a possible structure within certain data sets may be recognized even without the initial knowledge of the number of classes and the expected differences. In contrast, a supervised pattern recognition requires *a priori* knowledge about the classes contained within the training samples, *i.e.* which sample belongs to which class, such as, samples from disease and from healthy patients. Consequently, unsupervised pattern-recognition techniques are exploratory methods for data analysis, seeking inherent similarities in the data, and grouping them in a "natural" way. This approach allows unexpected grouping within a training set may be discovered often not initially evident, as for a group of disease-related samples that might additionally separate into two or more distinctly different classes.

Supervised pattern-recognition techniques are different, as they group data into predefined classes during the training procedures, thereby allowing a more precise classification within the class boundaries. Clearly, each approach has strengths and weaknesses rendering a general recommendation impossible. Efforts have been made to combine different pattern-recognition methods for improved classification results [9, 10]. In general, sufficient accuracy and robustness of classification and predictive regression models has to be evaluated with an appropriate set of validation samples prior to the analysis of unknowns.

1.3.1 Unsupervised pattern recognition

Principal Component Analysis (PCA)

PCA constitutes the most basic "work horse" of all of multivariate data analysis. The starting point is an X-matrix with n objects and p variables (an n by p matrix) (Figure 1.1), often called the "data matrix" or the "data-set". The objects can be the observations, samples or experiments, while the variables typically are "measurements" of each object. In our case, the n objects are NMR spectra of samples, while the p variables are integrations of spectra sections, called "buckets", of a well defined size.

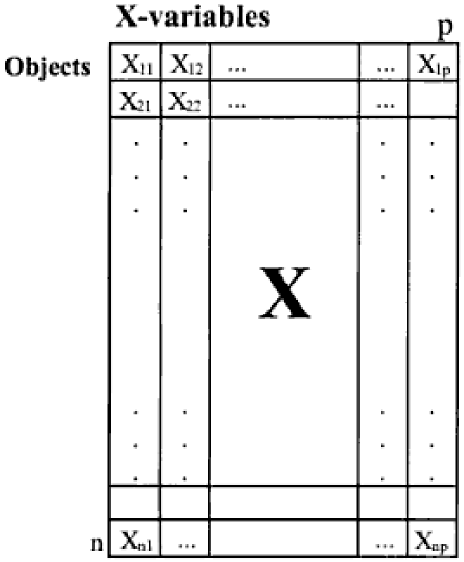


Figure 1.1: X matrix or data matrix consisting of n observations (n NMR spectra) and p variables (p spectral regions "buckts").

The purpose of PCA, so as of all MVA techniques, is to decompose the data in order to detect and model the "hidden phenomena" for which the concept of variance is very important. In fact, the fundamental assumption for this method is that the underlying directions with maximum variance are more or less directly related to the hidden phenomena. The data matrix X, with its p buckets columns and n spectra rows, can be represented in a Cartesian (orthogonal) coordinate system of dimension p called the "variable space" or, in this case, the "spectroscopic space", meaning the space spanned by the p variables corresponding to the buckets. The dimension of this space is p, but the dimension related to the rank of the matrix representation (mathematically: the number of independent basis vectors; statistically: the number of independent sources of variation within the data matrix) may be often less than p. PCA seeks this operative or effective dimensionality by a linear coordinate transformation from the variable space into a space which is spanned by a lower number of new coordinates, called "principal components" (PC_S), which, in turn are related to directions of largest variances in the ensemble (Figure 1.2). The first principal component (PC_1) explains most of the variance, the second (PC_2) the second most, etc. Therefore, PCA is a powerful data-reduction technique that can condense original data (with a large number of initial variables) to a dataset with only few variables reflecting the most relevant analytical information.

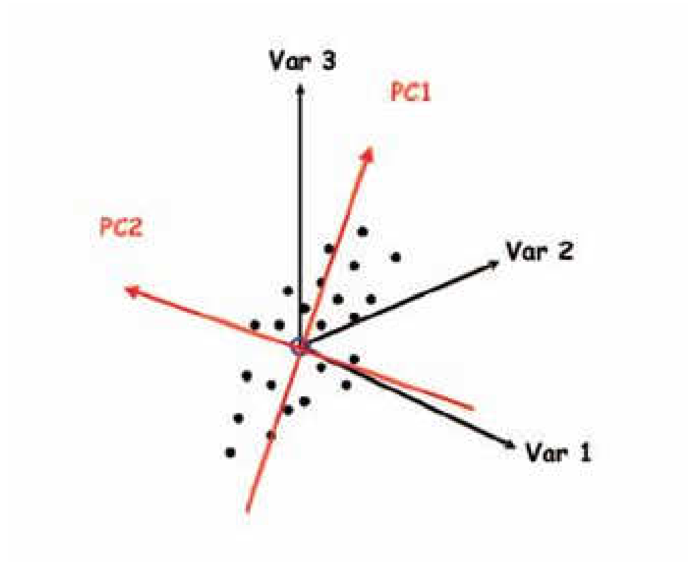


Figure 1.2: Representation of all observations in the data matrix in a 3D space where the computed principal components are shown as vector arrows.

By looking into two-dimensional subspaces like PC_1 *vs.* PC_2, one could see if all spectra have similar positions (scores) with respect to the corresponding part of the variance (Figure 1.3). The corresponding plots are called "scores plots".

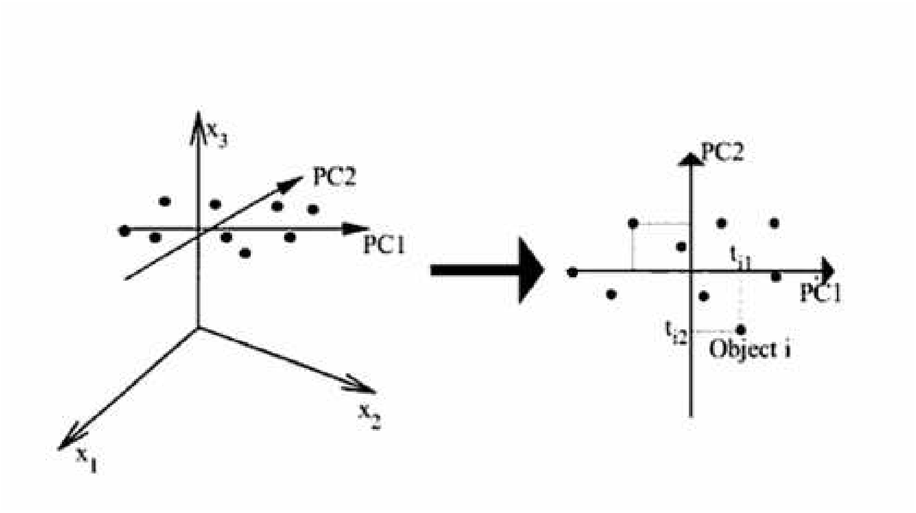


Figure 1.3: Representation of all observations from the variable space to the principal components space. Such PC system consists of a number of PC_S, each lying along a maximum variance directions in decreasing order. Scores plot are obtained as projection of observations onto the PC_S axes.

A further step is to look for further structures in the ensemble by reducing the variance space such that most of the total variance (like 99.5%) is explained and the rest is regarded as noise. The reduced space is called "model space". By calculating quantities like distance to model of each spectrum it is possible to check if all spectra are still similar or if some spectra appear outside this model space. This is also the basis for classification. The relation between the variables in the new principal component space and original spectroscopic space are described by the so-called loadings (ref. Section 1.4). By studying one or two-dimensional loadings plots it is possible to understand how buckets contributed to the construction of the new principal component space. A high loading of a bucket (variable) indicates that the corresponding area (or peak) in the spectrum was important. The loadings plots provide the link between statistical and spectroscopic interpretation of the phenomena in the ensemble. This is essential because PCA itself reveals statistical phenomena but does not explain the reason for these phenomena, for example in chemical terms. This interpretation remains to be done after the PC_S calculation.

1.3.2 Supervised pattern recognition

Projection to Latent Structures Discriminant Analysis (PLS-DA)

PLS-DA is a discriminant method derived from PLS regression models [11] (see next Section). Here, the threshold for separating two classes is calculated using the observed distribution ($P_1, P_2 ... P_m$; m = number of classes; P_m = probability that the spectra belongs to class m) of the predicted values, and the Bayesian theorem, which calculates the probability of one object belonging to a certain class by use of the ratio $\frac{P_i}{\sum P_m}$, for discriminating different classes. Barker et al. describe how PLS-DA statistically connects with discriminant analysis, and may thereby serve as a discriminant tool [11]. For classification, PLS is guided by among-group variance, while PCA, which is guided only by the total variance, cannot discriminate among-group from within group variance. Compared to PCA, it is clear that PLS-DA provides favorable discrimination, especially if the within-group difference dominates over among-group difference. In recent studies, this model was successfully used to discriminate artherosclerotic and normal aorta tissues in rabbit models [29, 48].

1.3.3 Multivariate regression

Principle component regression (PCR)

During PCR, PCA is used to compress and decompose the original spectra generated from training samples into fewer variables (PC_S) capturing the relevant variances within the data set, and then using the scores derived from the training data to create a quantitative model. During the prediction of unknowns, the score vectors of the unknown are derived based on their unique spectra, and regressed against the *PC* vectors obtained from the calibration samples for retrieving a quantitative prediction of the unknown concentration. PCR was also successfully implemented as a classification tool by Haaland et al., and was used to classify cell and tissue samples [12].

Projection to Latent Structures (PLS) regression

PLS also starts out with an ensemble of spectra, which is translated into the X matrix, commonly called the "bucket table" where the number of *p* variables is the number of buckets. However, a second information table is needed. It could comprise other spectroscopic data or any other sort of data, like concentration measurements, arbitrary id numbers, disease characterizations etc. This secondary table is commonly called Y matrix or Y table (Figure 1.4).

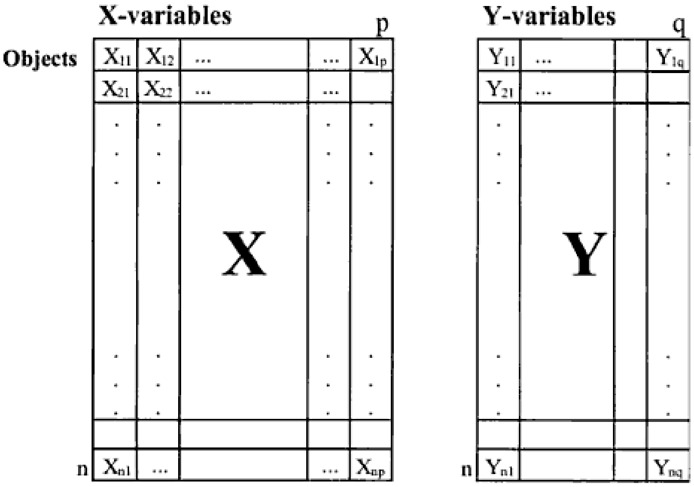


Figure 1.4: X matrix containing data and observations, and Y matrix containing, for each observation, data related to sample information like concentrations or disease classifications.

The number of Y variables (also called response variables or *q* variables) is identical to the number of columns in the Y table. Unlike PCA, which detects the direction of maximum variance in the X matrix, PLS tries to find the best correlation between the X and Y matrices using relevant linear combinations of variables in the X and Y tables. It detects that part of the variance in the

X table which fits best the data in the Y table in an iterative way. While in PCA the user has to decide the number of principal components he wants to work with (typically such that most of the variance in the ensemble is explained), in PLS he has to define the number of PLS components (factors) that should be used to model the Y table. This number is often not obvious. In principle, it should be chosen such that the non-explained variances in X and Y space approach a minimum, and such that the PLS model has good predictive capabilities. Unlike the number of principal components in PCA, the number of PLS factors must be carefully chosen. The results of a PLS calculation are presented in similar ways compared to PCA (again, we get scores and loadings plots of the X table data). However, there are a number of further plots which need interpretation, e.g. showing the correlation between X and Y tables or the prediction power of the model. Similarly to PCA, the model building process in PLS is to find the correct statistical variables (e.g. number of PLS factors), and the right spectra that should stay in the model. Once the model is established (calibrated) it is used to analyze new spectra with missing Y table information and use the constructed model to predict it. This is extremely valuable if the Y table would have been expensive to obtain otherwise, or if it can not be experimentally obtained at all.

There is a second interesting usage of PLS motivated by the following situation. Ensembles often contain different groups of spectra, say normal/abnormal or originate from different samples, say kidney/liver etc. One then would like to see these groups in a PCA analysis, e.g. as different clusters in a scores plot. However, PCA is designed to find the maximum variance in the ensemble but not necessarily that part of the variance that results in the best discrimination. To enforce this, it is of course possible to perform a spectroscopic analysis first and find signals responsible for discrimination, and then use these signals in a subsequent PCA. Alternatively, it is possible to supply a Y table which contains discriminating information (in the most simple case just 0 and 1). A PLS then detects that part of the variance in the ensemble, which fits best to the Y table. A scores plot of the ensemble data may possibly show a good discrimination. How safe is such a proceeding, it depends on the application. With two indistinguishable groups in the ensemble, a PLS using a Y table with 0 and 1 will not provide a good discrimination and the correlation plots between X and Y data would indicate poor correlation. If the ensemble in fact contains two groups of spectra, PLS with a corresponding Y table can indeed improve discrimination. This should however be confirmed by spectroscopic or other data, otherwise a not solid discrimination could be overemphasized.

Orthogonal Projection to Latent Structures Discriminant Analysis (O2PLS-DA)

O2PLS is a multivariate regression method that extracts linear relationships from two data blocks, X and Y, by removing the structured noise [13, 14]. In particular, O2PLS decomposes the systematic variation in the X-block into two model parts: the so called predictive part, which models the correlations between X and Y, and another called the orthogonal part, which is not related to Y. Like other PLS regression techniques, O2PLS can be used to perform discriminant analysis by introducing suitable dummy variables. The main advantage in using O2PLS-DA technique is the reduction of the model complexity. For m classes, the dimension of the predictive space is *m-1*, and the classification model can be investigated by using only *m-1* latent components. Useful visualization tool, as the correlation plot or S-plots, can be used to highlight the role of the X-variables in the classification model.

1.4 Plots and data visualization

As stated in the previous Sections, multivariate methods allow investigation of the relationships between all variables in a single context. These relationships can be displayed in plots like time series, histograms and pair-wise scatter plots.

Model overview plots

Model overview plot could be presented as an histogram showing how the cumulative explained variance (R^2 value) gets larger as the number of the PC_S increases on horizontal axis (Figure 1.5). The number of PC_S for the model should be such that R^2 (sum of squares of all the X matrix variables explained by the extracted components) and Q^2 (the cumulative cross validated R^2) values are somewhere in the flat asymptotic part of curve histogram.

Influence plots

Influence plot shows spectra in a diagram where the vertical axis is a measure of how far a spectrum is from the model space (off model distance). If a spectrum is in the upper part of this display it is most likely outside the model space. The horizontal axis is a measure how far a spectrum is from the model center, after being projected into the model space (in model space). If a spectrum appears on the right side, it has a strong influence on the model.

The two lines displayed inside the plot are so-called 95% confidence limits. Spectra inside these limit belong to the model with a probability of 95%.

Scores plots

Two dimensional scores plots of the form PC_i *vs.* PC_j (e.g. PC_1 *vs.* PC_2*)* show how the spectra are distributed in the corresponding sub-space (Figure 1.6). This plot is used to see whether spectra are gathered in groups or are outlying from others. Dominant effects in the PCA may typically be seen in plots that involve the first few PC_S. Sometimes effects in higher PC_S are equally important; so with PC_1, PC_2 and PC_3 a 3D scores plot can be visualized. It could, for example, indicate strong unexpected signals in a spectrum but present in only very few spectra. By checking the influence plot or all scores plots it can be seen whether higher PC scores plots should be considered.

Loadings plots

Loading plot shows how PC_S are related to the original buckets. The 1D loadings plot of a principal component looks like a spectrum. Peaks indicate those buckets (and therefore spectral regions) which contributed significantly to that principal component. 1D loadings plots, e.g. of PC_1 show how the original variables (buckets) contributed to the construction of a PC. They look like a 1D spectrum and the largest peaks indicate the strongest contributions. 2D loadings plots (Figure 1.7), e.g. of PC_1 and PC_2 relate loadings of the different PC_S to each other. Each point in such a plot corresponds to a pair of buckets. A combined interpretation of scores and corresponding loadings plots can for example show the buckets responsible for an outlying behavior. Combined interpretation means to look for spectra which are outlying along a certain direction, and for loadings which are lined up along the same direction. For example, if a spectrum is outlying in a particular position in the plot, the loadings points into the same direction indicate the resonances responsable for spectrum outlying.

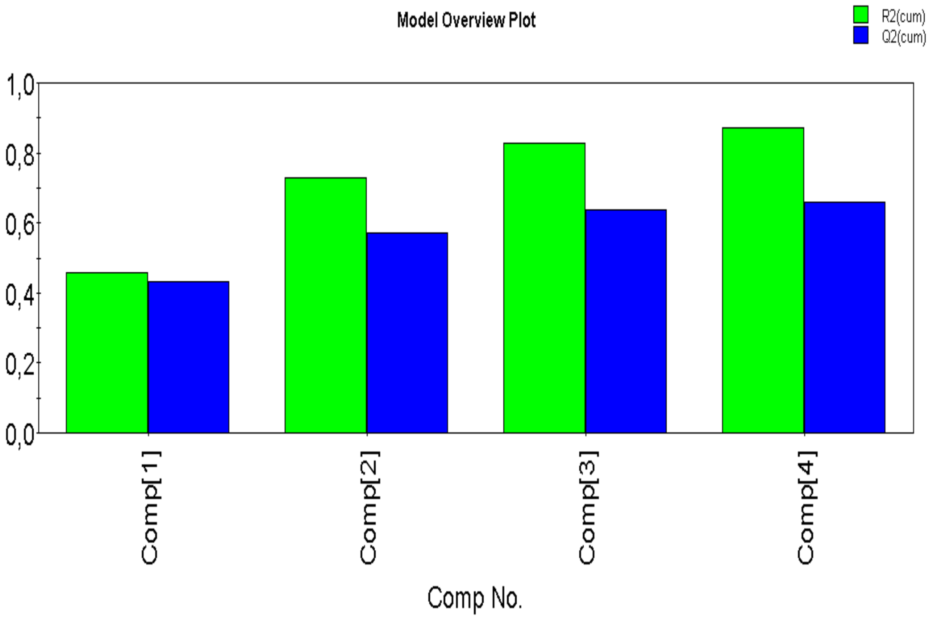


Figure 1.5: Model overview plot: R^2 and Q^2 values are parameters describing how the new PC_S components fit the PCA model.

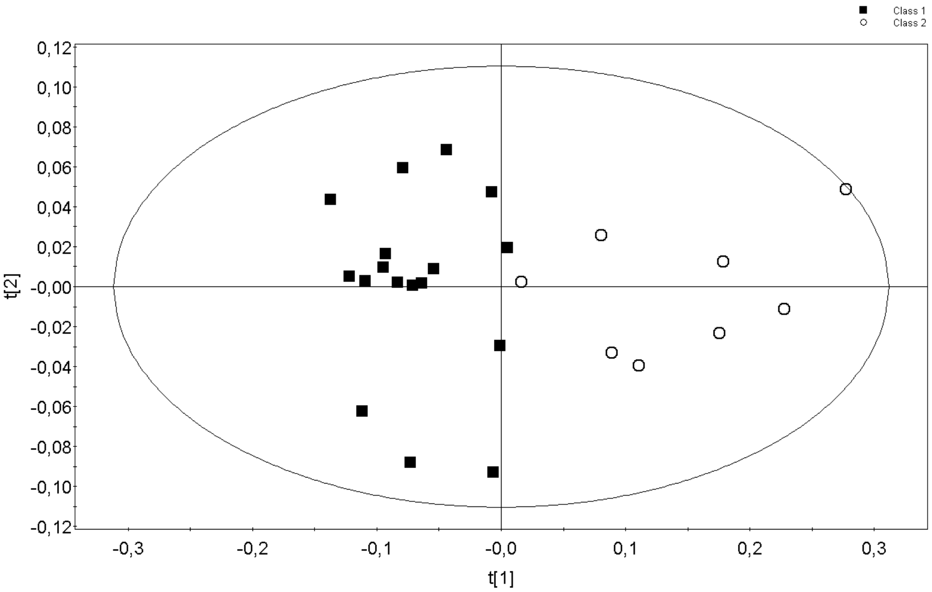


Figure 1.6: PCA scatter plot PC_1 *vs.* PC_2 of two representative class samples.

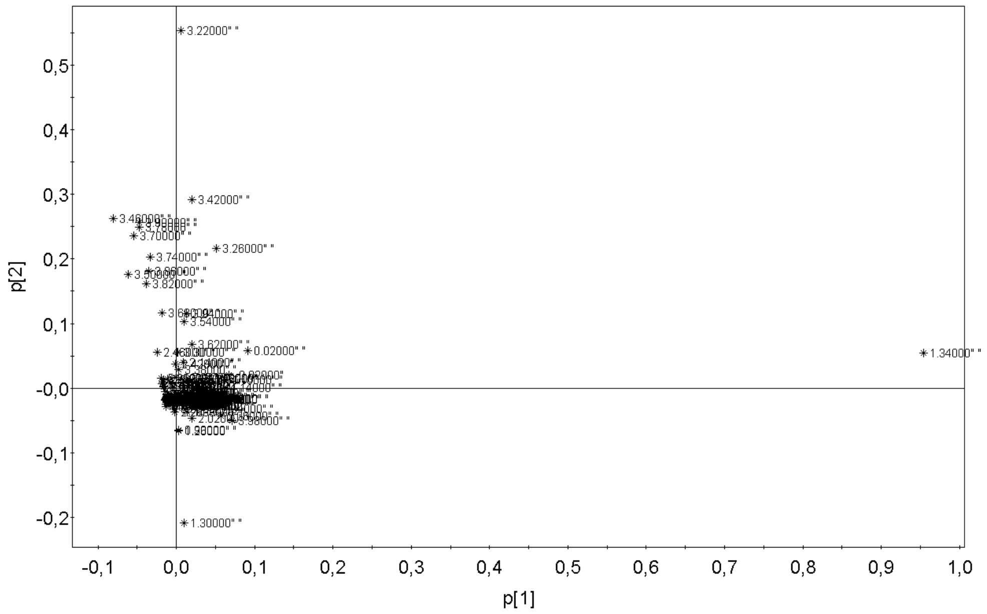


Figure 1.7: Scores scatter plot reporting the buckets responsible for the samples distribution of PCA in Figure 1.6.

1.5 Applications

1.5.1 a) Human hepatocellular carcinoma

The human hepatocellular carcinoma (HCC) is one of the most common malignancies whose incidence is steadily increasing worldwide [15, 16] (Figure 1.8). The liver is also the most frequent site of metastatic colonization, and hepatic metastasis are far more common than primary liver cancers in Western countries [17]. Because of its aggressiveness, early detection of HCC is crucial to schedule more effective therapeutic options and improve patients' survival. The most commonly encountered differential diagnosis in liver is HCC versus intrahepatic cholangiocarcinoma or metastatic adenocarcinoma. Moreover, small hepatic lesions ($\leq$ 1.5 cm in diameter) are frequently difficult to characterize, and diagnostic inaccuracy may lead to incorrect patient treatment. Magnetic Resonance Imaging (MRI) has been shown to effectively differentiate benign and malignant small hepatic lesions with moderate to good interobserver agreement [18, 19]. Yet, the clinical importance of these lesions often remains unknown until biopsy or follow-up imaging is performed months later [20]. Serological markers (such as alpha fetoprotein) can be useful in narrowing the differential diagnosis when they are markedly elevated but a substantial number of patients unfortunately do not have high levels of these markers at the time of presentation. Therefore, a tissue diagnosis is often required, because the presence of hepatic metastasis may substantially alter prognosis and therapy [21].

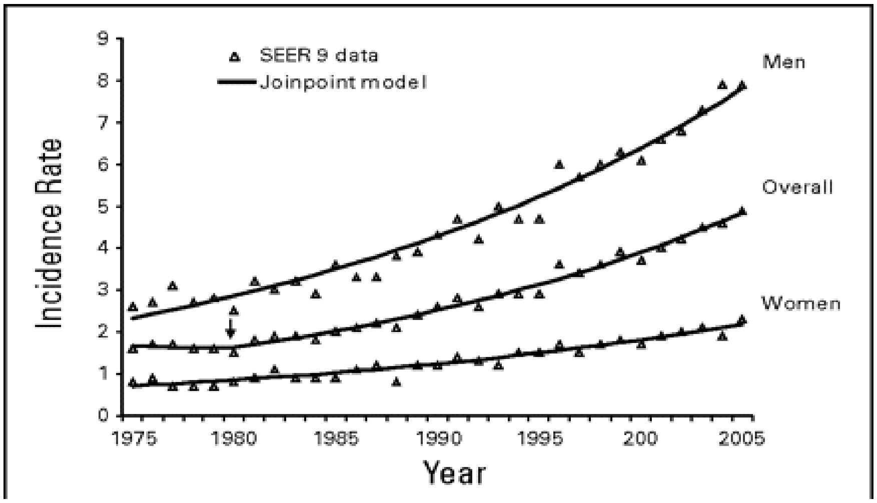


Figure 1.8: Annual age-adjusted incidence rates per 100,000 and trends, all hepatocellular carcinoma cases and by sex, 1975 to 2005 (Surveillance, Epidemiology, and End Results 9 [SEER9]).

Histopathological evaluation of biopsy samples plays a key role in achieving an accurate diagnosis, and fine needle aspiration biopsy of liver has gained increasing acceptance as the diagnostic procedure of choice, and is reported to be safe, minimally invasive, accurate and cost effective [20]. A possible disadvantage of the biopsy-based histopathology is represented by the difficulties in its use as a screening approach for early tumor detection. On the other hand, MRI and all the commonly-used imaging techniques, which are widely accepted as screening tests, provide limited biochemical information (i.e., metabolite composition), which may be useful to discriminate the different hepatic lesions at the molecular level. Evaluation of intracellular metabolic profiles of hepatitis C virus (HCV) infected liver, HCC and metastases is lacking and NMR spectroscopy profiles could contribute to clarify these aspects. NMR is an established analytical tool extensively used for probing the metabolic status of biological samples [22, 23, 1], and provides a "metabolic fingerprint" useful to investigate physiopathological states. As pointed out in the previous sections, the presence of discriminating elements in an NMR spectrum or in spectra belonging to the same class can be tested with multivariate data analysis, which allows a thorough comparison of sets of spectra [24]. As shown in this chapter, some of the most often used techniques to identify models for possible groups as well as to predict a probable class membership for new observations are based on PCA or multivariate regression methods as O2PLS to perform discriminant analysis [25]. As it will be described in the Chapter 4, we used multivariate data analysis to gain insight into hidden phenomena and trends in ensembles of different hepatic tissue spectra which would not be obvious in the usual spectroscopic view. Such an analysis will also point out the most relevant NMR signals for the classification of tissue spectra, clearly indicating changes in concentration of a specific metabolite as well as its relative variation.

In-vitro studies conducted on tissue extracts have shown that high-resolution NMR improves both spectral resolution and sensitivity, yielding more detailed metabolite information [13, 14]. On the contrary, *in-vivo* NMR can detect non-invasively biochemical changes in human cancers [26], liver diseases such as chronic hepatitis [27], cirrhosis and carcinoma [28, 29]. However, spectral resolution and sensitivity makes *in-vivo* NMR of limited value for the identification and quantification of metabolites [30]. A useful diagnostic strategy could be represented by a combination of *in-vitro* and *in-vivo* NMR compared to histological analysis in order to follow-up variations of distinctive lesions classified by high-resolution NMR spectra. We here followed the biochemical progression of human hepatic lesions through NMR-based analysis of primary (HCC) and secondary (metastases from colorectal carcinoma) liver tumors, cirrhotic tissues, and non-cirrhotic normal liver tissues adjacent metastases, achieving a metabolic differentiation of the various pathological conditions based upon the variation of the intracellular lactate/glucose ratio, thus suggesting that such a signal pattern may act as a potential marker for assessing pathological hepatic lesions.

1.5.2 b) Exhaled breath condensate

Exhaled breath condensate (EBC) is a simple, noninvasive and useful tool to study the biochemical and inflammatory molecules in the airway lining fluid [31]. Obtained by cooling exhaled air from spontaneous breathing, EBC predominantly contains water vapour and collects volatile and nonvolatile substances from the lower airways [32]. As such, it can also be considered a matrix for analysis of environmental toxicants and for evaluation of exposure monitoring [33]. Very few data are available on EBC metabolite composition; often single inflammatory molecules are analysed by ELISA and spectroscopic methods.

Since NMR, coupled with pattern recognition methods, has been proved to be a powerful tool for biofluids to probe the metabolic status [34, 1, 23, 35] and to investigate different diseases [36, 37, 38, 39], we applied it to characterize EBC metabolic profile.

Recently, EBC of asthmatic children has been investigated by NMR and statistical analysis [40]. To date, there are several recommendations on the methodological approach to EBC collection, but its standardization is not completely defined, as most inflammatory mediators, obtained through tracheostomies, are similar to those collected in the mouth [41, 42].

The aims of the present study were:

1. To validate the NMR metabonomic approach to analysis of EBC in adults, assessing the role of pre-analytical variables (saliva and disinfectant contamination) potentially influencing EBC and evaluating the stability and reproducibility of samples;

2. To evaluate the possibility of discriminating healthy subjects from patients with airway disease.

As detailed in Chapter 4, in total, 36 paired EBC and saliva samples, obtained from healthy subjects, laryngectomized patients and chronic obstructive pulmonary disease (COPD) patients, were analyzed by means of ^{1}H-NMR spectroscopy followed by principal component analysis. The effect on EBC of disinfectant, used for reusable parts of the condenser, was assessed after different washing procedures. To evaluate intra-day repeatability, eight subjects were asked to collect EBC and saliva twice within the same day. All NMR saliva spectra were significantly different from corresponding EBC samples. EBC taken from condensers washed with recommended procedures invariably showed spectra perturbed by disinfectant. Each EBC sample clustered with corresponding samples of the same group, while presenting intergroup qualitative and quantitative signal differences (94% of the total variance within the data). In conclusion, the nuclear magnetic resonance metabonomic approach could identify the metabolic fingerprint of exhaled breath condensate in different clinical sets of data. Moreover, metabonomics of exhaled breath condensate in adults can discriminate potential perturbations induced by pre-analytical variables.

Chapter 2

CAKE: Monte *CA*rlo pea*K* volume *E*stimation

Contents

This chapter is based on the paper: R. Romano, D. Paris, F. Acernese, F. Barone, A. Motta. *Fractional volume integration in two-dimensional NMR spectra: CAKE, a Monte Carlo approach.* J Magn Res 192 (2008) 294-301.

2.1 Introduction

NMR spectra can provide quantitative analysis of a sample, and a standard 1D ^{1}H-NMR spectrum is often used to obtain a reliable evaluation of peaks. However, as the complexity of the sample increases, resonance overlap becomes a serious problem that easily degrades the accuracy of the analysis, and 2D NMR data are required to gain sufficient discrimination of resonances. Quantification of NMR spectra is also fundamental in the new emergering field of metabolomics/metabonomics [43, 34], and in the structure and dynamics of proteins in solution [44]. This widespread requirement of deriving quantitative information from NMR data has prompted the need to find methods for accurate and precise integration procedures both for 1D and 2D spectra. This paper describes a new simple method for peak volume integration in 2D spectra, which appears to be particularly suited for overlapping peaks. Quantitative information in NMR spectra is brought by peak areas [45]. Two methods of peak integration are often used: direct summation of spectral data points and peak parameter search by curve fitting. In the absence of a model for the peak shape, direct summation appears to be the only practical technique. It is not, however, adaptable to (partially) overlapping peaks, and introduces two kinds of systematic errors. One is due to the approximation caused by the assimilation of the integral of a continuous function with a finite sum [46]; the second one is caused by the parts of the peaks that are left outside of the integration range [47].

Ideally, an efficient integration method should be applicable even when in the presence of peak overlap or artifacts. Many of the available NMR processing and analysis packages achieve volume integration by direct summation of all data points within a polygonal bounding the peak. This procedure requires a reliable definition of the peak area: the circling should be as large as possible to enable for a complete integration, but also small enough to minimize inclusion of artifacts (baseplane rolls, t_1 noise, tails of other peaks). As such, the idealized procedure appears to be restricted to well-resolved peaks. In automated protocols, a possible way to define the area integration makes use of the observation that the slope of a peak height decreases monotonically with the distance to the peak center, at which point it approximates zero [48, 49]. A similar approach defines the peak integration area using an iterative region-growing algorithm [50, 51, 52], which recognizes all data points that are part of a given peak, and the integration is performed on a user-defined threshold level. This procedure works quite satisfactorily even for overlapping peaks, as long as the peak maxima are visibly resolved and therefore recognizable by the peak-picking procedure. In a different approach, the peaks are fitted by a set of reference peaks defined by the user [53, 54, 55]. In order to obtain accurate line shapes and integrals in one dimension, it is necessary to apply a nonlinear curve-fitting procedure [56, 45]. Although this protocol is probably best suited in cases where peaks strongly overlap, it hinges on the careful definition of suitable reference peaks and selection of initial fitting parameters by the user.

A general approach for peak integration would be to exploit the peak symmetry as a criterion to evaluate the peak volume. Symmetry considerations have previously been used for pattern recognition in 2D NMR spectroscopy [57], and only rarely for the analysis of in-phase peaks as in NOESY and TOCSY experiments. The program AUTOPSY used symmetry for automated peak picking in multi-dimensional NMR spectra of proteins [58]. Here we propose CAKE, a novel integration method based on peak symmetry. After a 2D Lorentz-Gauss time domain filtering, the spectral lines are converted into Gaussian lines, therefore presenting a cylindrical or elliptical symmetry. By assuming the vertical axial symmetry of individual peaks (a peak with a unique center corresponds to its maximum), the volume is obtained by multiplying a selected volume fraction by a factor R, which represents a proportionality ratio between the total and the fractional volume, optimized by Monte Carlo techniques. This *minimalistic* approach warrants that the fractional volume can be chosen so as to minimize the effect of overlap in complex NMR spectra. When applied to simulated and experimental 2D in-phase peaks with different degrees of overlap, CAKE (Monte *CA*rlo pea*K* volume *E*stimation) obtains an unbiased volume estimation. It is shown that, compared with the direct summation procedure, the fractional volume approach yields rather good estimates of the peak volumes, even for significant overlap, as long as a single contour level and its center arising from a single peak can be detected.

2.2 The fractional peak method

2.2.1 Peak line shapes in two-dimensional NMR

In high-resolution NMR the frequency domain line shapes are closely approximated by a Lorentzian function. Neglecting coherence transfer echoes, the signal envelope of a 2D experiment can be assumed to have a biexponential form [57]

$$s^{(e)}(t_1, t_2) = s^{(e)}(0,0)\exp(-\lambda^{(e)}t_1)\exp(-\lambda^{(d)}t_2) \tag{2.1}$$

with rates $\lambda = 1/T_2$ in the evolution (e) and detection (d) periods. Such time-domain envelope, decaying exponentially in both dimensions, lacks cylindrical symmetry about the origin $t_1 = t_2 = 0$. After a 2D Fourier transformation, the corresponding 2D absorption peak shows a Lorentzian shape, whose sections, taken parallel to either axis yield pure 1D absorption Lorentzian line shapes. The asymptotic decay is proportional to $(\Delta\omega_{tu}^{(e)})^{-2}$ and $(\Delta\omega_{sr}^{(d)})^{-2}$ on sections parallel to one of the frequency axes, while it is proportional to the inverse fourth power in the bisecting planes [with $(\Delta\omega_{tu}^{(e)})$ and $(\Delta\omega_{sr}^{(d)})$, frequency offset in evolution (e) and detection (d) periods with respect to resonances $\omega_{tu}^{(e)}$ and $\omega_{sr}^{(d)}$]. This lack of cylindrical or elliptical symmetry has been called "star effect", and can be removed by a 2D Lorentz-Gauss transformation [57], which yields a 2D absorption mode peak shape with cylindrical or elliptical symmetry (Figure 2.1 and 2.2).

By using a weighting function

$$h(t_1, t_2) = \exp(+\lambda_1 t_1)\exp(+\lambda_2 t_2)\exp(-\sigma_1^2 t_1^2/2)\exp(-\sigma_2^2 t_2^2/2) \tag{2.2}$$

with σ being an adjustable parameter, the envelope of Eq. 2.1 becomes

$$s^{e}(t_1, t_2) = s^{e}(0,0)\exp(-\sigma_1^2(t_1^2/2))\exp(-\sigma_2^2(t_2^2/2)). \tag{2.3}$$

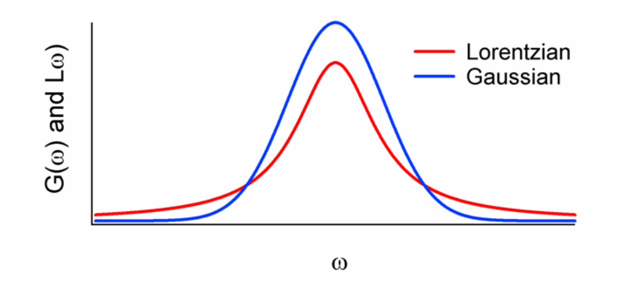


Figure 2.1: 1D profiles of Lorenztian and Gaussian peaks.

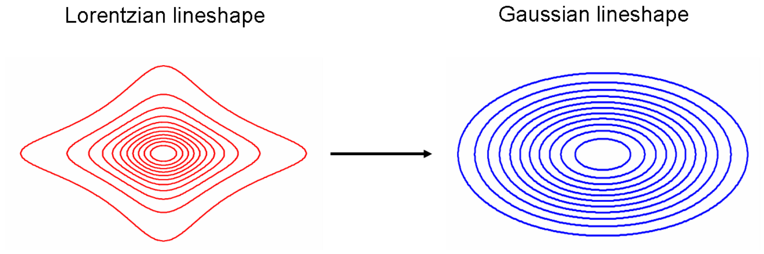


Figure 2.2: Removal of the so-called "star effect" of a Lorentzian peak by a 2D Lorentz-Gauss transformation.

After a 2D transformation, a Gaussian line shape is obtained

$$S(\omega_1,\omega_2) = s^{(e)}(0,0)(\frac{2\pi}{\sigma_1\sigma_2})\exp{(-\frac{\Delta\omega_1^2}{2\sigma_1^2})}\exp{(-\frac{\Delta\omega_2^2}{2\sigma_2^2})}. \tag{2.4}$$

The contours are circular for $\sigma_1 = \sigma_2$ and elliptical for unequal widths. It is important to underline that 2D Lorentz-Gauss transformation is useful only if the dispersive components in peaks with mixed phase are suppressed, and this can be achieved with pure phase spectra (i.e. either pure 2D absorption or pure 2D dispersion peaks) [57]. It must also be emphasized that the elliptical symmetry of Gaussian signals is obtained only in phase-sensitive displays, and if the absolute amplitude of a Gaussian signal is calculated, a peak shape is obtained which features again a star effect.

In most practical applications, the complete analytical expression for a discrete Fourier transform NMR spectrum is a sum of complex, non-Lorentzian functions ([45, 59]). However, if the acquisition time t_2 is large, compared to the relaxation time of the slowest decaying resonance ($t_2 \geq 1/R_{2,j}$), and the sweep width is large compared to the relaxation rate $R_{2,j}$ as well as the frequency range of the spectrum $\nu_j - \nu$, a true Lorentzian spectrum is obtained [60]. Nevertheless, this discrete Fourier transform spectrum requires correction of a pseudobaseline stemming from the first point of the FID and of a frequency-dependent phase distortion of the spectrum (for details see Refs. [59, 45]). Accordingly, a phased, baseplane corrected unsaturated resonance line in solution is closely approximated by a Lorentzian function. Convolution of the time domain with exponential, sine, cosine functions, does not alter the line shape after transformation [61], and preserves the frequency of its maximum. This shape has been useful in peak fitting procedures applied to experimental data [60]. As stated above, a 2D Lorentzian line lacks cylindrical or elliptical symmetry, which can be achieved by a 2D Lorentz-Gauss transformation. Gaussian filtering transforms a Lorentzian frequency-domain function of width ω_0 into a Gaussian frequency-domain function of width $\rho\omega_0$, where ρ is typically less than unity, and it has been found that $\rho = 0.66$ is usually close to optimum [62].

Bearing in mind the power of Lorentz-Gauss tranformation and the symmetry of the Gaussian line, the CAKE algorithm aims at integrating a peak relying upon its axial symmetry, even when in drastic overlapping conditions. The idea is that the volume can be estimated by integrating a non-overlapping fraction of the peak obtaining a reasonable approximation of volume in cases where cross peaks overlap. Therefore the major assumption in this study is that the Lorentzian signal is transformed into a Gaussian line by a Lorentz-to-Gauss transformation. For in-phase peaks of TOCSY and NOESY spectra, such a transformation is well-suited, especially considering that the multiplet structure of the in-phase components is only barely resolved and a maximum signal-to-noise ratio is usually required to detect even weak signals [57].

Figure 2.3A shows the contour plot of a Gaussian peak. The arbitrary angle $A\widehat{O}B$ (a "slice" selected in a non-overlapping region and centered on the center of mass), defines the area A_{F_i} of a peak fraction for each $i-th$ level bound curve; such an angle identifies a fractional volume V_{F} in the three-dimensional representation. Because of the axial symmetry, for each $i{-}th$ level the fractional volume V_{F} relates to the total volume V_T as the fractional area of each level relates to the corresponding total area A_{T_i}. From the equation

$$V_{\mathrm{T}} = \frac{A_{\mathrm{T}_i}}{A_{\mathrm{F}_i}} \cdot V_{\mathrm{F}}, \tag{2.5}$$

true for each couple of level bound areas, if $R_i = \frac{A_{\mathrm{T}_i}}{A_{\mathrm{F}_i}}$, the total volume of a peak can be obtained by multiplying a fractional volume by the corresponding R_i factor.

It is common experience that experimental 2D peak shapes are quite close to an ellipse. Therefore, Eq. (2.5) is still valid if the right angle $A\widehat{O}B$ delimits $\frac{1}{4}$ of the ellipse by lying on the semimajor and the semiminor axes (Figure 2.3C).

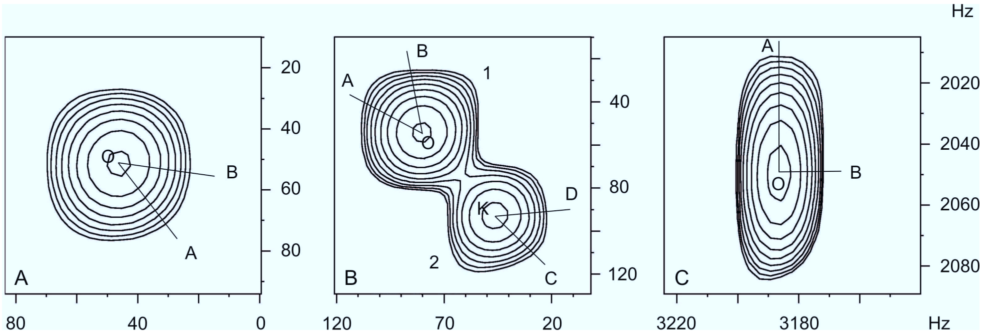


Figure 2.3: Contour plots of simulated isolated (A) and overlapping (B) Gaussian peaks. In (A), the arbitrary angle $A\widehat{O}B$ defines a fraction of the peak area, selected in a non-overlapping region, and centered on the center of mass. In (B), $A\widehat{O}B$ and $C\widehat{K}D$ select a fraction of peaks 1 and 2, respectively. (C) Experimental Gaussian cross-peak. The right angle $A\widehat{O}B$ selects a fractional area corresponding to $\frac{1}{4}$ of the total area.

In particular, by defining the ellipse eccentricity as $e = \sqrt{1 - \frac{b^2}{a^2}}$, where b and a are the semiminor and the semimajor axes (assuming $b < a$), $0 \leq e \leq 1$ and $e = 0$ in the case of a circle. More generally, it can be demonstrated that Eq. (2.5) applies with a good approximation to eccentricity $e \leq 0.5$, which corresponds to a difference $< 10\%$ between axes, and a circle well approximates the ellipse. For eccentricity $e > 0.5$, Eq. (2.5) can be safely used if the polygonal $A\widehat{O}B$ identifies a region symmetrical with respect to one of the semiaxes. The advantage of this approach becomes apparent for overlapping Gaussian peaks. Here, the integration is biased by the presence of the overlapping region that affects both volumes. In contrast, the "slice" $A\widehat{O}B$ of peak 1 (Figure 2.3B), selected in a non-overlapping region, has very little contribution, if any, from peak 2, and therefore its fractional volume can mostly be attributed to peak 1. The same is true for $C\widehat{K}D$ slicing peak 2 (Figure 2.3B), whose fractional volume can mostly be attributed to peak 2. Therefore, if we integrate the volume fraction identified by $A\widehat{O}B$ and calculate the corresponding R_1 constant, it should be possible to estimate the unbiased volume of each peak. From Figure 2.3B, the second most internal (highest) level of peak 1, essentially arises from peak 1, and the effect of peak 2 on that level is negligible. Consequently, the R_1 constant can be obtained from the ratio between the total area (A_{T_1}) and the fractional area (A_{F_1}) of that level, $A_{\mathrm{T}_1}/A_{\mathrm{F}_1}$. Analogously, for peak 2 the fractional volume identified by $C\widehat{K}D$ can be considered, and its second highest level can be chosen to obtain the respective factor R_2 (Figure 2.3B).

2.2.2 The R factor estimation

In order to estimate the R factor for a selected fraction of a peak, an internal level attributable to the peak has to be chosen. Denoted by A_{T} the total level area and by A_{F} the fractional level area, the ratio $R = A_{\mathrm{T}}/A_{\mathrm{F}}$ can be obtained by a Hit-or-Miss Monte Carlo technique [63, 64]. Let us denote by (lx_i, ly_i), with $i = 1, 2, ..., N$, the vertex coordinates of the polygonal P_{level} relative to a contour level, by (c_x, c_y) the coordinates of its center point, and by α_1, α_2 two rays with their common origins in (c_x, c_y). The fractional area A_{F} is therefore defined by the intersection of the polygon P_{level} and the area delimited by the rays (Figure 2.4). Furthermore, let us denote by lx_{min} and lx_{max} the minimum and maximum lx_i coordinates, and by ly_{min} and ly_{max} the minimum and maximum ly_i coordinates, respectively. Two pseudo random numbers x_r and y_r are now uniformly extracted in the intervals $[lx_{\mathrm{min}}, lx_{\mathrm{max}}]$, and $[ly_{\mathrm{min}}, ly_{\mathrm{max}}]$, respectively. The extraction is continued until a number $N_{\mathrm{A_T}}$ of points (x_r, y_r) is internal to the polygonal P_{level}. If an extracted point (x_r, y_r) is also inside the area A_{F}, then the number of fractional hits $N_{\mathrm{A_F}}$ is augmented by one. Of course, being the (x_r, y_r) pairs uniformly extracted in the rectangle $[lx_{\mathrm{min}}, lx_{\mathrm{max}}] \times [ly_{\mathrm{min}}, ly_{\mathrm{max}}]$, the ratio $R = A_{\mathrm{T}}/A_{\mathrm{F}}$ will be estimated by the ratio $R = N_{\mathrm{A_T}}/N_{\mathrm{A_F}}$.

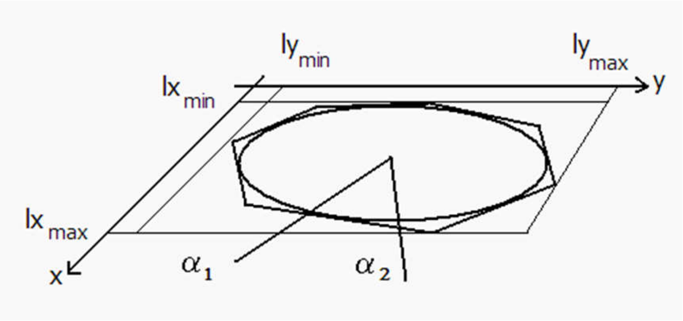


Figure 2.4: Total level area A_{T} and fractional level area A_{F} defined by the intersection of the polygon P_{level} and the area delimited by the rays α_1, α_2.

2.2.3 The Monte Carlo integration

In principle, any method is suitable to integrate the selected fractional volume. However, the simple sum can be biased because of the small region and the limited number of points within the selected area. Accordingly, the Monte Carlo Hit-or-Miss technique appears to be more suitable. Let us denote by (px_i, py_i), with $i = 1, 2, 3, 4$, the vertex coordinates of the quadrilateral P_{base}, which is the base of a prism of height h and that contains the fractional volume

V_{F} (in particular, $px_1 = c_x$, and $py_1 = c_y$, while other two points are chosen on the α_1 and α_2 rays). Furthermore, let $px_{\min}$ and $px_{\max}$ be the minimum and maximum px_i coordinates, and $py_{\min}$ and $py_{\max}$ the coordinates corresponding to the minimum and maximum py_i, respectively. Two pseudo random numbers x_r and y_r are uniformly extracted in the intervals $[px_{\min}, px_{\max}]$ and $[py_{\min}, py_{\max}]$, respectively. The extraction is continued until the extraction number $N_{\mathrm{P_{base}}}$, which represents the number of points (x_r, y_r), is internal to the quadrilateral of base P_{base}. If a point (x_r, y_r) is internal to the quadrilateral of base P_{base} and to the polygonal base P_{level}, a cubic interpolation gives the peak $p(x, y)$ values in the point (x_r, y_r), and another pseudo random number ρ is uniformly extracted in the interval $[0, 1]$. If $\rho{\cdot}h \leq p(x_r, y_r)$, that is, if $\rho{\cdot}h$ is a point internal to the fraction volume V_{F}, the number of volume hits N_V is augmented by one. If V_{P} is the prism volume (Figure 2.5), calculated by the software, then the fractional volume V_{F} is estimated as

$$V_{\mathrm{F}} = N_{\mathrm{V}}{\cdot}V_{\mathrm{P}}/N_{\mathrm{Pbase}} \tag{2.6}$$

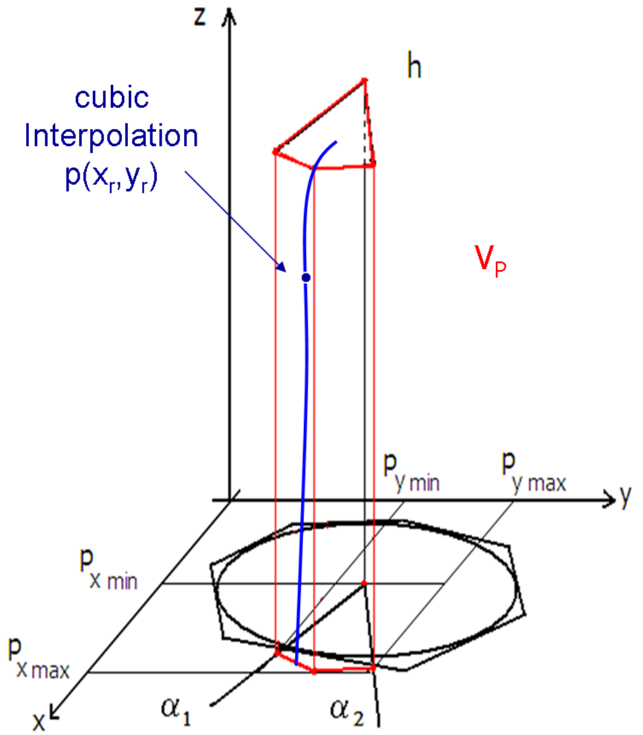


Figure 2.5: Prism of volume V_{P} that contains the fractional peak volume V_F.

Chapter 3

Fast NMR techniques in metabonomics

Contents

3.1 Introduction

NMR has found an increasingly broad range of applications in different fields of research ranging from physical and material sciences to chemistry, biology, and medicine. Because it interacts with nuclear spins by using very weak electromagnetic fields, NMR is virtually the only technique that provides atomic-level information without disturbing the chemical properties of the molecules and materials under investigation. This enormous versatility has been possible because of the development of a wide range of NMR tools through the years. Among the major achievements one should cite Fourier-transform NMR that had a dramatic effect on the experimental sensitivity of NMR [65], and the introduction of multidimensional NMR spectroscopy by Jeener [66] and Ernst [67] in early seventies.

In recent years, NMR spectroscopy faces a number of new challenges, such as the investigation of the structure and dynamics of biological molecules of increasing size and complexity, the characterization of protein-complexes, as well as the study of kinetic features of biochemical processes in the cell. This requires further technical and methodological improvements in terms of experimental sensitivity, spectral and temporal resolution. New advanced NMR pulse sequences and acquisition schemes are thus required that make optimal use of the improved instrumental performance, and are best adapted to the scientific problems in mechanistic systems biology. It has to be pointed out that the wide variety of methods recently developed for fast data acquisition are mostly addressed to protein structure elucidation and protein-ligand kinetic investigations. Therefore, fast NMR acquisition schemes are shaped and configured on relatively large molecules. In such context, this chapter will explore the application of a fast-pulsing NMR experiment for metabolic profile characterization, thus requiring the optimization of a recent pulse sequence for fast HMQC acquisition, called SOFAST-HMQC [2, 3] (band-Selective Optimized Flip-Angle Short Transient heteronuclear multiple quantum coherence), of small molecules such as metabolites.

3.2 Fast multidimensional NMR spectroscopy

Multidimensional NMR experiments are crucial for the study of biomolecular structure and dynamics as they provide the required resolution to extract spectral parameters for individual nuclear sites in the molecule. While in 1D NMR the time evolution of nuclear spin magnetization is detected directly via the electric current induced in a receiver coil, the evolution in a so-called indirect time domain is monitored by stepwise increments of a delay in the pulse sequence. As a consequence of this time increments procedure, the experimental time required for the acquisition of an nD NMR spectrum increases by *ca.* 2 orders of magnitude per additional dimension. Therefore, even if the inherent sensitivity is sufficient, complete sampling of the indirect time domain grid imposes lower limits on the experimental times: several minutes for 2D, several hours for 3D and so on. Therefore, new acquisition schemes are required for a more rapidly data recording, taking care to obtain a sufficient signal-to-noise ratio. In order to speed up multidimensional NMR data acquisition, the sampling problem can be resolved either by limiting the number of data points (sparse or non-uniform sampling techniques), or by reducing the duration of each repetition of the experiment (fast pulsing techniques). Most of the existing fast acquisition techniques are based on the first solution, incomplete sampling of the indirect dimensional time space. Examples are non-uniform data sampling combined with non-linear processing schemes [68, 69], reduced dimensionality or projection NMR [70, 71, 72, 73], and Hadamard NMR [74, 75] where data sampling is realized directly in the frequency domain. All of these methods basically allow recording of multidimensional correlation spectra in an experimental time ranging from a few minutes up to several hours.

The ultimate solution to the NMR data sampling problem has recently been proposed and experimentally demonstrated by Frydman and co-workers [76]. Their ingenious concept of "single-scan" NMR allows recording of any multidimensional NMR spectrum within a single repetition of the experiment. Despite the high potential of single-scan NMR for future biomolecular applications, this technique currently requires a very high intrinsic sensitivity and spectrometer hardware optimized for both NMR spectroscopy and imaging purposes. On the other hand, for application to proteins in aqueous solution, several scans are generally required to yield good water suppression and acceptable signal to noise in a few seconds of experimental time.

NMR fast pulsing techniques present an alternative way to reducing acquisition times. The main idea is to shorten the time delay between successive scans (recycle delay) to achieve higher repetition rates and thus collect the same number of scans in less time. Of course, the number of data points to be recorded can also be reduced as discussed above, which makes fast-pulsing techniques fully compatible with sparse sampling approaches. A recycle delay is required to allow relaxation of the excited spins (usually ^{1}H) towards their thermodynamic equilibrium, and to build up sufficient ^{1}H polarization to be used for the next scan.

In order to keep the experimental sensitivity high enough while using fast repetition rates, some spectroscopic tricks are required. A first approach has become known as longitudinal relaxation enhancement [77]. Such method is based upon the fact that the efficiency of ^{1}H spin-lattice relaxation is increased if nearby ^{1}H are unperturbed by the pulse sequence, so that they can take up some of the energy put into the system *via* dipole-dipole interactions (nuclear Overhauser effect, NOE), or via hydrogen exchange. In practice, the relaxation enhancement is realized by selectively manipulating a subset of the proton spins of interest in a well defined spectral region throughout the pulse sequence, thus ensuring that the spin states of all other protons that are not directly involved in the coherence transfer pathways of a particular experiment remain unperturbed. This yields reductions in effective longitudinal ^{1}H relaxation times from a few seconds to a few hundred milliseconds. In some circumstances, *e.g.*, in HMQC experiments, the sensitivity of fast-pulsing experiments can be even further enhanced by adjusting the excitation flip angle to the socalled Ernst angle [57, 78]. Both effects have been combined in the SOFAST experiment [2, 3] that allows one to record 2D ^{1}H-^{15}N or ^{1}H-^{13}C correlation spectra of proteins in only a few seconds, thus opening new avenues for real-time investigations of protein kinetics at atomic resolution. We explored the potential of such experiment for metabolic profiling issue by applying it to cell samples for fast detection of metabolites.

3.3 SOFAST-HMQC

The introduction of SOFAST-HMQC sequence by Shanda and Brutscher represents an alternative technique for fast acquisition of 2D heteronuclear correlation spectra. The sequence is realized by using very short interscan delays therefore combining the advantages of a small number of radio-frequency pulses, Ernst-angle excitation, and longitudinal relaxation optimization [77, 79] to obtain an increased signal to noise ratio for high repetition rates of the experiment. Since SOFAST-HMQC uses standard data sampling in the indirect dimension, it has the further advantage of being therefore easily implemented on any commercially available high-field NMR spectrometer. Figure 3.1 shows the basic pulse scheme to record SOFAST-HMQC spectra.

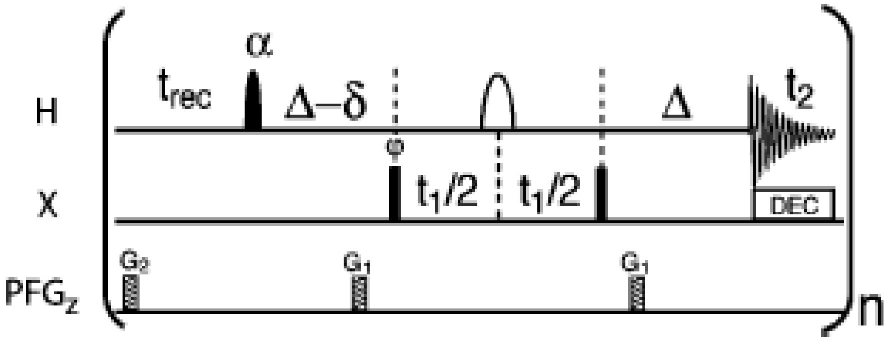


Figure 3.1: SOFAST-HMQC experiment to record ^{1}H-X (X=^{15}N or ^{13}C) correlation spectra of proteins. Filled and open pulse symbols indicate 90° and 180° rf pulses, except for the first ^{1}H excitation pulse applied with flip angle α. The variable flip-angle pulse has a polychromatic PC9 shape, and band-selective ^{1}H refocusing is realized using an r-SNOB profile. The transfer delay Δ is set to 1/(2JHX), the delay δ accounts for spin evolution during the PC9 pulse, and t$_{rec}$ is the recycle delay between scans.

3.3.1 General aspects

This pulse sequence provides the required high sensitivity to perform fast heteronuclear ^{1}H-X correlation experiments of macromolecules by using very short recycle delays (t$_{rec}$). The main features of SOFAST-HMQC are the following:

- The HMQC-type ^{1}H-X transfer steps require only few rf pulses which limits signal loss due to B$_1$-field inhomogeneities and pulse imperfections. Rf pulses reduction will be especially important if the experiment is performed on a cryogenic probe, where B_1-field inhomogeneities are more pronounced.

- The band-selective ^{1}H pulses reduce the effective spin-lattice relaxation times (T_1) of the observed proton spins. The presence of a large number of non-perturbed ^{1}H spins, interacting with the observed ^{1}H via dipolar interactions (*NOE* effect), significantly reduces longitudinal relaxation times whereby the equilibrium spin polarization is more quickly restored.
- The adjustable flip angle of the proton excitation pulse allows further enhancement of the available steady-state magnetization for a given recycle delay.

3.3.2 Ernst-angle excitation

The repetition rate of an NMR pulse sequence depends on the delay t_{rec} between the first pulse of one scan and the first pulse of the next scan. If the spin system is saturated by fast rf pulsing, short interscan delays (t_{rec}) lead to a significant loss in signal intensity. Ernst and co-workers developed an elegant technique to optimize the sensitivity in fast pulsed 1D one-pulse NMR experiments by the application of a non-90° flip-angle [57], known as the Ernst angle [80, 44]. Maximal signal for an interscan delay t_{rec}, and longitudinal relaxation time T_1, is obtained by the application of an excitation angle β_{Ernst} given by:

$$\cos(\beta_{Ernst}) = \exp(\frac{-t_{rec}}{T_1}) \tag{3.1}$$

T_1 is the effective spin-lattice relaxation time constant assuming mono-exponential polarization recovery. The longitudinal equilibrium magnetization M_{eq} in dependence of the thermal equilibrium magnetization M_0 is

$$M_{eq} = M_0 \frac{(1 - \exp(-t_{rec}/T_1))}{(1 - \exp(-2t_{rec}/T_1)} \tag{3.2}$$

The signal resulting from a single rf pulse applied to M_{eq} with a flip-angle β_{Ernst} is

$$Signal = M_{eq} \sin(\beta_{Ernst}) \tag{3.3}$$

and the signal-to-noise ratio per measurement time, referred to as the sensitivity of the single pulse experiment [57], is

$$Sensitivity = Signal/\sqrt{t_{rec}} \tag{3.4}$$

In the case of SOFAST-HMQC sequence (Figure 6.1) Equation 3.1 becomes

$$\cos(\beta_{Ernst}) = \exp(\frac{-T_{rec}}{T_1}) \tag{3.5}$$

with T_{rec} the effective ^{1}H longitudinal relaxation delay including the inter-scan delay (t$_{rec}$), the acquisition times $t_{1/2}$ and t_2 , and the transfer delay Δ (Figure 6.1) and the S/N per unit experimental time, neglecting transverse spin relaxation effects and other sources of signal loss, is then given by

$$S/N \propto \frac{(1-\exp(-T_{rec}/T_1))}{1-\exp(-T_{rec}/T_1)\cos(\beta)} \cdot \frac{\sin(\beta)}{\sqrt{nT_{Scan}}} \tag{3.6}$$

with β the effective flip angle $\beta = \alpha - 180°$ taking into account the effect of the ^{1}H refocusing pulse, and T_{Scan} the time required for a single scan including the pulse sequence duration, acquisition time, and the inter-scan delay (t_{rec}).

3.3.3 Proton band-selective pulses

The performance of SOFAST-HMQC critically depends on the choice of the pulse shapes for the band-selective excitation and refocusing pulses on the ^{1}H channel. Actually, the longitudinal relaxation optimization enhancement effect is strictly related to the number and type of the applied proton pulses. For this purpose, Shanda and co-workers [3] used only 2 (band-selective) ^{1}H pulses in SOFAST-HMQC thus ensuring minimal perturbation of the undetected proton spins, and providing higher enhancement factors than observed with other longitudinal relaxation optimized pulse schemes [77]. More over, since the water resonance is outside the selected ^{1}H pulse bandwidth, the WATERGATE-type [81] pulse sequence element G$_1$-180°(^{1}H)-G$_1$ (Figure 6.1) yields efficient water suppression within a single scan. The selective proton manipulation also removes coupling evolution between excited ^{1}H spins and passive ^{1}H spins from frequency bands that are not perturbed by the selective pulses.

As spin refocusing pulse, Shanda and co-workers [3] first chose r-SNOB profile [82] for it presents the advantage of a short pulse length thus reducing signal loss due to transverse spin relaxation [2]. Afterwards, they tested other pulse shapes and found that, for ^{1}H-^{15}N correlation spectra, a REBURP (Figure 3.2) profile yields higher sensitivity despite a 3-times longer pulse duration. Experimental comparison of r-SNOB and REBURP performance in ^{1}H-^{15}N SOFAST-HMQC showed signal increase of up to 50% observed when using REBURP instead of r-SNOB for short scan times. Such result depend on better off-resonance performance of REBURP, resulting in less perturbation of the aliphatic ^{1}H spin polarization and, as a consequence, shorter longitudinal relaxation times of the amide proton spins.

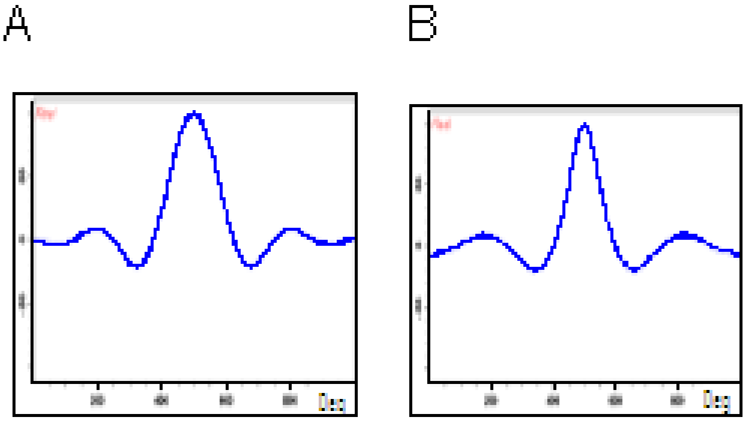


Figure 3.2: Excitation shaped pulses profiles. A) PC9 pulse; B) RE-BURP pulse.

The most band-selective "top-hat" pulse shapes commonly used for NMR spectroscopy, *e.g.* BURP [83], Gaussian pulse cascades [84], or SNOB [82], have only been optimized for discrete flip angles of 90° or 180°, and generally are not useful for variable flip angle excitation purposes. In contrast, polychromatic (PC) selective pulses have been shown to perform well for a whole range of flip angles [85]. These PC pulses are based on a series of simultaneously applied, frequency shifted basic pulse elements. For the SOFAST application, Shanda and co-workers used the PC9 excitation pulse shape (Figure 3.3), which has the required "top-hat" excitation profile for flip angles $0° < \alpha < 120°$.

Moreover, unlike other band-selective excitation pulses that yield "pure-phase" transverse magnetization, the PC9 pulses produce phase that is a linear function of the frequency offset. So, Shanda and Brutscher proposed to replace a PC9 pulse by the combination of a pure-phase excitation pulse followed by a delay δ. The chemical shift and scalar J$_{HX}$ coupling evolution occurring during this delay δ can be accounted for by adjusting the subsequent transfer delay of the HMQC sequence to $1/(2J_{HX}) - \delta$ (Figure 6.1). If the delay δ has been properly adjusted prior to data acquisition no first-order phase correction is required in the ^{1}H dimension. Otherwise, pure-phase spectra can still be obtained by applying a first order phase correction.

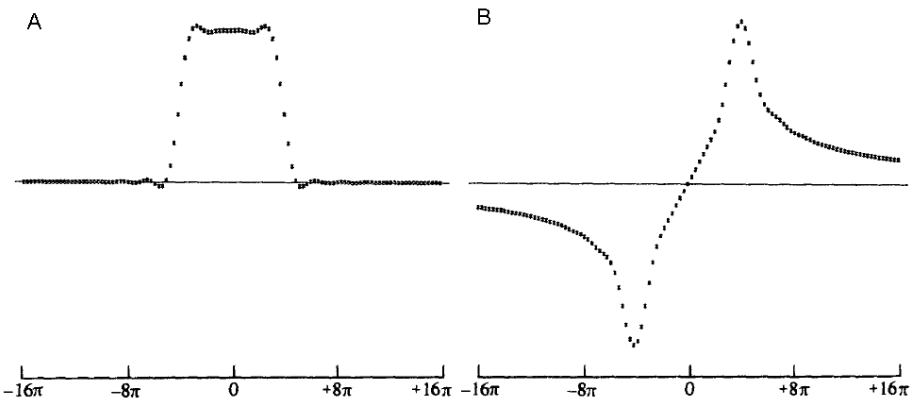


Figure 3.3: Simulated frequency-domain response of the polychromatic pulse PC9 consisting of nine radiofrequencies spaced $\Delta f = 1/T$ apart with relative intensities of 1:2:2:2:2:2:2:2:1. A) Absorbtion; B) dispersion.

3.3.4 Application to protein

The SOFAST-HMQC pulse sequences of Figure 6.1 have been designed to provide high sensitivity for fast repetition rates. To examine the performance of the SOFAST-HMQC experiment for the desired short interscan delays, Shanda and co-workers measured 1D spectra of ^{15}N-labeled ubiquitin. Figure 3.4 shows the measured S/N ratios for constant experimental time as a function of the duration of a single repetition of the experiment T_{scan} (taking into account the length of the pulse sequence, data acquisition time, and recycle delay) for ubiquitin sample acquired at 600 MHz on a spectrometer equipped with a standard probe (Figure 3.4a) and at 800 MHz on a spectrometer equipped with cryoprobe (Figure 3.4b). Such spectra provide only information on the average signal to noise ratio obtained by the different pulse sequences. Each intensity point was obtained by scaling all spectra to the same noise level according to the number of applied scans, and integrating the spectral intensity over the range 7.0-9.5 ppm. The curves are therefore representative of the average behavior of the experiment for all amide sites in the protein.

The SOFAST-HMQC data (Figure 3.4) for three different flip angles (90°, 120°, and 150°) are compared to results from a sensitivity-enhanced (se) water-flipback (wfb) HSQC pulse sequence, and from a longitudinal relaxation optimized HSQC (LHSQC) experiment [77].

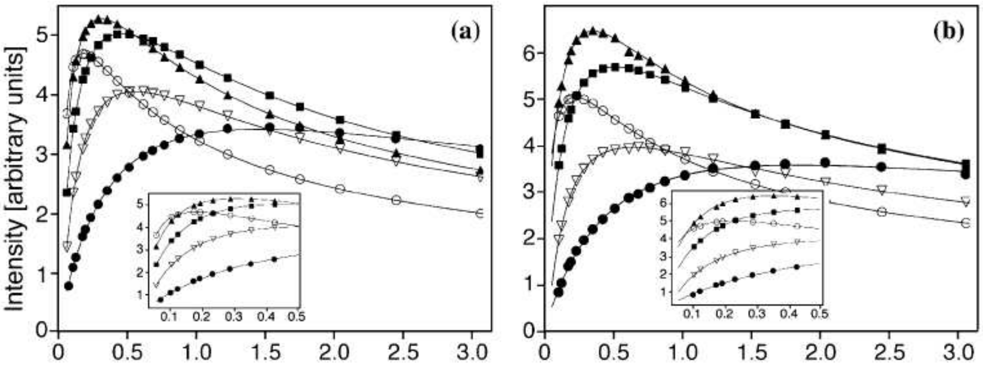


Figure 3.4: Signal-to-noise ratios per unit time (intensity) plotted as a function of the scan time (T_{scan}) obtained with different ^{1}H-^{15}N correlation experiments for (a) ubiquitin (8.6 kDa, 2 mM, 25 °C, pH 6.2) at 600 MHz, (b) ubiquitin at 800 MHz. The intensities were extracted from 1D spectra recorded using the SOFAST-HMQC sequence of Figure 6.1 (t_1=0) with flip angles of a α= 90° (■), 120° (▲) and 150° (○), LHSQC (▽), and se-wfb HSQC (●). Band-selective ^{1}H pulses in the SOFAST-HMQC and LHSQC experiments were centered at 8.0 ppm covering a bandwidth of 4.0 ppm. Variable flip angle excitation and refocusing in SOFAST-HMQC were realized using a PC9 pulse of 3.0 ms and a REBURP pulse of 2.03 ms, respectively.

The principal conclusions from those experimental results are the following:

- Using optimized acquisition parameters (scan time, flip angle) and moderate t_1 acquisition times, SOFAST-HMQC yields the most sensitive ^{1}H-^{15}N correlation spectra of folded proteins.

- SOFAST-HMQC provides a much higher sensitivity than se-wfb-HSQC using the same scan times, and a similar sensitivity as se-wfb-HSQC recorded with optimized inter-scan delays.

The SOFAST features showed in Figure 3.4 could be used as guidelines for setting up SOFAST-HMQC experiments. For practical applications the authors recommended to fix the scan time (recycle delay) and then optimize the flip angle of the PC9 excitation pulse experimentally by recording a series of 1D SOFAST-HMQC spectra varying the power level (flip angle) of the PC9 pulse.

3.4 Real-time cell $^1\boldsymbol{H}$-$^{15}\boldsymbol{N}$ metabolic profile

NMR is a well-established technique for monitoring metabolism in living cells. They are often investigated by 1D NMR spectroscopy, therefore benefiting of real-time measurements since all spectral frequencies are excited by a single scan. However, 1D NMR lacks the resolution needed to cope with the degeneracy of the NMR resonance frequency and a reasonable S/N ratio, the latter because of the short acquisition time required for the short lifetime of samples. The lack of resolution can be circumvented by 2D spectroscopy that, compared with 1D, does yield higher resolution, but is intrinsically time-consuming because data acquisition for the second dimension spans at least several minutes. As discussed before, the total experimental time will be given by the product of the number of scans N_{scan}, required for a proper sampling of the indirect domain, and the single-scan duration (the repetition time) T_{scan}, which includes the spin relaxation time necessary to restore the thermal equilibrium before the next additional measurement. This recycle delay is therefore associated with the ^{1}H spin-lattice relaxation time T_1, and, depending on its duration, acquisition times can be of the order of minutes, yielding total experimental times of hours.

Cells are able to survive and stay suspended in the solvent medium for several hours, but, after only few minutes, oxygen starvation changes their metabolism and decreases the cytoplasmic pH [86]. Therefore, long acquisition times may detect small molecules originating from an "average" metabolism that does not correspond to the physiological state of the cell. For samples with short lifetime data acquisition must be rapid, and fast-acquisition 2D techniques, as those used to study the structure and dynamics of proteins in solution are required [87]. Two different strategies have been put forward for fast acquisition spectroscopy: the "single-scan" NMR [76, 88] and the SOFAST-HMQC. The single-scan approach is able to record any multidimensional NMR spectrum within a single repetition of the experiment, but with current spectrometer hardware it typically lacks in sensitivity, resolution, and/or sufficient gradient strength over extended periods of time. Alternatively, the SOFAST method is able to drastically reduce T_{scan} by relaying on accelerated T_1 of the spins of interest [77] and on optimized flip-angles (*e.g.*, the Ernst angle [57]) to enhance the steady-state magnetization of the excited spins [78]. As pointed out in the previous section, Brutscher and co-workers have combined these features into single 2D and three-dimensional NMR protocols [2, 3, 89, 90], showing that it is possible to reduce T_{scan} down to 100 ms, obtaining 2D ^1H-^{15}N or ^1H-^{13}C correlation spectra in the range of seconds and with high S/N ratio.

Because of its adaptability to routine spectrometers, we have investigated the possibility of using the SOFAST-HMQC approach to explore cellular metabolism in ^{15}N-labeled cells. In Chapter 6 we report that the SOFAST experiment allows acquisition of 2D ^{1}H-^{15}N correlation spectra of small metabolites directly in living cells in few seconds, with a high S/N ratio, therefore affording a picture of the "instantaneous" in-cell metabolism. In particular, we have applied the SOFAST-HMQC experiment to ^{15}N-labeled diatoms cells, which are unicellular algae with silicified cell walls.

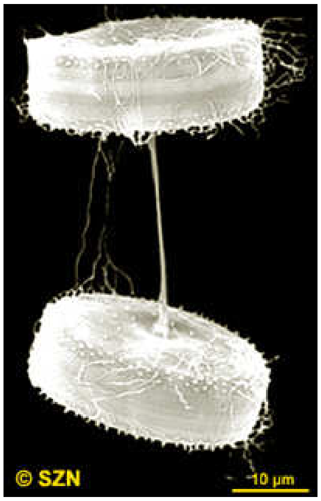


Figure 3.5: *Thalassiosira rotula* image from SEM microscope.

They are at the base of the marine food web, and are the major contributors to phytoplankton biomass worldwide. In response to favorable light and nutrient conditions, diatoms rapidly divide and form large blooms, and as blooms propagate, nutrients are depleted, growth ceases, and cells sink to the deep ocean. The sinking diatom blooms fuel the biological carbon pump and export carbon from the atmosphere to the deep ocean. Despite this, little is known about the molecular underpinnings of diatom biology. As a part of a long-running project, we have recently undertaken a study of the metabolic profile of *Thalassiosira rotula* (Figure 3.5) to understand how diatoms acquire nutrients, how they respond to stress, and how they activate chemical defense and chemical signaling that regulates algal bloom. Although useful information can be achieved by investigating the metabolic profile of polar and lipophilic extracts, in-vivo studies of *T. rotula* cells in (artificial) sea water are expected to yield a more reliable understanding of the metabolic pathways.

On the other hand, the presence of salt in the artificial sea water culture medium, used to suspend the cells in the NMR tube, will cause resonance broadening, and this, together with the degeneracy of the resonance frequency, will make 1D spectroscopy useless. *T. rotula* cells can easily be cultured on unlabeled and ^{15}N-labeled media, and this warrants that a sufficient number of colonies can rapidly be obtained to test the potential application of the SOFAST-HMQC sequence to ^{15}N-labeled cells. The 2D correlation spectra obtained for *T. rotula* cells in 10-15 seconds with a high S/N ratio suggest that fast acquisition techniques introduced for proteins can be easily extended to other living cell systems, monitoring the metabolism under physiological or stressing conditions in the emerging fields of metabolomics and metabonomics [91, 35].

Chapter 4

NMR metabolic profile experiments

Contents

This chapter is based on the following papers:
a) D. Paris, D. Melck, M. Stocchero, O. D'Apolito, R. Calemma, G. Castello, F. Izzo, G. Palmieri, G. Corso, A. Motta. *Monitoring liver alteration during hepatic tumorigenesis by NMR profiling and pattern recognition*, submitted to Metabolomics.
b) G. de Laurentiis, D. Paris, D. Melck, M. Maniscalco, S. Marsico, G. Corso, A. Motta and M. Sofia. *Metabonomic analysis of exhaled breath condensate in adults by Nuclear Magnetic Resonance spectroscopy.* Eur Respir J 2008; 32: 1-9.

4.1 Materials and methods: a) hepatocellular carcinoma

Specimens collection

Liver tissues were collected from patients with diagnosis of hepatocellular carcinoma (HCC) developed on liver hepatitis C virus (HCV) related cirrhosis (CIR) or liver metastasis from colorectal carcinoma (MET-CRC). The portions of the surgically excised samples that were addressed to NMR spectroscopy consisted of HCC tissues (HCC; N = 17), with the corresponding

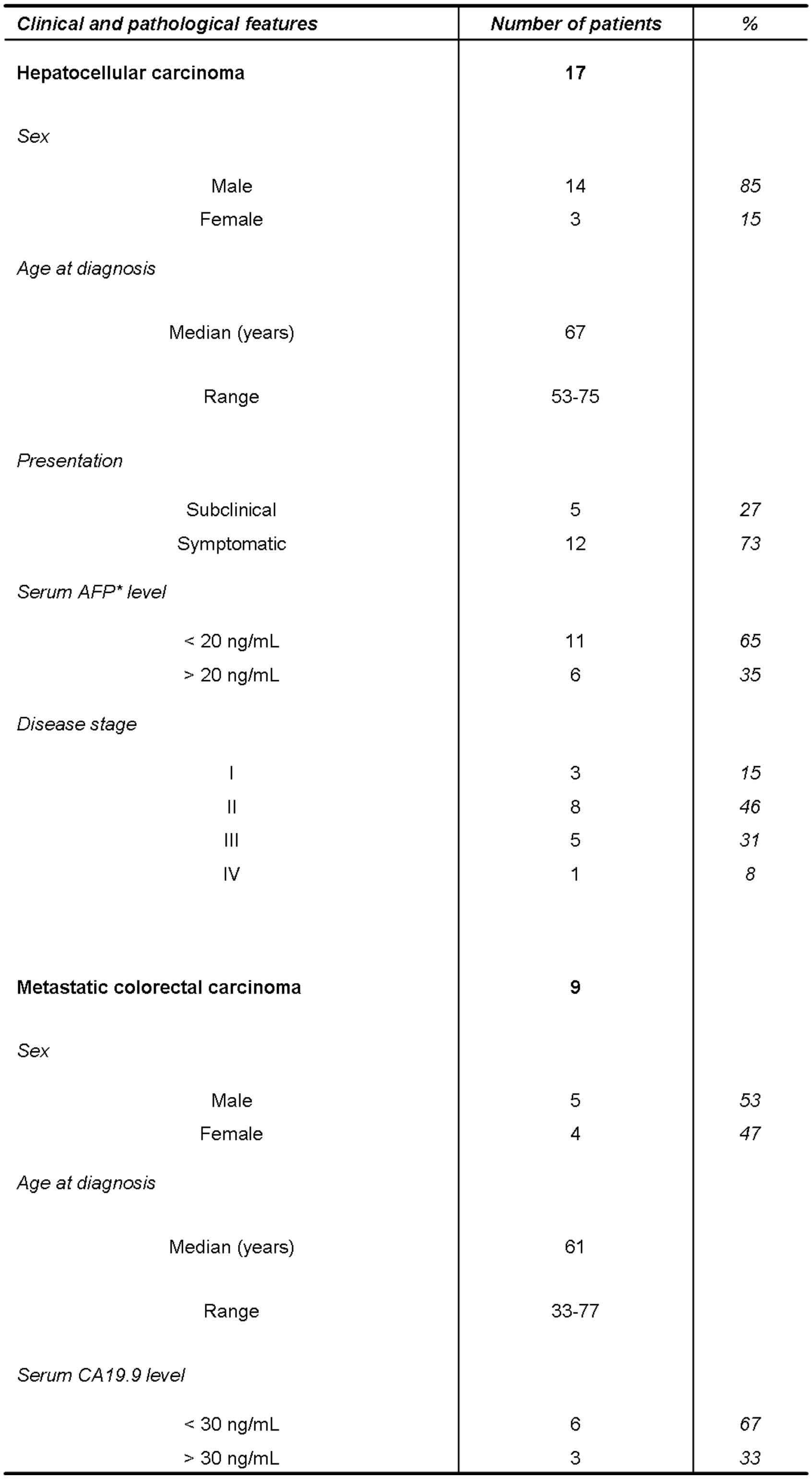
HCV -related cirrhotic tissues (CIR; N = 17), tissues from liver metastases (MET-CRC; N = 9), and the corresponding adjacent non-cirrhotic liver tissues plus two liver tissues from healthy subjects (NT; N = 11). All samples were frozen in liquid nitrogen in order to immediately "quench" any metabolic reaction and preserve metabolite concentrations. Tissues were stored at -80 °C until extraction to prevent any metabolic decay. Pathological evaluation was performed on each case, histopathological classification was based on the criteria of World Health Organization; disease status at the time of diagnosis was defined depending on clinical staging as assessed by medical history, physical examination, and instrumental tests. A written informed consent for tissue sampling was obtained before the analysis from cancer patients. The study was reviewed and approved by the ethical review board at the National Cancer Institute - G. Pascale Foundation - of Naples. The main characteristics of cancer patients are presented in table of Figure 4.1 .

Sample preparation

Tissues were mechanically disrupted to deproteinize the sample and permanently halt the metabolism. The procedure allowed extraction of only the metabolites of interest (e.g., lipids, carbohydrates, amino acids and other small metabolites) while leaving others compounds (e.g., DNA, RNA, proteins) in the tissue pellet. Combined extraction of polar and lipophilic metabolites was carried out by using methanol/chloroform as suggested by the Standard Metabolic Reporting Structures working group [92]. It appears to be the preferred choice for metabonomic NMR studies considering yield, reproducibility, ease and speed, as perchloric acid extracts show a large sample-to-sample variation [93], especially for particularly lipid-rich tissues such as liver and brain [93, 94]. Homogenization of 30 mg of frozen tissue samples was carried out in 8 ml/g of wet tissue of methanol and 1.70 ml/g per wet tissue of water (all solvents were cold) with UltraTurrax for 2 min on ice. Then, 4 ml/g wet tissue of chloroform were added and the homogenate was stirred and mixed, on ice, delicately using an orbital shaker for 10 min (the solution must be mono-phasic). Then, other 4 ml/g wet tissue of chloroform and 4 ml/g wet tissue of water were added and the final mixture was shaken well and centrifuged at 12000 g for 15 min at 4 °C. This procedure separates three phases: a water/methanol phase at the top (aqueous phase, with the polar metabolites), a phase of denatured proteins and cellular debris in the middle and a chloroform phase at the bottom (lipid phase: with lipophilic compounds). The upper and the lower layers of each sample were transferred into glass vials and the solvents were removed under a stream of dry nitrogen and stored at -80 °C until required. For one-dimensional (1D) and two-dimensional (2D) homonuclear NMR experiments the polar extracts were resuspended in 700 μl Phosphate Buffer Saline (PBS, pH 7.4) with D_2O 10% for lock procedure, and then transferred in an NMR tube. For 2D heteronuclear ^{1}H-^{13}C experiments, the polar fraction was resuspended in 700 μl of D_2O.

NMR measurements

1D spectra were recorded at 600.13 MHz on a Bruker Avance-600 spectrometer, equipped with a TCI CryoProbe^{TM} fitted with a gradient along the Z-axis, at a probe temperature of 27°C and acquired at the Institute of Biochemical Chemistry in Pozzuoli (Napoli). 1D proton spectra were acquired by using the excitation sculpting sequence [95]. We used a double-pulsed field gradient echo, with a soft square pulse of 4 ms at the water resonance frequency, with the gradient pulses of 1 ms each in duration, adding 1024 transients of 16384 points with a spectral width of 7002.8 Hz. Time-domain data were all zero-filled to 32768 points, and prior to Fourier transformation, an exponential multiplication of 0.6 Hz was applied. Clean total correlation spectroscopy (TOCSY) [96, 97, 98] spectra were recorded using a standard pulse sequence, and incorporating the excitation sculpting sequence for water suppression. In general, 320 equally spaced evolution-time period t_1 values were acquired, averaging 4 transients of 2048 points, with 7002.8 Hz of spectral width. Time-domain data matrices were all zero-filled to 4096 points in both dimensions, thus yielding a digital resolution of 3.42 Hz/pt. Prior to Fourier transformation, a Lorentz-to-Gauss window with different parameters was applied for both t_1 and t_2 dimensions for all the experiments. TOCSY experiments were recorded with spin-lock period of 64 ms, achieved with the MLEV-17 pulse sequence. Spectra were referred to 0.1 mM sodium trimethylsilylpropionate (TSP), assumed to resonate at $\delta = 0.00$ ppm. The natural abundance 2D ^{1}H-^{13}C Heteronuclear Single Quantum Coherence (HSQC) spectra were recorded on the Avance-600 spectrometer operating at 150.90 MHz for ^{13}C, using an echo-antiecho phase sensitive pulse sequence using adiabatic pulses for decoupling [99, 100]. 128 equally spaced evolution time period t_1 values were acquired, averaging 48 transients of 2048 points and using GARP4 for decoupling. The final data matrix was zero-filled to 4096 in both dimensions, and apodized before Fourier transformation by a shifted cosine window function in t_2 and in t_1. Linear prediction was also applied to extend the data to twice its length in t_1. Spectra were referred to the lactate doublet (βCH3) resonating at 1.33 ppm for ^{1}H, and 20.76 ppm for ^{13}C.



Clinical and pathological features	***Number of patients***	*%*
Hepatocellular carcinoma	**17**	
Sex		
Male	14	*85*
Female	3	*15*
Age at diagnosis		
Median (years)	67	
Range	53-75	
Presentation		
Subclinical	5	*27*
Symptomatic	12	*73*
Serum AFP level*		
< 20 ng/mL	11	*65*
> 20 ng/mL	6	*35*
Disease stage		
I	3	*15*
II	8	*46*
III	5	*31*
IV	1	*8*
Metastatic colorectal carcinoma	**9**	
Sex		
Male	5	*53*
Female	4	*47*
Age at diagnosis		
Median (years)	61	
Range	33-77	
Serum CA19.9 level		
< 30 ng/mL	6	*67*
> 30 ng/mL	3	*33*

Figure 4.1: Characteristics of cancer patients (*AFP, alpha-fetoprotein).

Statistical and multivariate data analysis

High resolution ^{1}H-NMR spectra were automatically data reduced to integrated regions ("buckets") having equal width of 0.04 ppm over the spectral region between 0.04 and 9.40 ppm by using AMIX 3.6 software package (Bruker Biospin, Germany). The residual water resonance region (4.72 - 5.10 ppm) was excluded and the integrated region was normalized to the total spectrum area. To differentiate liver tissues through NMR spectra, we carried out a multivariate statistical data analysis using projection methods. The integrated data reduced format of the spectra was imported into SIMCA-P+ 12 package (Umetrics, Umea, Sweden), and Principal Component Analysis (PCA) and Orthogonal Projection to Latent Structures Discriminant Analysis (O2PLS-DA) were performed. Mean-centering was applied as data pre-treatment for PCA, while Pareto scaling and mean-centering were used prior to O2PLS-DA. Both the ANOVA and the t-test were used for statistical analysis of the signals selected for quantification.

4.2 Results

NMR experiments

NT, CIR, HCC and MET-CRC underwent a dual-phase extraction, and the aqueous fractions were investigated by high-resolution NMR. Typical spectra of NT (trace A), CIR (trace B), HCC (trace C) and MET-CRC (trace D) are reported in Figure 4.2. Although isolated resonances can readily be assigned to specific metabolites by comparing their chemical shifts with literature data [101, 102], line overlapping prevented the complete spectral identification. This required homo- and heteronuclear 2D experiments such as TOCSY (Figure 4.3) to identify ^{1}H-^{1}H connectivities, and ^{1}H-^{13}C HSQC (Figure 4.4) for directly bonded ^{1}H and ^{13}C nuclei. Thus, we were able to identify all resonances by a comparison with literature data and with NMR spectra of standards acquired in separate experiments. The ^{1}H assignments are reported in table of Figure 4.5. Inspection of 4.2 shows clear visible differences among NT, CIR, HCC and MET-CRC. The spectral region from 0.5 to 3.00 ppm contains signals assigned to leucine, valine, threonine, alanine, lysine, glutamate, glutamine, and some organic acids such as lactate, acetate and succinate. The region from 3.0 to 4.5 ppm includes signals attributed to creatinine, choline, arginine, phosphoethanolammine, phosphocholine, glycerolphosphatidilcholine, α-glucose, trimethylamine-N-oxide, glycine, glycogen, *myo*-Inositol and glycerol, and represents the most variable region. The 4.5-7.5 ppm region, together with the residual water signal eliminated by the specific pulse-sequence used in the experiment, contains the resonances of β-glucose, fumarate, tyrosine, histidine and phenylalanine. The region 5.5-6.4 ppm does not contain signals, and as such it has been omitted from 4.2.

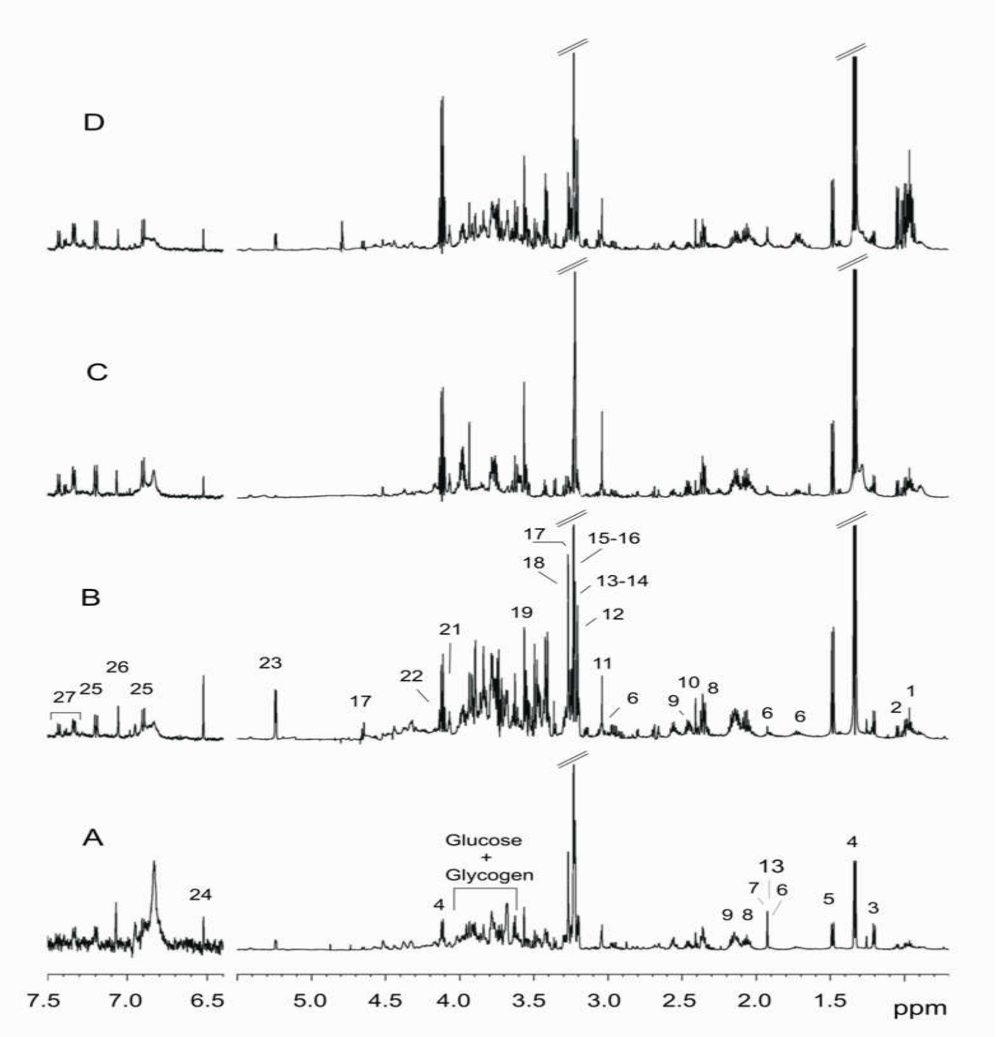


Figure 4.2: Representative aliphatic ^{1}H-NMR spectra of all liver tissue extracts used in this study (spectra scaled to TSP): (A) control non-tumoral adjacent to metastasis (NT) and (D) metastasis from the same patient (MET-CRC); (B) cirrhotic adjacent to HCC (CIR) and (C) HCC from the same patient (HCC). Numbers labels: 1, Leucine; 2, Valine; 3, Threonine; 4, Lactate; 5, Alanine; 6, Lysine; 7, Acetate; 8, Glutamate; 9, Glutamine; 10, Succinate; 11, Creatine; 12, Choline; 13, Arginine; 14, Phosphoethanolamine; 15, Phosphocholine; 16, Glycerophosphocholine; 17, β-Glucose; 18, Trimethylamine-N-oxide; 19, Glycine; 20, Glycogen; 21, *myo*-inositol; 22, Glycerol; 23, α-Glucose; 24, Fumarate; 25, Tyrosine; 26, Histidine; 27, Phenylalanine.

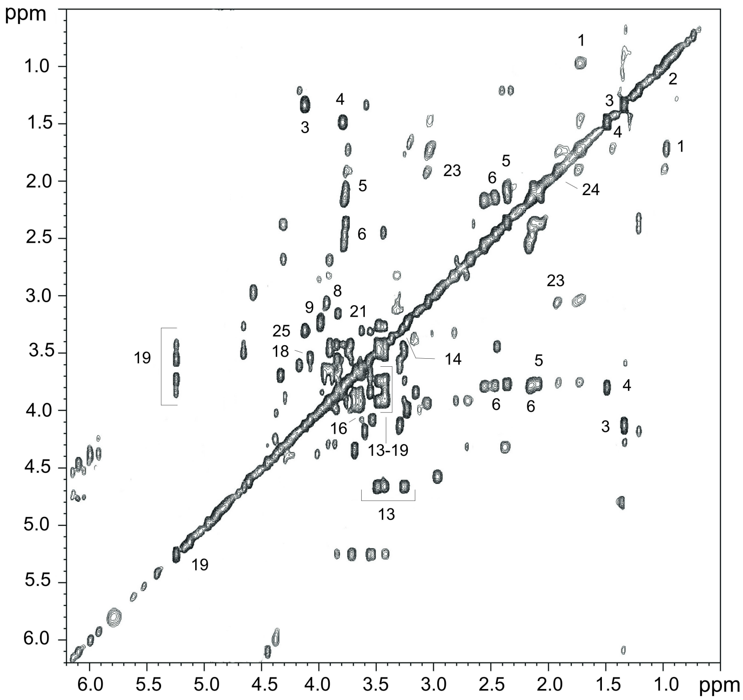


Figure 4.3: Typical TOCSY spectrum of HCC extract sample; for metabolites identification see Figure 4.5 caption.

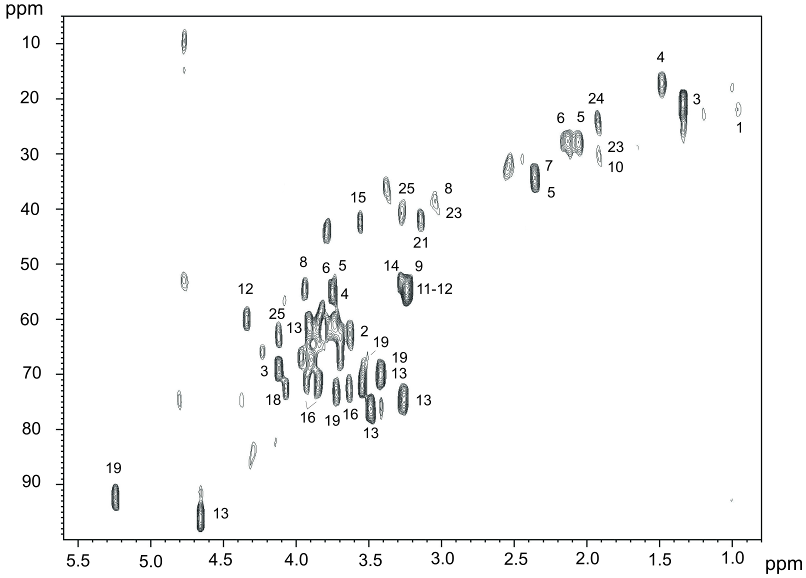


Figure 4.4: Example of ^{1}H-^{13}C HSQC spectrum of HCC sample; for metabolites identification see Figure 4.5 caption.

Entry	Metabolite	δ ^{1}H	δ ^{13}C	Group	Entry	Metabolite	δ ^{1}H	δ ^{13}C	Group
1	Leucine	0.96 1.72	22.03 40.50	δCH_3 βCH_2	15	PC[a]	3.23 3.61 4.17	54.69 66.96 58.78	$N(CH_3)_3$ NCH_2 OCH_2
2	Valine	1.04 3.63	17.86 62.52	γCH_3 αCH	16	GPC[a]	3.23 3.69 4.33	54.69 66.02 59.94	$N(CH_3)_3$ NCH_2 OCH_2
3	Threonine	1.20	21.93	γCH_3	17	β-glucose	4.64 3.26 3.48 3.40 3.47 3.90	96.40 74.60 76.22 70.70 76.80 61.40	C1H C2H C3H C4H C5H C6H
4	Lactate	1.34 4.13	20.81 69.12	βCH_3 αCH	18	TMAO[a]	3.27	55.20	$N(CH_3)_3$
5	Alanine	1.48 3.75	17.37 55.00	βCH_3 αCH	19	Glycine	3.56	42.52	CH_2
6	Lysine	1.91 1.71 3.05	30.95 26.70* 39.36	βCH_2 δCH_2 εCH_2	20	Glycogen	3.63 3.93 3.87 3.86	72.82 72.96 72.47 62.09	C2H C3H C5H C6H
7	Acetate	1.91	24.30	CH_3	21	*Myo*inositol	4.06	72.70	C2H
8	Glutamate	3.77 2.06 2.35	55.13 27.70 34.30	αCH βCH γCH_2	22	Glycerol	3.69 3.84 4.11	63.91 71.64 63.28	C1H C2H 1-CH_2
9	Glutamine	3.75 2.15 2.45	55.13 27.60 32.00	αCH βCH_2 γCH_2	23	α-glucose	5.24 3.54 3.72 3.42 3.84 3.78	92.40 72.20 73.30 70.03 71.61 62.37	C1H C2H C3H C4H C5H C6H
10	Succinate	2.41	34.97	$\alpha,\beta CH_2$	24	Fumarate	6.52	136.10*	$\alpha,\beta C{=}C$
11	Creatine	3.04 3.93	38.50 54.50	$N(CH_3)$ $N(CH_2)$	25	Tyrosine	3.93 3.14 6.89 7.18	57.00 41.92 131.70* 117.70*	αCH βCH C3,5H ring C2,6H ring
12	Choline	3.20 4.07	54.60 56.50	$N^+(CH_3)_3$ $CH_2(OH)$	26	Histidine	7.06 7.78	117.70* 136.70*	C4H C2H ring
13	Arginine	1.91 3.22	30.45 41.40	βCH_2 δCH_2	27	Phenylalanine	3.27 3.90 7.39 7.43 7.33	40.76 56.80 130.33* 130.30* 128.60*	βCH_2 αCH C2,6 ring C3,5 ring C4 ring
14	PE[a]	3.22 4.00	41.10 61.10	CH_2 CH_2					

Figure 4.5: List of ^{1}H and ^{13}C chemical shift (δ, ppm) of metabolites found in ^{1}H-TOCSY and ^{1}H-^{13}C-HSQC-NMR spectra of HCC, metastasis and adjacent non-involved liver tissues. [a] Abbreviations: GPC, glycerophosphocholine; PC, phosphocholine; PE, Phosphoryl-ethanolamine; TMAO: Trimethylamine-N-oxide. * Expected chemical shift.

Principal Component Analysis

Notwithstanding the use of 2D spectra, visual inspection alone did not warrant meaningful observations of the metabolite distribution. To obtain statistically relevant biochemical information from NMR data, we first applied multivariate data analysis based on pattern recognition methods to all spectra by comparing each tissue with the anothers. Therefore, we applied PCA on spectra of NT and CIR in order to evaluate their metabolomic profiles. Figure 4.6 shows the PCA results as scores (Figure 4.6A) and loadings plots (Figure 4.6B) for the first two principal components from spectra of CIR (filled squares, ■) and NT samples (empty squares, □).

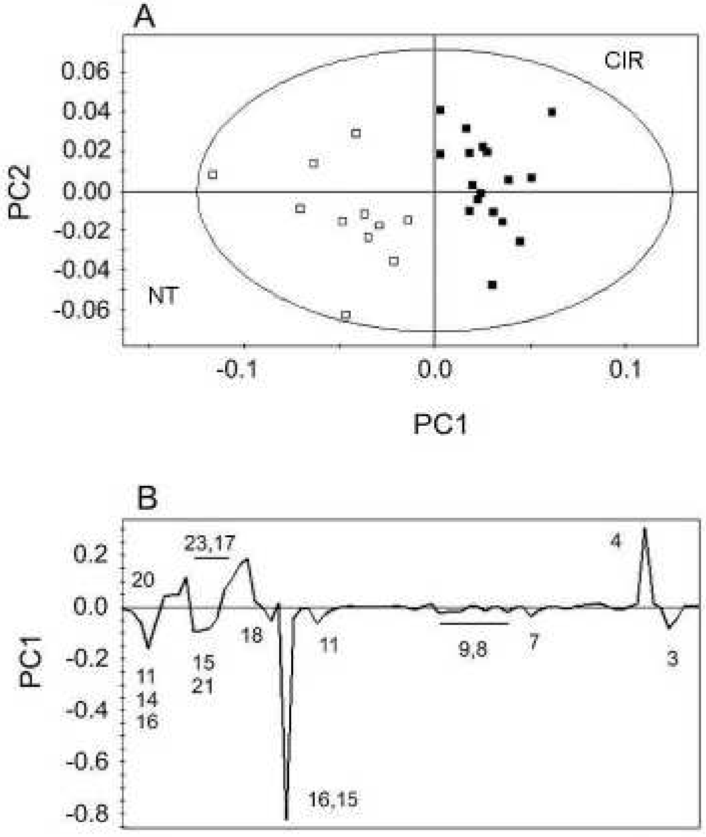


Figure 4.6: PCA comparison of non-tumoral (NT) with cirrhotic tissues (CIR). (A) Scores plot (R^2=73.14%) for CIR (■) and NT (□). The major metabolic signals that differentiate the two classes are shown in the loadings plot (B), where numbers refer to metabolites as labeled in Figure 4.2.

Clustering is observed from the scores plot PC_1 *vs.* PC_2 (Figure 4.6A), where PC_1 and PC_2 explained 73.14% of the total variance within the data. The metabolic signals responsible for the differentiation of the two classes can be identified from loadings plot (Figure 4.6B) associated with the PCA. Compared with NT tissue extracts, CIR showed increased concentrations of lactate (Figure 4.2 for labeling), α-/β-glucose, and glycogen, with decreased concentration of Thr, acetate, Glu, Gln, creatine, PC, GPC, TMAO, and *myo*-Inositol. Applying PCA to the spectra of liver metastasis (■), they resulted separated from those corresponding to non-cirrhotic normal liver (□), as depicted in the scores plot PC_1 *vs.* PC_2, which explains 90.78% of the total variance (Figure 4.7C). The loadings plot in Figure 4.7D shows the major alterations of the metabolic signals responsible for the separation. In particular, metastasis differentiated from the non-cirrhotic normal liver for high level of Leu, Thr, lactate, Ala, acetate, Glu, Gln, Gly, GPC, PE, and *myo*-Inositol, and for lower level of α-/β-glucose and glycogen.

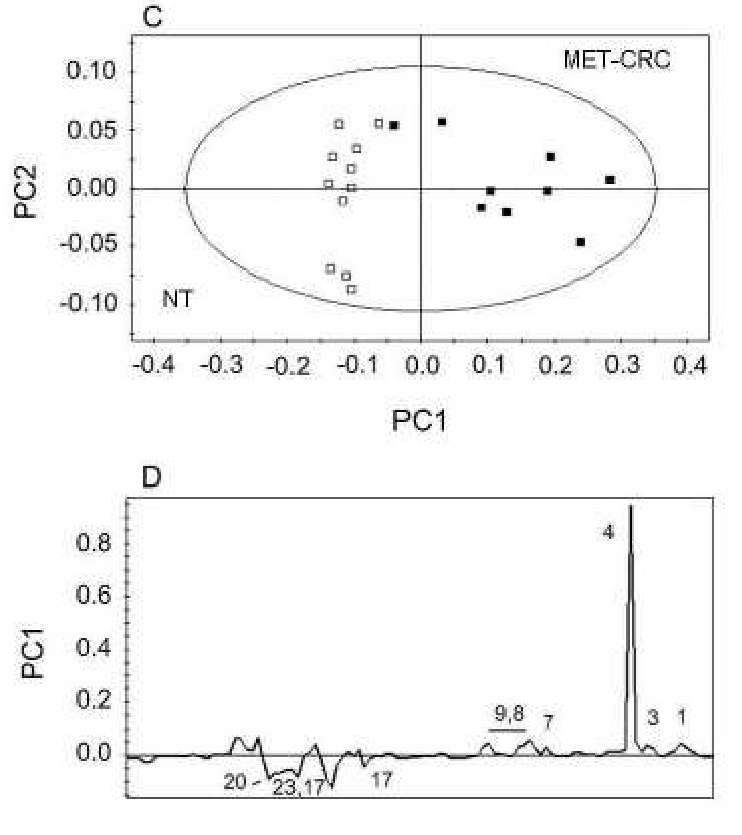


Figure 4.7: PCA comparison of non-tumoral (NT) with metastasis tissues (MET-CRC). The scores plot C (R^2=90.78%) distinctly shows a separation for metastasis (■) and non-cirrhotic (□) tissues along the PC_1 axis. The loadings plot (D) shows the major signals that determined difference in the clustering, numbers refer to metabolites as labeled in Figure 4.2.

As it can be seen in the scores plot (Figure 4.8A), PCA successfully classified HCC tissues (□) from the CIR strains (■) through two PCA components, which explained 70.93% of the variance within the dataset. The separation was due to an increase of Leu, Thr, lactate, Ala, acetate, Glu, Gln, PC+GPC and PE, and to a decrease of creatine, α-/β-glucose and glycogen in HCC (Figure 4.8B).

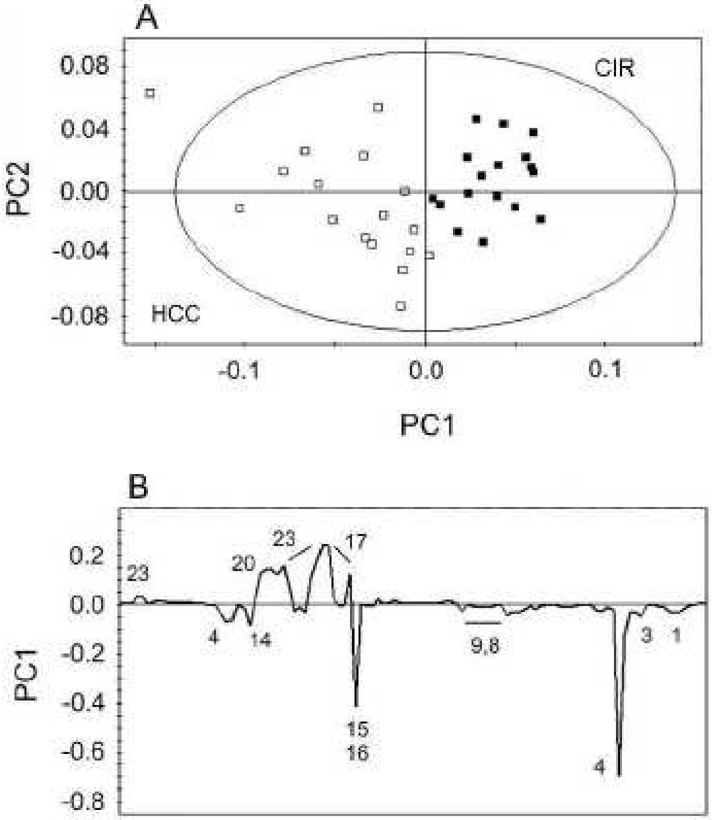


Figure 4.8: PCA comparison of HCC with cirrhotic tissues (CIR). HCC (□) and the corresponding cirrhotic (■) samples separated in the scores plot A (R^2=70.93%) along the PC_1 axis, by means of the loadings plot B. Numbering as in Figure 4.2 caption.

Furthermore, we readily distinguished HCC (■) from metastases (□), as shown by the scores plot PC_1 *vs.* PC_2, where the two components explained 83.79% of the total variance within the data (Figure 4.9C). The associated loadings plot shows differences of the metabolite concentration which determined such clustering (Figure 4.9D). Compared to metastasis, HCC tissues had higher levels of α-/β-glucose and glycogen, with lower levels of Leu, Thr, lactate, acetate, Glu, creatine, TMAO, *myo*-Inositol, Gly, GPC and

PE. Finally, we performed PCA of the whole dataset by extending pattern

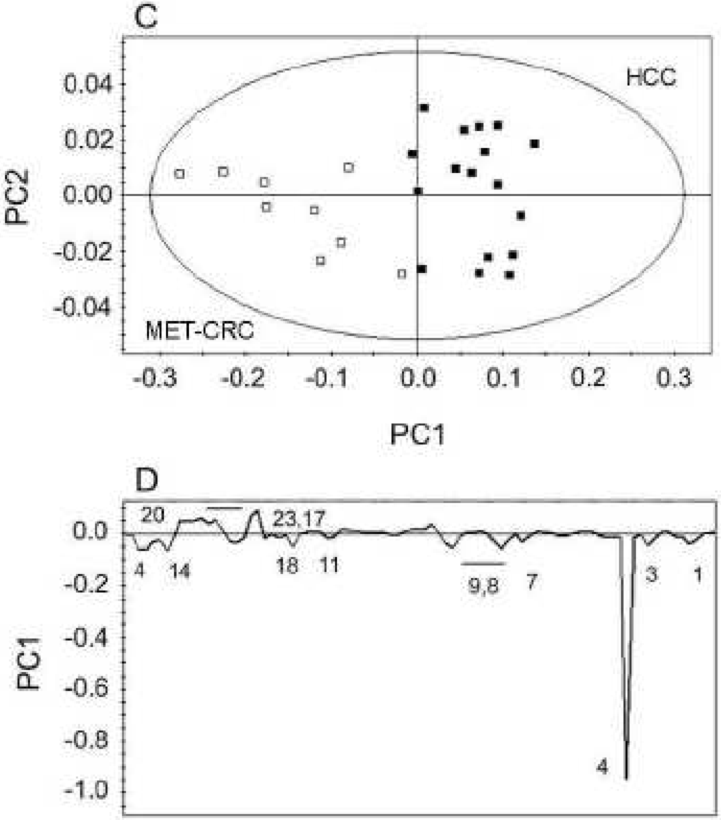


Figure 4.9: PCA comparison of HCC with metastasis tissues (MET-CRC). The scores plot C (R^2=83.79%) displays HCC spectra (■) and metastasis (□) spectra in two clusters along the PC_1 axis according to the signals in the loading plot D, which highlights the signals involved in the clustering. Numbering as in Figure 4.2 caption.

recognition technique to all classes. Figure 4.10 shows the scores plot PC_1 *vs.* PC_2 and explains 77.94% of the total variance.

Although clusterings displayed in Figures 4.6, 4.7, 4.8 and 4.9 clearly separated different pairs of hepatic tissues, the whole model is more controversial as it appears in the scatter plot of Figure 4.10. For that reason we performed an OPLS-DA analysis.

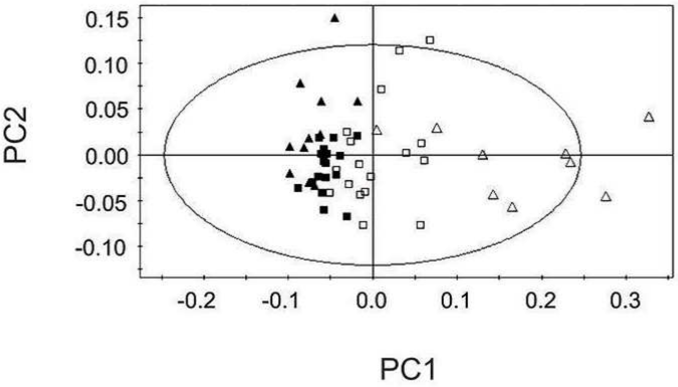


Figure 4.10: PCA showing the metabolic differences within each individual group of tissues, namely NT (▲), CIR (■), HCC (□) and MET-CRC (△).

Orthogonal Projection to Latent Structures Discriminant Analysis

To better construct a four tissue classes model and to understand the role of the X variables ("buckets") in the class separation, and to prove the potential of the NMR representation in assigning new samples to a specific class, we constructed an O2PLS-DA model, which resulted in three predictive and three orthogonal components (R^2=0.65 and Q^2=0.35).

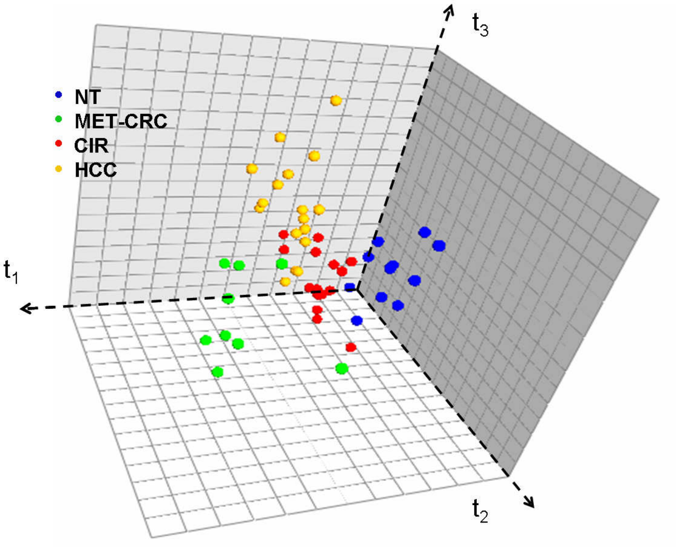


Figure 4.11: 3D score plot showing the class separation of the different group of tissues, namely NT (blue), CIR (red), HCC (yellow) and MET-CRC (green).

In the 3D score plot (Figure 4.11) the four tissue classes appear sufficiently separated in clusters, although the model seems to be robust for the MET-CRC samples (R^2=0.82 and Q^2= 0.63), but weaker for the HCC (R^2=0.55 and Q^2=0.26) and CIR samples (R^2=0.58 and Q^2=0.17). However, the latent structure corresponding to the predictive part of the model can be used to explain the relationships between X-variables and class separation.

The p(corr)/q(corr) plot (Figure 4.12) is a useful tool to identify the variables responsible for the tissues class separation. The $p_i(corr)_j$ parameter is the correlation coefficient between the t_i predictive score vector and the X_j variable, and can be considered as a measure of the similarity between the t_i score vector and the X_j variable. On the other hand, the $q_i(corr)_j$ parameter corresponds to the correlation coefficient between the t_i predictive score vector and the dummy variable representing the class j, and allows its representation in the same plot of the X variables. Figure 4.12 indicates that the first principal component is very similar to variables corresponding to "buckets" 1.34, 4.10, 3.90 and 3.82 ppm. In particular, a progressive increase of the 1.34 ppm variable can be observed starting from the NT class, through the CIR and the HCC up to the MET-CRC samples (Figure 4.13A). On the contrary, the 3.90 ppm variable shows an opposite trend through the four classes (Figure 4.13B).

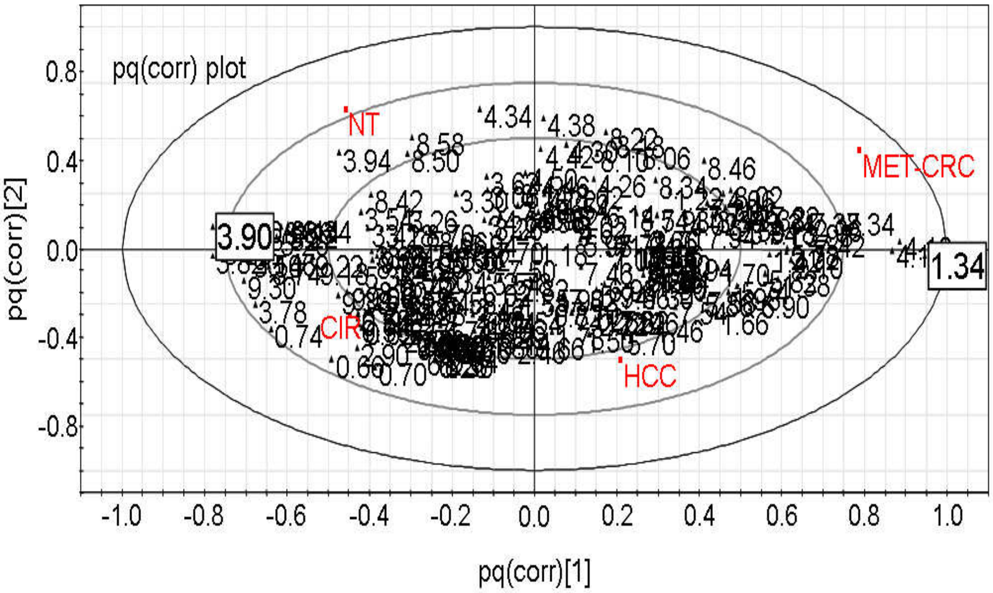


Figure 4.12: Identification of variables responsible for the tissues class separation: pq(corr) plot with all variables ("buckets").

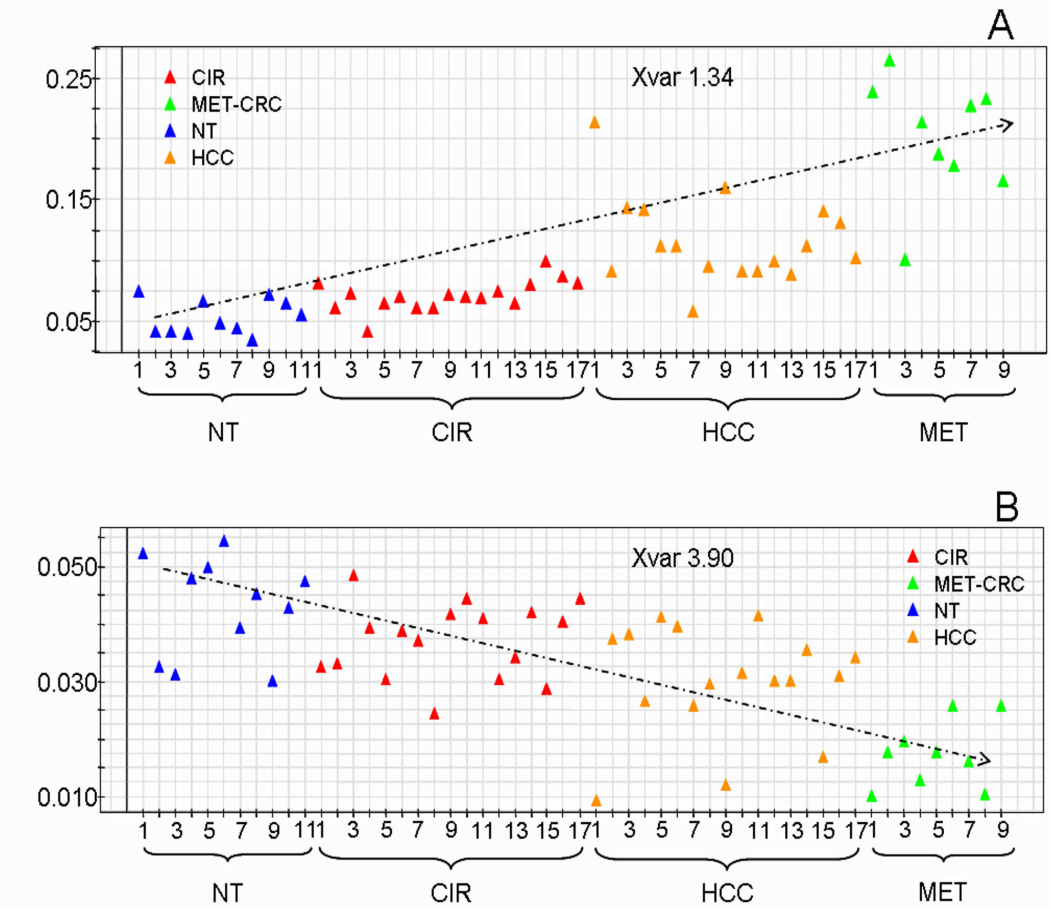


Figure 4.13: Identification of variables responsible for the tissues class separation. (A) and (B) variation of the "buckets" relative to the most significant signals at 1.34 ppm (lactate) and 3.90 ppm (α-glucose), respectively, showing a progressive increase of the 1.34-ppm variable, and a corresponding decrease of the 3.90-ppm variable. Samples are identified by a color code.

In order to build a Naïve Bayes classifier the three predictive score vectors were used to obtain a new representation of the sample space. The prediction performance of the classifier was evaluated by complete cross-validation (four groups). It showed just 7.4% of incorrect prediction (4/54 samples), while 92.6% of samples were correctly predicted (50/54 samples). The four samples were incorrectly classified as belonging to adjacent classes: one NT sample was predicted as CIR (1/11 NT); two CIR samples were predicted as HCC (2/17 CIR), and one HCC was predicted as MET-CRC (1/17). For a two-class model, O2PLS-DA is able to obtain a powerful classification and detect potential markers [14]. In this case, only one component is needed to explain the variation between the two classes, and the predictive score vector t can directly be used to highlight resonances ("buckets") acting as potential markers. This could easily be achieved by building the S-plot, in which p(corr) is plotted against the predictive loading vector p of the model, and only the variables having an absolute p/p_{err} ratio > 1.7 (where p_{err} is the error on p estimated by jack-knife in cross-validation) will be considered.

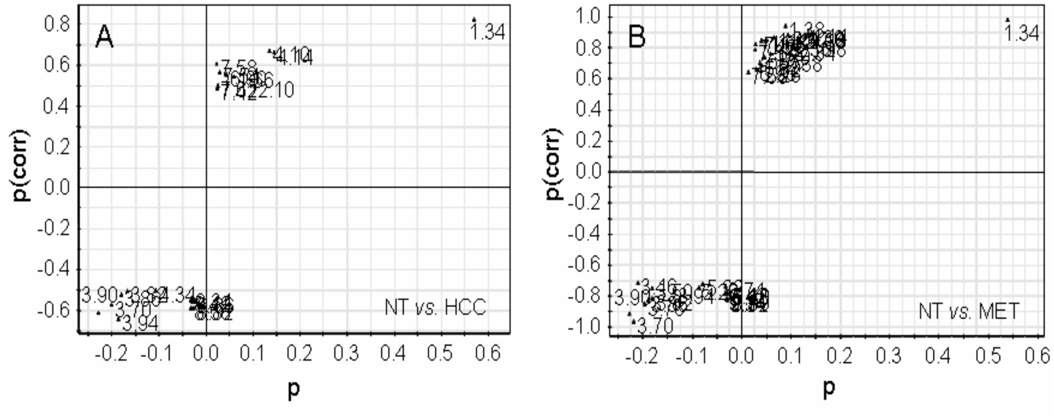


Figure 4.14: S-plots reporting p(corr) against the predictive loading vector p of the model: (A) NT *vs.* HCC; (B) NT *vs.* MET-CRC. All models indicated the signals at 1.34 and 3.90 ppm, as the principal discriminating variables.

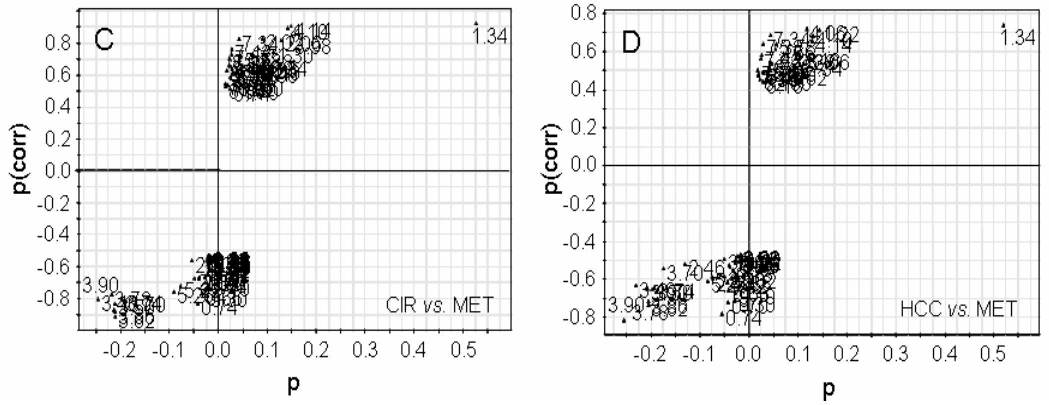
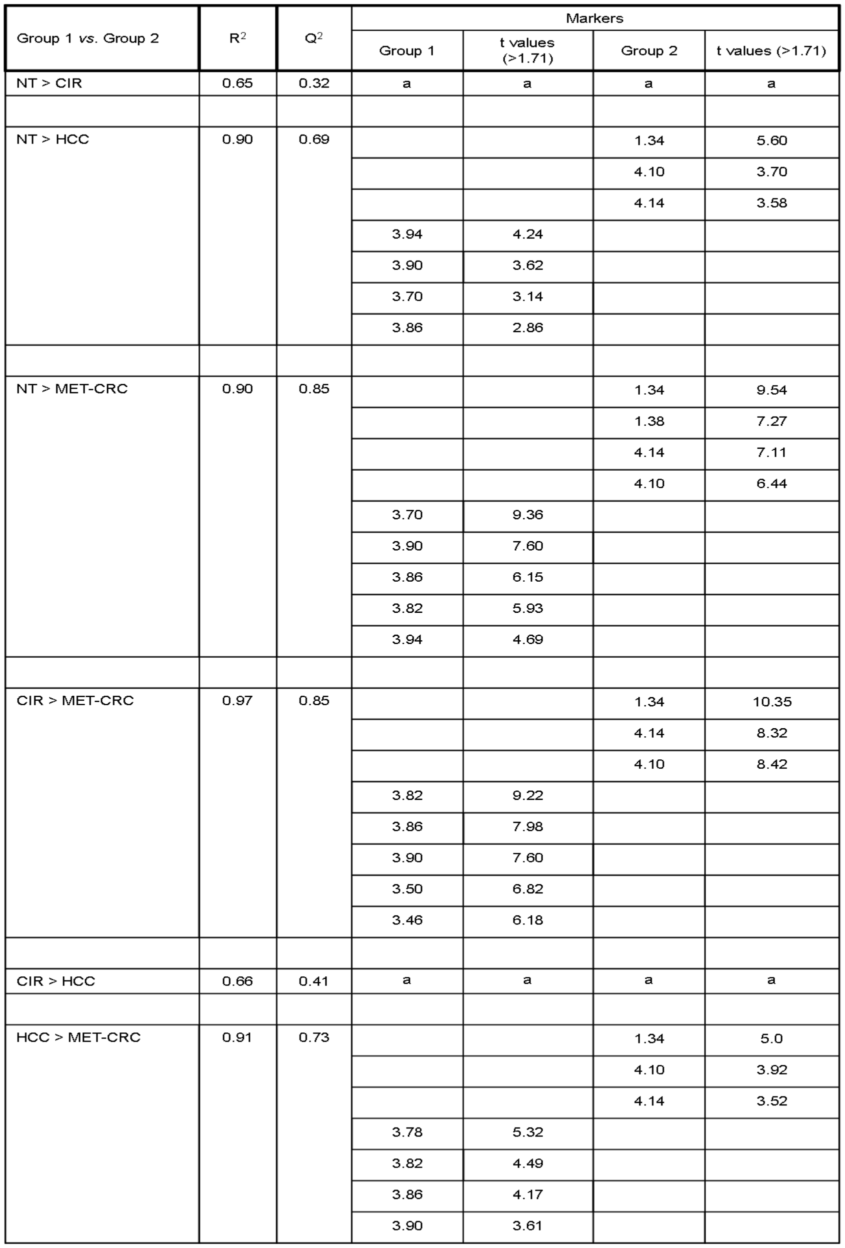


Figure 4.15: S-plots reporting p(corr) against the predictive loading vector p of the model: (C) CIR *vs.* MET-CRC; and (D) HCC *vs.* MET-CRC. All models indicated the signals at 1.34 and 3.90 ppm, as the principal discriminating variables.

Six models were considered, each corresponding to a pair of sample classes. Figure 4.14 shows the S-plots of NT vs. HCC (panel A) and NT *vs.* MET-CRC (panel B) while Figure 4.15 shows the S-plots of CIR *vs.* MET-CRC (panel C) and HCC vs. MET-CRC (panel D). All models indicated the signals at 1.34 and 3.90 ppm, stemming from the lactate and the glucose, respectively, as the principal variables discriminating both MET-CRC and HCC from NT samples, and CIR and HCC from MET-CRC. These models can all be considered robust having high Q^2 values (> 0.69). On the contrary, the NT *vs.* CIR and the CIR *vs.* HCC models did not show any discriminating variable as a putative marker. Table reported in Figure 4.16 summarizes all parameters related to the O2PLS-DA models.



Group 1 *vs*. Group 2	R^2	Q^2	Markers			
			Group 1	t values (>1.71)	Group 2	t values (>1.71)
NT > CIR	0.65	0.32	a	a	a	a
NT > HCC	0.90	0.69			1.34	5.60
					4.10	3.70
					4.14	3.58
			3.94	4.24		
			3.90	3.62		
			3.70	3.14		
			3.86	2.86		
NT > MET-CRC	0.90	0.85			1.34	9.54
					1.38	7.27
					4.14	7.11
					4.10	6.44
			3.70	9.36		
			3.90	7.60		
			3.86	6.15		
			3.82	5.93		
			3.94	4.69		
CIR > MET-CRC	0.97	0.85			1.34	10.35
					4.14	8.32
					4.10	8.42
			3.82	9.22		
			3.86	7.98		
			3.90	7.60		
			3.50	6.82		
			3.46	6.18		
CIR > HCC	0.66	0.41	a	a	a	a
HCC > MET-CRC	0.91	0.73			1.34	5.0
					4.10	3.92
					4.14	3.52
			3.78	5.32		
			3.82	4.49		
			3.86	4.17		
			3.90	3.61		

Figure 4.16: Summary of O2PLS-DA parameters from the six pairs of models analyzed. [a] No discriminating variables were found as a putative marker.

If a particular class can be considered as a control, it is possible to gain information about the variables that discriminate each class, with respect to the control, using the so called SUS-plot (Shared and Unique Structure plot). Assuming the NT samples as control, the p(corr) vectors estimated for each two classes models, separately including the NT class, were used to represent the X-variables in the SUS-plot (Figure 4.17). Since the NT *vs.* CIR model was not robust enough to be understood in terms of single variables, we limited our analysis to NT, HCC and MET-CRC classes. We found that the same signals separate both HCC and MET-CRC samples from the control, while no unique signals discriminate these two classes. In particular, the buckets located at 1.30-1.38 ppm and 4.00-4.14 ppm; which contain the lactate signals, are elevated in both HCC and MET-CRC classes, suggesting the lactate as the putative marker. On the contrary, the buckets at 3.70-4.00 ppm, containing the glucose signals, are prominent in NT class, suggesting the glucose as the putative marker. Therefore, both metabolites primarily contribute to the classification of the different groups, showing an opposite trend among the groups. In particular, the lactate level increases from NT group, through CIR and HCC, to reach the highest value in the liver MET. On the contrary, the signals of glucose progressively decrease from NT group, through CIR, HCC and MET-CRC group, which shows the lowest intensity.

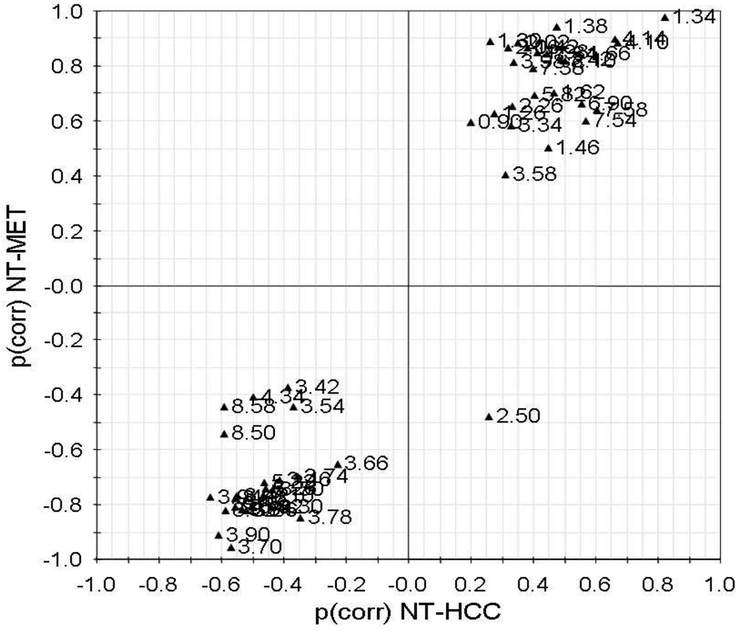


Figure 4.17: SUS plot of NT, HCC, and MET-CRC classes. Assuming the NT samples as control, the p(corr) vectors estimated for each two classes models, separately including the NT class, were used to represent the X-variables.

Quantification and statistical significance

To confirm the parallel trend of these two putative markers (increased lactate and decreased glucose), we integrated the ^{1}H-NMR isolated signals of lactate (βCH_3, 1.33 ppm) and α-glucose (C1H, 5.24 ppm) in all tissue samples. We only considered the α-glucose, which represents *ca.* 36% of total glucose, because the remaining 64%, corresponding to the β form, gives a signal at 4.65 ppm, and as such it is strongly perturbed by the pulse sequence used for water peak (4.68 ppm) suppression in the NMR experiments. The peak area of lactate and α-glucose was scaled to the molar concentration taking into account that they represent the lactate methyl group and the glucose isomer, and calculated the lactate/glucose molar ratio. Figure 4.18 illustrates the lactate/glucose molar ratio for each patient sample. The analysis of variance (ANOVA with Bonferroni correction) has been applied, and statistically significant differences were observed for the lactate/glucose ratio of NT vs. MET-CRC ($p < 0.001$), CIR vs. MET-CRC ($p < 0.001$) and HCC vs. MET-CRC ($p < 0.001$).

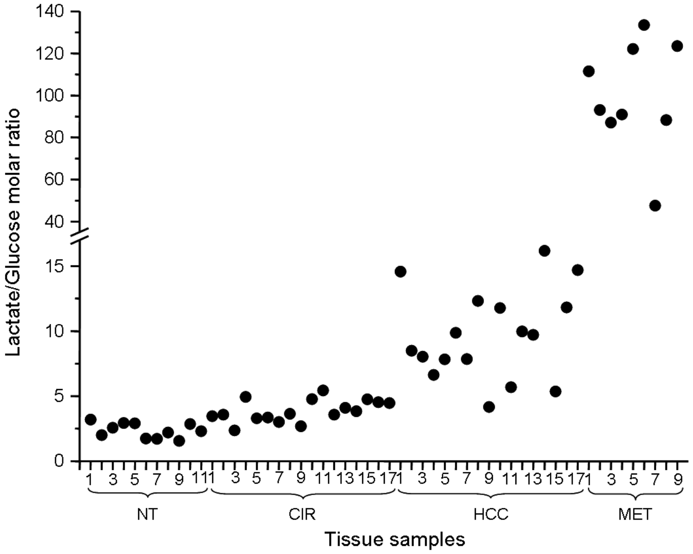


Figure 4.18: Lactate-α-glucose molar ratio for each patient sample. Statistically significant differences were observed for the lactate/glucose ratio of NT *vs.* MET-CRC ($p < 0.001$), CIR *vs.* MET-CRC ($p < 0.001$) and HCC *vs.* MET-CRC ($p < 0.001$). The vertical axis has been cut to highlight the variations for NT, CIR and HCC samples, all with a ratio <15.

4.3 Discussion

In this study we have used high-resolution ^{1}H-NMR spectroscopy to investigate the metabolite composition of human hepatic tissue extracts of 17 patients affected by hepatocellular carcinoma HCV-related (HCC), and 9 patients affected by liver metastases from colorectal carcinoma (MET-CRC). As a control we used cirrhotic liver tissues of HCC patients (CIR) and normal liver tissue of MET-CRC patients (NT), respectively. All spectral classes were visualized by PCA analysis, which also highlighted the "evolution" and relationship of the different pathological liver conditions represented by the four NMR data classes. The disease evolution is established along the PC_1 axis (Figure 4.12A), following the increase of the lactate (Figure 4.13B), and the remarkable decrease of glucose (Figure 4.13C). The progressive increase of lactate/glucose ratio along the PC1 axis is consistent with the enhanced conversion of glucose into lactate, through the different classes that represent different tissue conditions such as hypoxia and/or "aerobic glycolisis". Solid malignant tumors are characterized by pronounced tissue hypoxia [103] and enhanced formation of lactate [104], but many tumors exhibit a strong generation of lactate even in the presence of oxygen. This phenomenon, known as "aerobic glycolysis" or the "Warburg effect" [105], is generally considered the result of oncogenic alteration in glucose metabolism following malignant transformation [106], but its significance is still controversial [107]. An elevated lactate concentration in primary lesions at first diagnosis has been related to an increased risk of metastases in squamous cell carcinomas of the uterine cervix, of the head and neck, and in adenocarcinomas of the rectum [108]. Certainly no endogenous marker alone is able to predict the hypoxic status of the tumor, and we need to find, within hypoxic metabolic profiles, a pattern of signals (metabolites) that are expression of the pathological changes. However, our observations suggest that the metabolic shift towards enhanced glycolysis would already be present in the early stage, during multi-step hepatic tumorigenesis. Starting from liver cirrhosis, widely considered as precancerous lesions, the upregulation of glycolysis showed progressive rate of conversion in different hepatic conditions, thus indicating the metastasis group as the one, among all classes, requiring the larger amount of conversion in energy for its malignancy characteristics. Most probably, cell population with upregulated glycolysis could develop growth advantages which promote unconstrained proliferation and invasion [106].

The PCA analysis of variables shows that PC_1 separates NT from MET-CRC, and CIR from HCC, while PC_2 separates NT from CIR, and MET-CRC from HCC. These separations ideally identify two different "metabolic developmental trajectories", which, based on the changes in the NMR-visible metabolome, describe liver tumorigenesis (Figure 4.12). Starting from NT, it is possible to ideally draw an ideal line through CIR to HCC, according to a sequel of pathological liver alterations. Conversely, it is possible to connect NT directly to MET-CRC, according to the absence of any liver "intermediate" state. It is worth speculating about the possible applications of such metabolic trajectories. Firstly, the trajectory could be used to identify a specific pathological state by verifying when candidate metabolites deviate from the normal path. This could then be correlated with known morphological events providing insight into the progression towards HCC. Furthermore, the trajectory could define the point of HCC tumorigenesis where a limited number of genomic (DNA microarray) and/or proteomic studies could be carried out to better characterize the oncogenic changes. Secondly, comparison of metabolic trajectories can provide a suitable way to distinguish primary tumors from metastases. Thirdly, the effects of drug treatment could be assessed by determining if the pathological metabolic trajectory tends to the "normal" state. On this regard, the ^{1}H-NMR spectra provided quantitative data by integrating selected metabolite signals that were found to primarily contribute to the classification of the different groups. In particular, we identified the lactate/glucose ratio, which shows an opposite trend among subgroups and within each of them, therefore affording a reliable method for evaluating healthy or non-healthy status of the liver.

In this study the patients who developed HCC were also affected by chronic cirrhosis HCV-related. Hepatites C infection is the most frequent liver infection and is considered a pre-cancerous lesion of liver. HCV infection is also associated with an increased risk of glucose intolerance and diabetes maybe due to an impaired glucose homeostasis mediated directly by HCV proteins. Liver cirrhosis is a progressive fibrotic process that is characterized by the final necrosis of hepatocytes. In normal conditions, after carbohydrate digestion, blood glucose level rises, and in hepatocytes insulin acts so as to stimulate several enzymes and convert excess glucose into glycogen, thus preventing excessive osmotic pressure build up inside the cell. In fact, CIR samples (Figure 4.2B), compared to NT samples (Figure 4.2A), show an increased amount of lactate, and the lactate/glucose ratio is *ca.* 2 times that in NT (Figure 4.18). Hepatic transformations occur by sequential accumulation of genetic and molecular alterations, and HCC is often the result of a slow and progressive evolution going through the development of liver cirrhosis. The lactate in HCC samples is *ca.* 2 times higher than that in CIR samples, meaning that there is an alteration of the carbohydrate metabolism, with enhanced glycolysis and alteration of the tricarboxylic acid (TCA) cycle [15].

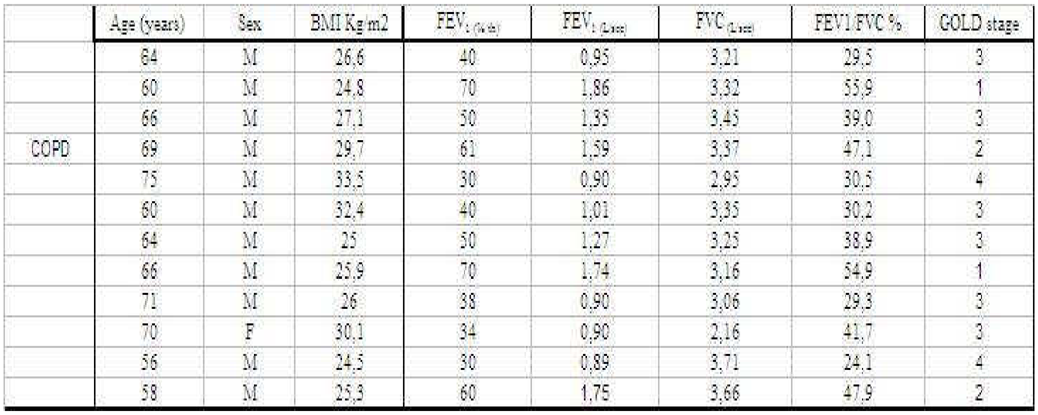
Metastasis formation is the result of a multi-step cascade of events occurring to cancer cells during tumor dissemination, which brings about considerable metabolic changes [109]. The large increase in lactate concentration as well as the decrease of intracellular glucose level was the predominant effect for the separation of metastases from HCC and NT (Figure 4.17), and the lactate/glucose ratio in MET-CRC ranges from 9 to 40 fold higher compared to HCC and NT, respectively (Figure 4.18), thus suggesting a role for the enhanced phenomenon of "aerobic glycolysis". Furthermore, the metastatic process for remodeling and altering extra-cellular matrix, tightly associated with cell proliferation, is consistent with the elevation of lactate, and has been already reported for metastasis in axillary lymph nodes in breast cancer and human cervical cancer [110].

The approach used in this study highlighted metabolic evolution of different liver diseases: cirrhosis, HCC, and liver metastasis. The analysis of such a wide range of specimen types indicated that the common discriminating factor, a progressive increase of lactate concentration, is coupled with changes in TCA cycle and alterations of the energy metabolism in the liver of CIR and HCC patients HCV-related. In addition, the raise of lactate is also coupled with a stronger elevation of lactate/glucose ratio of patients MET-CRC may be due to other metabolic mechanisms. In previous HR-MAS studies on intact tissues, the lactate resonance was discarded for possible anaerobic degradation of glucose induced during surgery or experiment [111]. Here all samples underwent the same treatment, and therefore we can safely exclude external factors altering the lactate levels. Furthermore, the dual extraction procedure used in our study allowed identification and quantification a much higher number of polar metabolites in comparison with protocols previously described for the NMR spectroscopy on intact tissues ex vivo [112].

4.4 Materials and methods: b) exhaled breath condensate

Specimens collection

A total of 36 paired EBC and saliva samples were collected from the following groups of subjects: 12 healthy subjects (HS; nine males, mean age 55.6±7.2 yrs); 12 laryngectomized patients (nine males, mean age 60.2±6.2 yrs); and 12 patients affected by chronic obstructive pulmonary disease (COPD; 11 males, mean age 64.9±5.7 yrs). All HS were nonsmokers, while the laryngectomized patients (who provided samples through a stoma, bypassing the pharynx entirely) and the COPD patients were ex-smokers (at least 24 months since smoking). All subjects presented no occupational or other pronounced exposure to organic solvents. The laryngectomized patients had been previously treated by laryngectomy for laryngeal carcinoma for at least one year prior (range 12-18 months) and did not have a history of chronic respiratory disease or recurrent exacerbations. COPD patients had received diagnosis in the past according to the Global Initiative for Chronic Obstructive Lung Disease guidelines [113]. The COPD anthropometric characteristics are summarized in table in Figure 4.19.



	Age (years)	Sex	BMI Kg/m2	FEV$_{1\ (\%\ th)}$	FEV$_{1\ (L/sec)}$	FVC$_{(L/sec)}$	FEV1/FVC %	GOLD stage
COPD	64	M	26,6	40	0,95	3,21	29,5	3
	60	M	24,8	70	1,86	3,32	55,9	1
	66	M	27,1	50	1,35	3,45	39,0	3
	69	M	29,7	61	1,59	3,37	47,1	2
	75	M	33,5	30	0,90	2,95	30,5	4
	60	M	32,4	40	1,01	3,35	30,2	3
	64	M	25	50	1,27	3,25	38,9	3
	66	M	25,9	70	1,74	3,16	54,9	1
	71	M	26	38	0,90	3,06	29,3	3
	70	F	30,1	34	0,90	2,16	41,7	3
	56	M	24,5	30	0,89	3,71	24,1	4
	58	M	25,3	60	1,75	3,66	47,9	2

Figure 4.19: Anthropometric characteristics of 12 patients affected by chronic obstructive pulmonary disease. BIM: body mass index; FEV1: forced expiratory volume in one second; % pred: predicted; FVC: Forced vital capacity; GOLD: Global Initiative for Chronic Obstructive Lung Disease; M: male; F: female. FEV1, FVC and FEV1/FVC were measured after bronchodilatation inhalation test.

None of the patients were on regular systemic or inhaled corticosteroid treatment. They were asked not to use long-acting β2-agonist and anticholinergic agents for a period longer than 12 h and 24 h, respectively, before EBC collection. All subjects were free from upper and/or lower airway infection for, at least, 4 weeks before the EBC collection. They refrained from food intake for 4 h before the test and from alcoholic drinks for 18 h before EBC collection. In laryngectomized patients, lower respiratory tract secretions were actively managed by selfsuctioning and cleaning before each EBC collection.

To assess within-day repeatability, eight subjects (four HS and four COPD patients) were asked to collect EBC and saliva twice within the same day (at times 0 h and 12 h). All subjects gave informed consent and the study protocol was approved by the Ethics Committee of the Monaldi Hospital (Naples, Italy).

EBC sampling

EBC was collected using an EcoScreen condenser (Jaeger, Wurzburg, Germany) as previously described [40] (Figure 4.20). Briefly, all subjects breathed through a mouthpiece (laryngectomized patients provided samples through the stoma) and a two-way nonrebreathing valve, which also served as a saliva trap, at normal frequency and tidal volume, while sitting comfortably and wearing a nose-clip for a period of 15 min. They maintained a dry mouth during collection by periodically swallowing excess saliva.

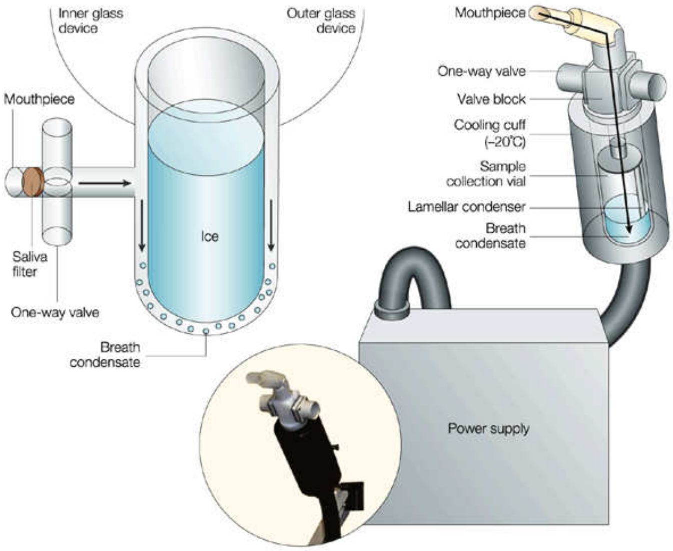


Figure 4.20: EBC schematic collecting system.

Condensate samples (3-4 ml) were immediately transferred into glass vials of 10 ml volume, closed with 20 mm butyl rubber lined with polytetrafluoroethylene septa, and crimped with perforated aluminium seals. Volatile substances, possibly deriving from extra-pulmonary sources [114, 115, 116], were removed by a gentle stream of nitrogen before sealing. Nitrogen was applied for a variable time (1, 3, 5, 10, 15 and 20 min); no difference was observed with spectra acquired after 1 min nitrogen exposure, but since such an interval appeared to be too short to avoid systematic errors, a 3 min interval was chosen. Nitrogen was used because the concentration of volatile solutes in EBC is dependent on their distribution between the saliva, exhaled air and droplets, and the condensate. This distribution can be altered by multiple factors, including minute ventilation, salivary pH, solubility, temperature and sample preparation [117]. Therefore, spectral differences may depend upon uncontrollable variables that prevent reliable quantification. The nitrogen stream also removes oxygen from solutions. Such a procedure, used for NMR protein structure determination [118], together with freezing of sealed samples in liquid nitrogen, immediately "quenches" metabolism at the collection time, and prevents any metabolic decay [37]. Samples were then stored at -80 °C until NMR analysis. Drying of the samples was avoided to circumvent irreversible solute precipitation and/or formation of insoluble aggregates, which were observed upon dissolving the dried condensate for NMR measurements.

Pre-analytical preparation of EBC condenser reusable parts

Before and after collection of each EBC sample, the reusable parts of the condenser (valve, salivary trap and lamellar condenser) were disinfected for 15 min using a solution of a 1.5% freshly prepared chemical agent (Descogen^{TM}; FILT GmbH, Berlin, Germany), and repeatedly flushed with water following the manufacturer's guidelines. To completely eliminate the disinfectant, parts already disinfected and washed were thoroughly rinsed for 15 min with pure grade ethanol (96%), thereafter exhaustively soaked with deionized distilled water for 15 min and dried under vacuum at 50 °C.

Salivary collection

Together with EBC collection, a salivary sample was taken in the same day. To avoid any interference from exogenous agents into the oral environment, the patients were asked to collect all saliva available ($\sim$ 2-4 ml), *i.e.* "whole" saliva expectorated from the mouth, into a plastic universal tube immediately after waking in the morning. As previously described by Silwood *et al.* [38], each patient was requested to refrain completely from oral activities (*i.e.* eating, drinking, tooth brushing, oral rinsing,smoking, etc.) during the short period between awakening and sample collection (<5 min). Each collection tube contained 15 μmol sodium fluoride, sufficient to ensure that metabolites were not generated or consumed *via* the actions of bacteria or bacterial enzymes present in whole saliva during periods of sample preparation and/or storage [39]. Specimens were transported to the laboratory on ice and immediately centrifuged (at 20,000×g at 4 °C for 15 min) on their arrival to remove cells and debris. Following this, a gentle nitrogen gas flow was applied for ~5 min to supernatants, which were then stored at -80 °C until measurements were made.

The ^{1}H-NMR profiles of salivary supernatant specimens subjected to analysis immediately after collection into the fluoride-containing tubes and rapid centrifugation were compared with those of the same samples stored as described previously, and no differences were discernible, *i.e.* none of the criteria investigated changed significantly during these periods of storage.

Sample preparation for NMR analysis

EBC samples were rapidly defrosted. To provide a field frequency lock, 70 μl of a deuterium oxide (D_2O) solution, containing 1 mM sodium 3-trimethylsilyl (2,2,3,3-$^2H_4$) propionate (TSP) as a chemical shift reference for ^{1}H spectra and sodium azide at 3 mM, was added to 630 μl of condensate, thus making 700 μl total volume. Saliva samples were rapidly defrosted and 70 μl of reference standard solution (D_2O-TSP) was added to 630 μl of sample.

NMR measurements

1D spectra were recorded on a Bruker Avance spectrometer (Bruker BioSpin GmbH, Rheinstetten, Germany) operating at a frequency of 600.13 MHz (^{1}H) and equipped with a TCI CryoProbe^{TM} (Bruker BioSpin GmbH), at a probe temperature of 27 °C. The water resonance was suppressed by using the noesypresat pulse sequence, called noesypr1d according to the manufacturers. It has the form - RD-90°-t-90°-t_m-ACQ, where RD is a relaxation delay, t a short delay, 90° represents a 90° radio frequency pulse, t_m the mixing time and ACQ the data acquisition period. In the present study acquisition conditions, the carrier frequency (O_1) value was set on the water resonance, the saturation power was 62 dB, t was 4 μs, t_m was 100 ms, the spectral amplitude was 7002.8 Hz, the time domain was 16 K, RD was 2.0 s and the number of transients was 256. This resulted in a total acquisition time of 14 min per sample. For processing, a line broadening of 0.6 Hz was applied and a real spectrum size of 32 K was used. Spectra were referred to TSP, assumed to resonate at a δ of 0.00 ppm.

Statistical analysis

High-resolution ^{1}H-NMR spectra were automatically data reduced to 200 integral segments ("buckets"), each of 0.02 ppm, using the AMIX software package (Bruker BioSpin GmbH). The resulting integrated regions were imported into the SIMCA package (Umetrics, Umea, Sweden) and used for statistical analysis and pattern recognition. Before pattern recognition analysis, each integral region is usually normalized to the sum of all integral regions of each spectrum; however, because of the presence of contaminant peaks, each bucket was normalized to the TSP peak of known concentration for a reference region of between 0.014 and -0.014 ppm. The correctness of the approach was tested by comparing the results with those obtained by referring to the sum of all integral regions of each contaminant free spectrum. No significant difference was observed between the two approaches; therefore, pattern recognition analysis was reliable with normalization to TSP. Data were preprocessed with the Centering scaling and then processed with PCA and partial least squares discriminant analysis (PLS-DA).

4.5 Results

Spectral differences between EBC and saliva

Figure 4.21 represents spectra of saliva (Fig. 4.21a, b and c) and EBC samples (Fig. 4.21d, e and f) from one HS (Fig. 4.21a and d), one laryngectomized patient (Fig. 4.21b and e) and one COPD patient (Fig. 4.21c and f). Saliva spectra were highly different from corresponding EBC samples and were notably dissimilar between patients: a visual examination establishes a correspondence between spectra from a HS (Fig. 4.21a) and a laryngectomized patient (Fig. 4.21b), but a difference from the COPD spectrum (Fig. 4.21c), which shows sharper lines. The most intense signals in the 0.0-3.2 ppm region of saliva were assigned according to previous studies [38, 101]. Resonance assignment was as follows: leucine δCH_{3s} (triplet) at 0.96 ppm; propionate βCH_3 at 1.04 ppm (triplet) and αCH_2 at 2.19 ppm (quartet); lactate βCH_3 at 1.32 ppm (doublet) and αCH at 4.11 ppm (quartet); threonine γCH_3 at 1.36 ppm (doublet); alanine βCH_3 at 1.47 ppm (doublet) and αCH at 4.20 ppm (quartet); acetate βCH_3 (singlet) at 1.93 ppm; βCH_2 of glutamate and glutamine at 2.10 ppm (multiplet); βCH_3 of pyruvate at 2.37 ppm (singlet); $\alpha,\beta CH_2$ of succinate at 2.41 ppm (singlet); εCH_2 of lysine at 3.06 (triplet); N-CH_{3s} of choline at 3.16 ppm and of phosphorylcholine at 3.23 ppm (both singlets); and N-CH_3 of taurine at 3.23 ppm (triplet).

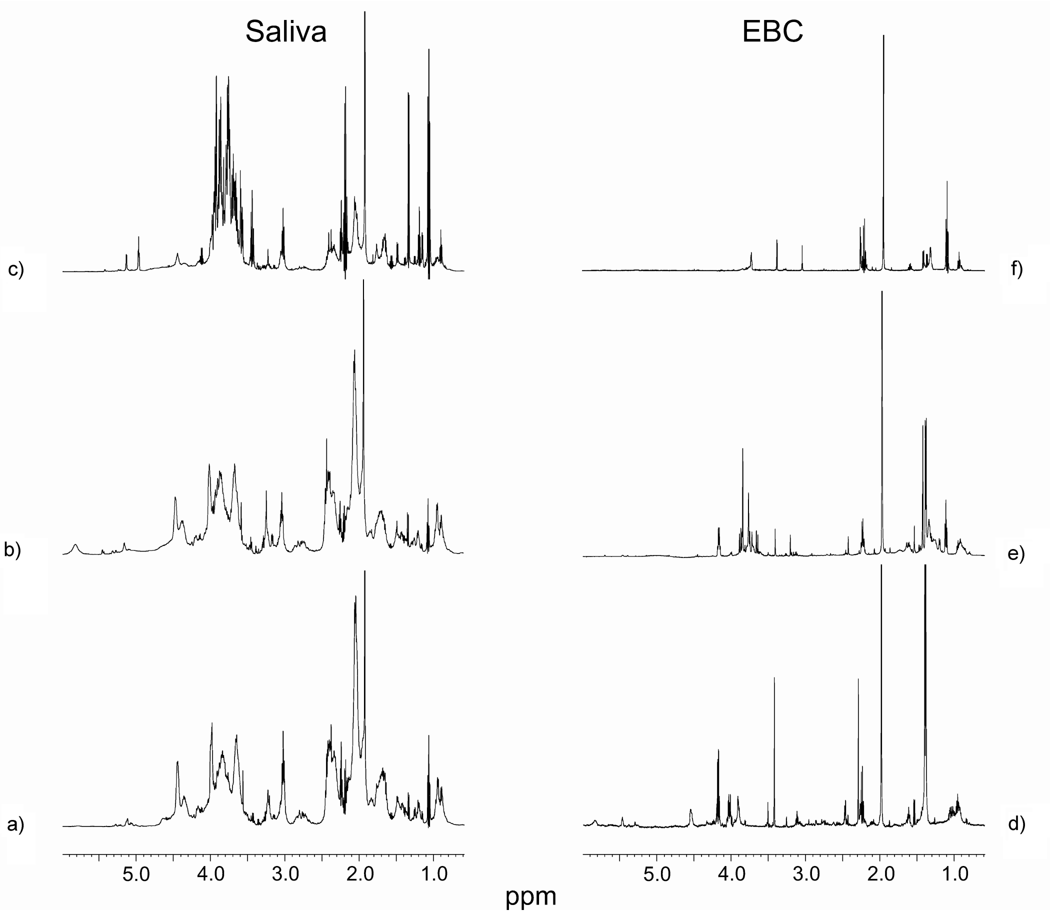


Figure 4.21: Representative one-dimensional ^{1}H-NMR spectra of saliva (a, b and c) and exhaled breath condensate (EBC; d, e and f) samples from healthy (a and d), laryngectomized (b and e) and chronic obstructive pulmonary disease (c and f) patients. The group of signals centered at 3.8 ppm in saliva spectra originates from carbohydrates and is not visible in the corresponding EBC spectra.

Signals between 3.3 and 6.0 ppm originate from carbohydrates and were virtually absent in the EBC spectra. Compared with saliva, EBC spectra presented fewer signals and, as observed for saliva, the COPD patient trace (Fig. 4.21f) appeared to be different from the HS (Fig. 4.21d) and laryngectomized patient (Fig. 4.21e) traces. Spectral differences between saliva and EBC were verified by PLS-DA analysis. Due to the complete absence of the carbohydrate signals in the EBC spectrum, the region 5.0 to 3.5 ppm was cut out from all spectra, partitioning the region between 3.5 and 0.8 ppm. Figure 4.22 shows the score plots of saliva and EBC samples from all subjects. Considering two PLS-DA components, it was possible to obtain a sample classification of 95% (samples correctly classified into different regions). In particular, while EBC samples were all clustered, the saliva samples of HS, laryngectomized and COPD patients were positioned differently from EBC and from each other. Such a separation comes mostly from signals resonating within the 3.5-2.9 and 2.1-1.7 ppm regions. EBC and saliva samples collected from eight subjects twice within the same day (at times 0 h and 12 h) demonstrated good within-day repeatability, showing no evident difference in resonances in the spectra.

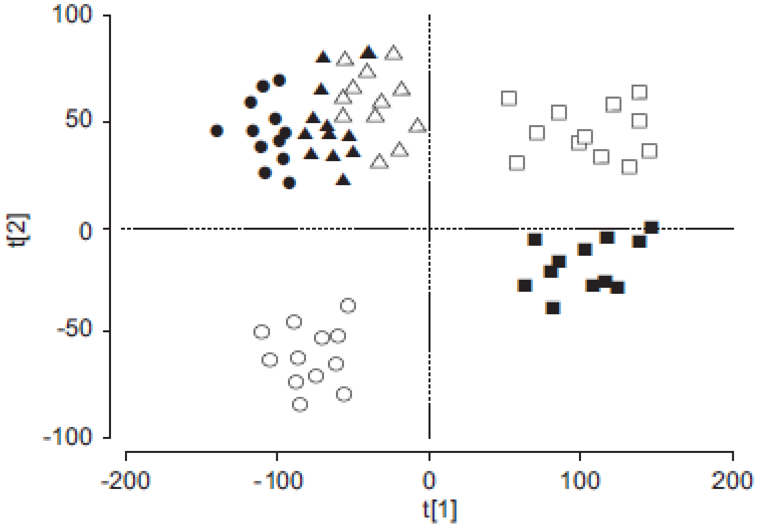


Figure 4.22: Partial least squares discriminant analysis (PLS-DA) scores discrimination for exhaled breath condensate (△: laryngectomized patients; ▲: healthy subjects (HS); •: chronic obstructive pulmonary disease (COPD) patients)and saliva (□: laryngectomized; ■: HS; ○: COPD). All variables were used and two PLS-DA components were retained in the model, obtaining a classification of ∼95%. The region 5.0 to 3.5 ppm, containing the carbohydrate signals, was cut out from the bucketing, and only the signals between 3.5 and 0.8 ppm were analyzed. t[1] and t[2] are the first two principal components.

Effects of disinfectant contamination on EBC spectra

Figure 4.23 shows the ^{1}H-NMR spectrum of Descogen^{TM} (Fig. 4.23a) with representative spectra of EBC samples contaminated by the disinfectant because of insufficient washing time (Fig. 4.23b and c). To completely eliminate the disinfectant, parts already disinfected and washed were thoroughly rinsed for 15 min with pure grade ethanol (96%), thereafter exhaustively soaked with deionized distilled water for 15 min and dried under vacuum at 50 °C (Fig. 4.23d).

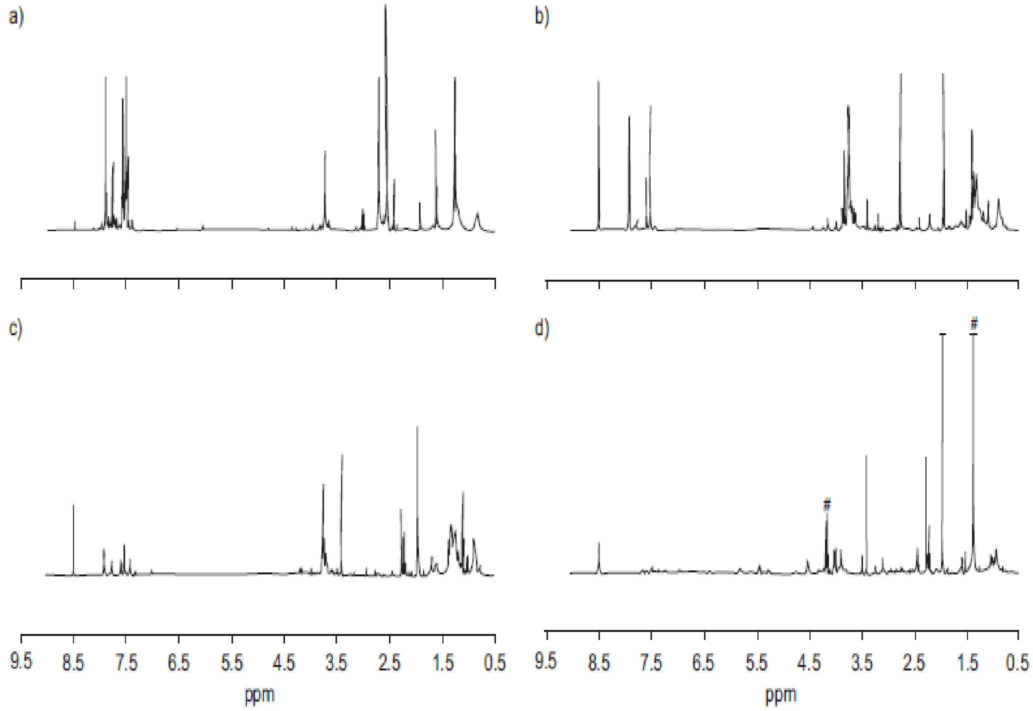


Figure 4.23: Contamination of exhaled breath condensate (EBC) samples by Descogen^{TM} (FILT GmbH, Berlin, Germany). a) ^{1}H-nuclear magnetic resonance spectrum of Descogen^{TM}, compared with b) spectra of EBC samples after partial washing (15 min), and c) intense water rinsing (30 min). d) Contamination was completely removed after the washing procedure using ethanol. The acetate signal at 1.93 ppm was cut in all EBC spectra. a) The vertical scale is one quarter the size of the other spectra. #: lactate resonances.

The resonances of the "saline" components of the disinfectant (citric acid, at 2.66 ppm in the Descogen^{TM} spectrum (Fig. 4.23a), and pentapotassium bis(peroxymonosulphate) bis(sulphate), highly soluble in water) disappeared completely after partial washing (15 min; Fig. 4.23b). However, minor unknown components, such as those giving signals in the 8.2-7.3 and 1.3- 0.7

ppm regions and the signal located at 3.2 ppm, appeared to be more persistent even after intense water rinsing (30 min; Fig. 4.23c). They were completely removed only after the washing procedure using ethanol (Fig. 4.23d). As the perturbation induced by the disinfectant contamination of EBC samples showed visible signals, two different contaminated sets of 12 EBC samples from all COPD patients were examined after partial washing (15 min, "high Descogen^{TM}"; Fig. 4.23b); and after intense water rinsing (30 min, "low Descogen^{TM}"; Fig. 4.23c). Since the region 8.5-7.0 ppm was absent in the "cleaned" EBC spectrum (Fig. 4.23d), as suggested by Carraro *et al.* [40], the region 4.5 to 0.5 ppm was used and the lactate signals were excluded (Fig. 4.23d). Considering two PLS-DA components, a classification of ~72% was obtained, with high-Descogen^{TM} and low- Descogen^{TM} EBC samples classified in two wide regions (Fig. 4.24). This suggests that the presence of the disinfectant at variable concentration affects the interpretation and the statistical analysis of the samples. However, if the presence of contaminant is ignored by a careful selection of the spectral regions to be used for statistical analysis, it is possible to correctly classify the samples. In fact, by selecting only the Descogen^{TM}- free regions of the spectra (3.5-2.9 and 2.1-1.7 ppm), all the samples could be correctly classified.

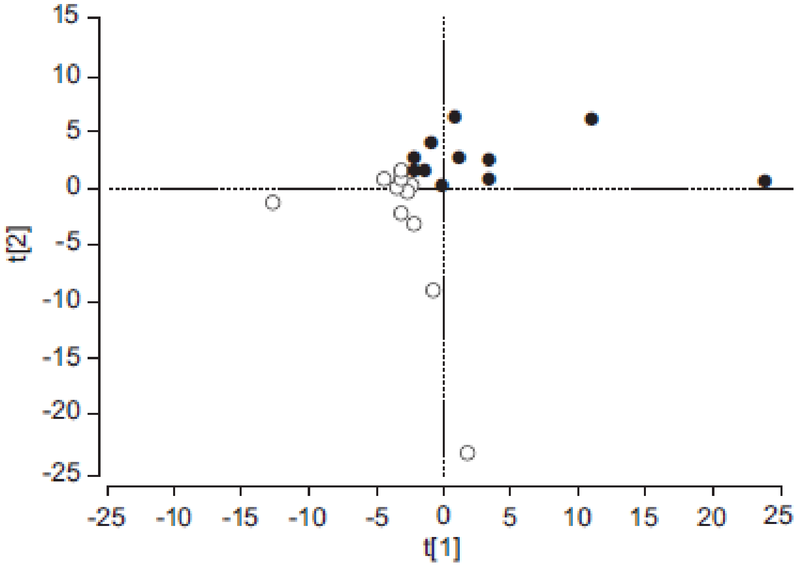


Figure 4.24: Partial least squares discriminant analysis scores discrimination for contaminated exhaled breath condensate (EBC) samples after different washing times; ○: high Descogen^{TM} (15-min rinsing); ●: low Descogen^{TM} (30-min rinsing). t[1] and t[2] are the first two principal components.

EBC spectral discrimination between HS, laryngectomized and COPD patients

The 3.5-1.7 ppm region of clean (*i.e.* Descogen^{TM}-free) EBC samples was used to investigate the metabolites characterizing EBC. Figure 4.25 depicts representative spectra of HS (Fig. 4.25a), laryngectomized patients (fig. 4.25b) and COPD patients (Fig. 4.25c).

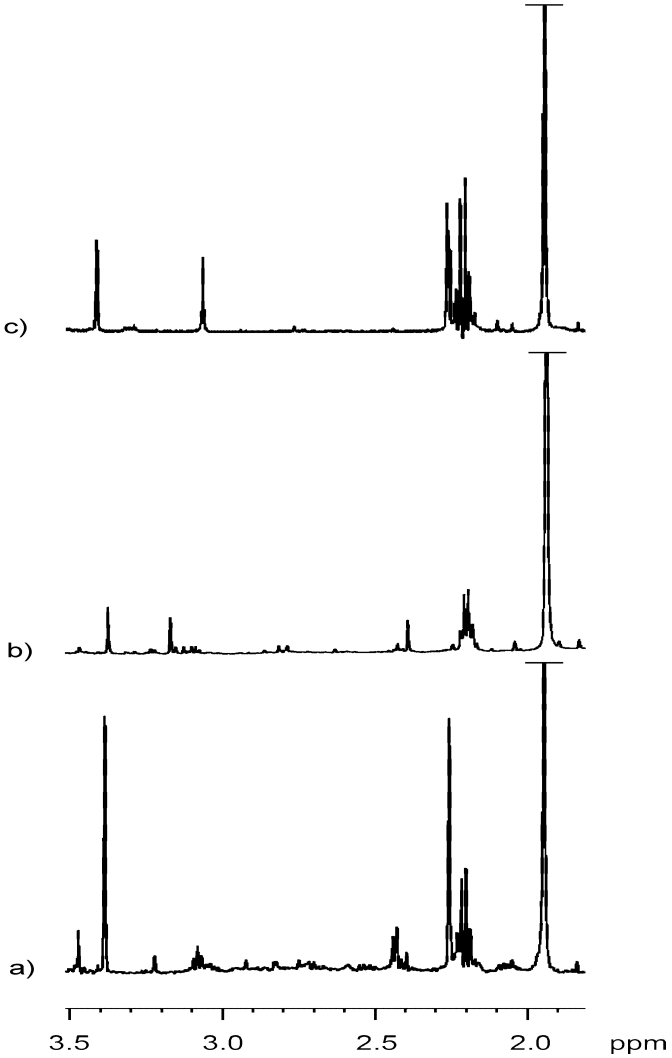


Figure 4.25: Representative ^{1}H-nuclear magnetic resonance spectra of contaminant-free exhaled breath condensate samples from a) healthy subjects, b) laryngectomized patients and c) chronic obstructive pulmonary disease patients. The acetate singlet at 1.93 ppm is cut by a horizontal bar.

Although the region contains few signals, the signals specifically characterize each patient subset, showing both quantitative (signal intensity) and qualitative (signal absence/ presence) differences. Differences in intensity were shown by the signals of: acetate βCH3 (singlet) at 1.93 ppm; propionate αCH2 at 2.19 ppm (quartet); pyruvate βCH3 (singlet) at 2.37 ppm; succinate α, βCH2 (singlet) at 2.41 ppm; glutamine γCH2 (multiplet) at 2.45 ppm; choline and phosphorylcholine N-CH_{3s} (singlets) at 3.16 and 3.23 ppm, respectively; methanol CH3 at 3.37 ppm (singlet); and trimethylamine-N-oxide (TMAO) N-CH_3 (singlet) at 3.44 ppm, as well as by the singlet at 3.03 ppm that most likely originated from N-CH_3 of creatine/creatinine. Pyruvate was present in the COPD spectrum (Fig. 4.25c) and was very intense in the HS spectrum (Fig. 4.25a), but barely visible in the laryngectomized spectrum (Fig. 4.25b). Succinate was small in the HS spectrum (Fig. 4.25a), bigger in the laryngectomized spectrum (Fig. 4.25b) but absent in the COPD spectrum (Fig. 4.25c). Glutamine was only present in the HS spectrum (Fig. 4.25a). The singlet at 3.03 ppm was only present in the COPD spectrum (Fig. 4.25c). Choline and phosphorylcholine were absent in the COPD spectrum (Fig. 4.25c), and TMAO was present in the HS spectrum (Fig. 4.25a), barely seen in the laryngectomized spectrum (Fig. 4.25b) and absent in the COPD spectrum (Fig. 4.25c). All these differences prompted a clear discrimination of HS, laryngectomized and COPD patients in three separate groups (Fig. 4.26).

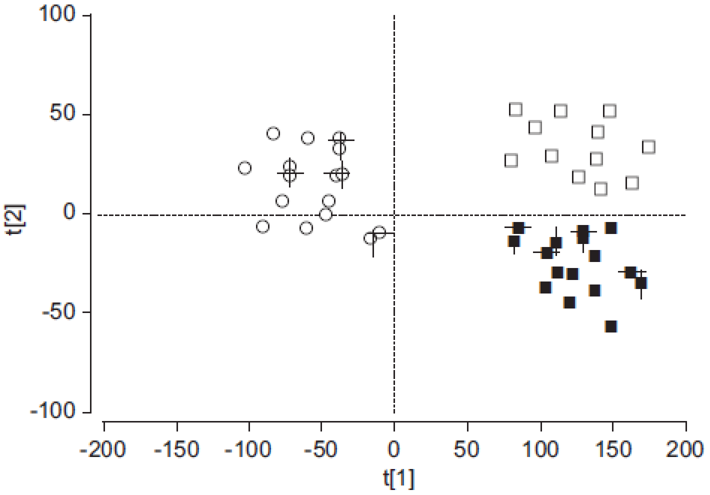


Figure 4.26: Partial least squares discriminant analysis (PLS-DA) scores discrimination for contaminant-free exhaled breath condensate samples. ■: healthy subjects; □: laryngectomized patients; ○: chronic obstructive pulmonary disease patients. Two PLS-DA components afforded a clear classification (~94%), with all samples correctly classified into three regions. Vertical and horizontal bars refer to samples collected in duplicate. t[1] and t[2] are the first two principal components.

4.6 Discussion

The present study demonstrates, for the first time, that NMR based metabonomics can be used to analyze EBC samples from adults, allowing a clear-cut separation between HS and patients with airway disease.

Although less sensitive than ELISA and mass spectrometry, NMR requires minimal sample preparation with a rapid acquisition time (~10-15 min). Furthermore, it is nondestructive and allows complete detection of observable metabolites ("sample metabolic fingerprint") at a reasonable cost.

The present data show that saliva is significantly different from the EBC samples and that the presence of identical metabolites in EBC and saliva does not hamper discrimination. By selecting the 3.5-0.8 ppm region (thereby excluding the carbohydrate signals absent in EBC), saliva spectra clearly differ from EBC (Fig. 4.22), notwithstanding the presence of some common metabolites (leucine, lactate, propionate, acetate, etc.). EBC standardizing guidelines [32] indicate that it is reasonable to assume that there is some degree of oral contamination of EBC, as saliva contains many of the mediators that are also present in the lower airways. Contamination of EBC is often proved by measuring the amylase level, but such a test is not specific and a negative signal does not completely exclude minute contribution from the mouth. To date, there are no data comparing the metabolic saliva composition and a lower airway derivate such as EBC, mainly because condensate samples have been screened for single, specific biomarkers and not as a whole. Indeed, combined saliva and EBC analysis by a metabonomics method has been recently advocated [116]. In light of these assumptions, the current authors also examined EBC from laryngectomized patients, which may represent a true saliva-free material from the lower airways, showing that in those subjects all saliva spectra strictly differed from corresponding EBC samples. Importantly, all EBC and saliva collected twice within the same day (12 h apart) showed good within-day repeatability (Fig. 4.26). Taken together, the data suggest that saliva contamination may play a minor role in the interpretation of EBC by NMR-based metabonomics. The influence of external contaminants was also considered, as the International Consensus on EBC recommends special care in the disinfection of reusable parts of condensers [31]. Upon standard cleaning, all EBC spectra presented signals corresponding to unknown inactive substances of the disinfectant. They persisted even after strong and repeated water soaking, and the presence of variable disinfectant concentration upon different cleaning levels may render classification less effective. Complete removal of the disinfectant signals was observed after washing the reusable parts with 96% ethanol and then rinsing thoroughly with distilled water for 15 min. EBC samples were "spiked" by partially washing the apparatus with water, after treatment with freshly prepared Descogen^{TM}, obtaining different degrees of EBC contamination. Since the citric acid signals were absent after partial washing (Fig. 4.23b), it is important to underline that the potentially toxic saline components of the disinfectant are easily removed from the condenser apparatus by water washing. However, the removal of interfering residual external contaminants is crucial for a correct EBC analysis. There are no data on the influence of residual disinfectant agents of reusable parts of EBC condensers. The influence of residual Descogen^{TM} on reported biomarker levels was not evaluated by an ELISA method, but the present authors suggest that the potential role of external contamination on the variability of some biomarkers [119, 120] should be evaluated. Significantly, by selecting specific regions of EBC spectra for statistical analysis, an efficient discrimination of samples was obtained. Although separation between HS and COPD patients can be achieved by either forced expiratory volume in one second measurements or clinically, the current authors evaluated the capability of NMR-based metabonomics to separate EBC subjects with airway diseases (COPD) from subjects without respiratory diseases. Five NMR signals appear to differentiate "respiratory" (COPD) from "non-respiratory" (HS and laryngectomized) subjects. As a comparison, Carraro *et al.* [40] reported the single acetate signal variation as distinctive in asthmatic children with respect to controls. They hypothesized that acetate increase might be related to increased acetylation of pro-inflammatory proteins in the extracellular space in the airway environment. Furthermore, they found that peaks in 3.2- 3.4 ppm regions of the NMR spectrum of asthmatic children were probably related to oxidised compounds. Heili-Frades *et al.* [121] have reported preliminary data on significant variations between NMR EBC spectra of normal and pathological cases with implications for correlative studies using spectral and clinical classification.

In the present study, by comparing EBC from respiratory (COPD) patients and non-respiratory (HS and laryngectomized) subjects, as well as acetate, four additional signal variations were found, which are likely to have included the methoxy compounds. It can be speculated that such variations could derive from the increased oxidative stress that is a hallmark of COPD, and these variations are usually investigated in EBC by measuring a limited number of markers [119, 120]. Also, the comparison between HS, laryngectomized and COPD EBC samples showed a clear-cut difference (Fig. 4.25) in the COPD patients compared with the other subjects. Figure 4.26 depicts a significant statistical difference along t[1] of COPD patients compared with HS and laryngectomized patients, who are less separated along t[2]. This could be interpreted by the fact that laryngectomized patients were not labeled as COPD before or after surgery; furthermore, mild airflow limitation was detected in only a few subjects (data not shown).

In conclusion, NMR-based metabonomics can safely be applied to exhaled breath condensate in adults, allowing an unambiguous definition irrespective of natural and/or artificial contaminants. In particular, the current authors report that nuclear magnetic resonance spectra of exhaled breath condensate, collected with a device using a salivary trap, do not show the presence of saliva signals. Furthermore, for the disinfectant medium currently used, a careful selection of the nuclear magnetic resonance region allows a clear statistical classification of samples, even for contaminated exhaled breath condensate samples. Finally, the present results suggest that condensate can be efficiently studied as a whole, and that nuclear magnetic resonance may become a leading diagnostic technique in this field.

Chapter 5

CAKE simulations and experimental tests

Contents

This chapter is based on the paper: R. Romano, D. Paris, F. Acernese, F. Barone, A. Motta. *Fractional volume integration in two-dimensional NMR spectra: CAKE, a Monte Carlo approach.* J Magn Res 192 (2008) 294-301.

5.1 Simulation tests

In order to test the CAKE algorithm, we simulated peaks of different shape and overlapping degree. First, we applied CAKE to simulated overlapping peaks of known volume with different overlapping degrees to optimize the number N_{Pbase} to determine the fractional volume V_F with the Hit-or-Miss method. Second, we tested CAKE integration on different elliptic NMR peak sections.

5.1.1 Simulations: bias *vs.* overlapping

We considered two Gaussian peaks centered at (x_i, y_i), of equation

$$G(x,y) = A_i exp[-\frac{(x-x_i)^2 + (y-y_i)^2}{2\sigma_i^2}] \qquad (5.1)$$

volume $V_i = 2\pi\sigma_i^2 A_i$ and with half-height width $\zeta_i = \sqrt{2\sigma_i^2 ln2}, \quad i = 1, 2,$ and addition of Gaussian noise. Denoting by

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{5.2}$$

the distance between the peak centers, it is possible to define the parameter η

$$\eta \equiv \frac{\zeta_1 + \zeta_2}{d} \tag{5.3}$$

as an index of the overlap, such that a large value corresponds to strong overlap. Setting the amplitude $A_1 = 50.0$ and the dispersion $2\sigma_1^2 = 2.0$ to obtain $V_1 = 100\pi$, the A_2 and $2\sigma_2^2$ values were changed so as to keep the volume V_2 constant ($V_2 = 100\pi$), with the overlap index being $0.8 \leq \eta \leq 1.5$. The contour plots of the simulated peaks are reported in Figure 5.1and Figure 5.2 for $\eta = 0.8$ (peak 1), and $\eta = 1.5$ (peak 2).

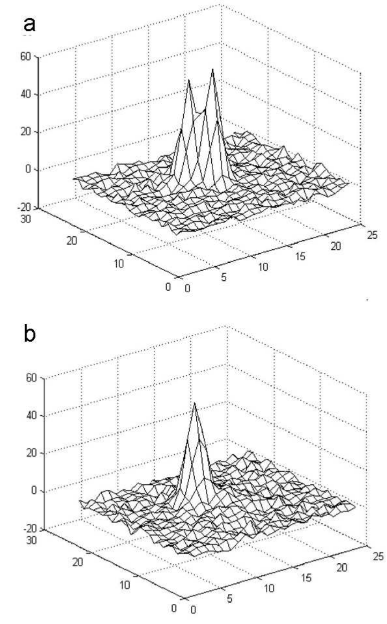


Figure 5.1: 3D Gaussian peaks with different degree of overlap (η): a) $\eta = 0.8$ and b) $\eta = 1.5$.

CAKE integration was compared with the standard one, obtained by summing the amplitudes of all data points within a polygonal bounding the peak. In order to establish the best number of extractions N_{P} in the Hit-or-Miss determination of R, and the best number of extractions $N_{\mathrm{P}_{\mathrm{base}}}$ in the Hit-or-Miss determination of the fractional volume, simulations were conducted in the extreme limit of $\eta = 1.5$.(Figure 5.2, peak 2). Figure 5.3 reports the percentage

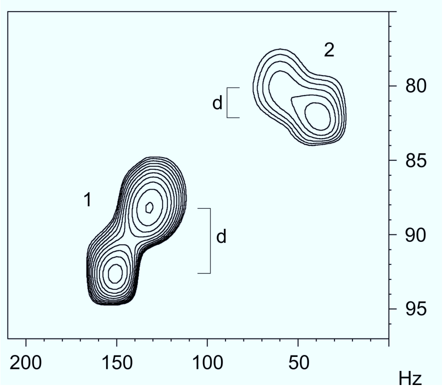


Figure 5.2: Contour plot of two Gaussian peaks with different degree of overlap (η): peak 1, $\eta = 0.8$ and peak 2, $\eta = 1.5$. For the definition of η see text. d is the distance between peak centers.

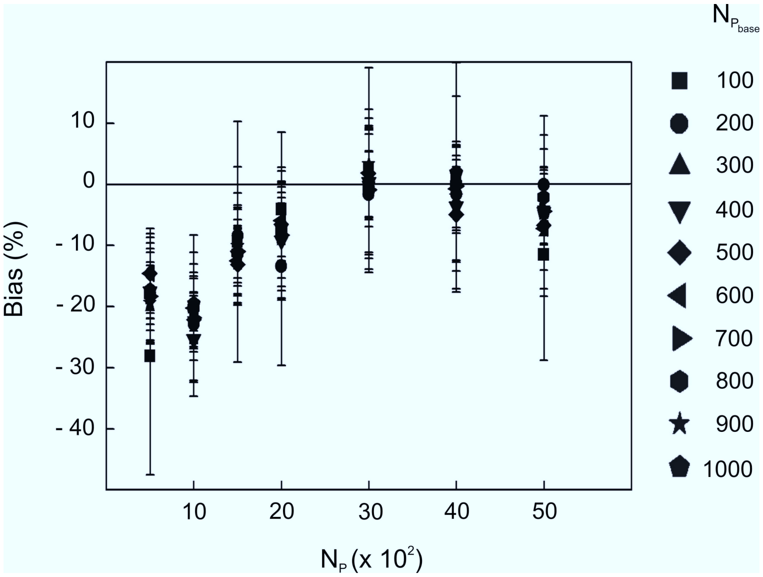


Figure 5.3: Percentage (%) of Bias as a function of the number of extractions (N_P) to estimate the R factor. For each N_P we tested several $N_{\mathrm{P}_{\mathrm{base}}}$ values to estimate the volume fraction, and they are indicated with corresponding symbols on the right.

of Bias *vs.* the number of extractions N_P, for different $N_{\mathrm{P}_{\mathrm{base}}}$ values ranging from 100 to 1000 (right column in Figure 5.3). As it can be seen, results become unbiased for $N_{\mathrm{P}} \geq 1500$, while, except for $N_{\mathrm{P}_{\mathrm{base}}} = 100$ (square symbol), the dependence on $N_{\mathrm{P}_{\mathrm{base}}}$ is negligible. Accordingly, the values $N_{\mathrm{P}} = 2000$, and $N_{\mathrm{P}_{\mathrm{base}}} = 500$ appear to be a good compromise between computing time and accuracy. The results of the simulations are reported as percentage of Bias *vs.* the degree of overlap for a signal-to-noise ratio (SNR) of 34.9±3.0 (Figure 5.4A) and 56.1±4.7 (Figure 5.4B). The standard integration (filled squares) was carried out by bounding the peak with an ellipse, while for the CAKE integration (filled circles) we used $N_{\mathrm{P}} = 2000$, and $N_{\mathrm{P}_{\mathrm{base}}} = 500$. In both cases, each integration was repeated 10 times.

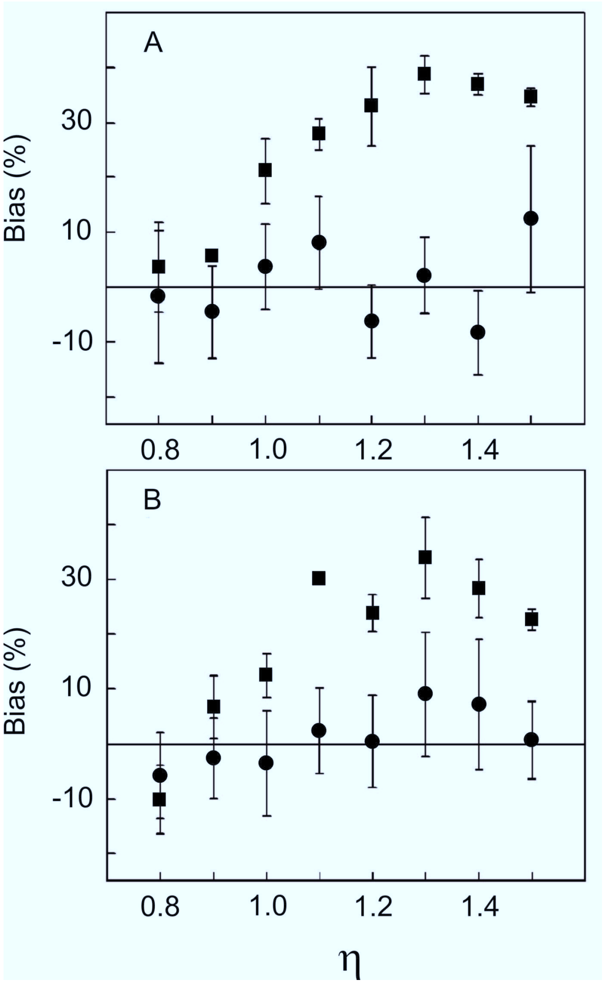


Figure 5.4: Simulation results expressed as percentage of Bias in volume estimation *vs.* the degree of overlap (η). Integration was achieved with the standard (■) and the CAKE (•) methods at different signal-to-noise ratios. (A) $SNR = 34.9 \pm 3.0$; (B) $SNR = 56.1 \pm 4.7$.

In Figure 5.4A ($SNR = 34.9\pm3.0$), the standard method gives unbiased integration values only for low overlap index $\eta \leq 0.9$. (Figure 5.2, peak 1), to become totally biased for $\eta \geq 1.0$. In contrast, CAKE always performs better, especially in the range $1.0 \leq \eta \leq 1.3$, which represents different degree of overlap commonly found in 2D spectra. Overall, the fractional method appears to be unbiased in the whole $0.8 \leq \eta \leq 1.5$ range, that is for strongly overlapping peaks and in the presence of a low signal-to-noise ratio ($SNR = 34.9\pm3.0$). Figure 5.4B reports the same simulations with a $SNR = 56.1\pm4.7$. The standard method performs well for $\eta \leq 0.9$, with a general trend very similar to that observed for lower SNR (Figure 5.4A). In contrast, the fractional method shows a general reduction of the bias percentage, with values generally lower than those obtained in the previous simulation. Taken together our results suggest that, regardless of the SNR, the CAKE method performs always better than the standard one.

5.1.2 Simulations: bias *vs.* eccentriciy

Since experimental 2D-peak shapes are close to elliptic, we tested CAKE on a simulated ellipse of known volume. In particular, we considered peaks of equation

$$S_i(\omega_1, \omega_2) = A_i(\frac{2\pi}{\sigma_{1i}\sigma_{2i}}) \exp{(-\frac{\Delta\omega_1^2}{2\sigma_{1i}^2})} \exp{(-\frac{\Delta\omega_2^2}{2\sigma_{2i}^2})} \tag{5.4}$$

volume $V_i = A_i$ and contour of eccentricity

$$e_i = \sqrt{1 - \frac{min(\sigma_{1i}, \sigma_{2i})}{max(\sigma_{1i}, \sigma_{2i})}} \tag{5.5}$$

with addition of Gaussian noise. Integration was carried out in two ways. The fractional area was firstly selected randomly (i.e. avoiding any symmetry), and, secondly, symmetrically with respect to any of the semiaxes of the elliptic peak. The random choise (Figure 5.5A) produced a scattered bias distribution between 0 and 20% for $0.8 \leq e \leq 0.74$, with a maximum of 25% for $e = 0.78$. For $0.8 \leq e \leq 0.9$, which corresponds to a ratio between semiaxes in the range of $0.45 \leq b/a \leq 0.60$, the average bias is 5%. This result appears to be relevant as the b/a value corresponds to the experimental elliptic shapes usually found in 2D spectra.

The symmetry selection of the fractional area (Figure 5.5B) shows a bias ≤10% for all eccentricity values, with the maximum at $e = 0.78$ reduced to 12%. For $0.8 \leq e \leq 0.9$ the average bias is very similar to that found for the random selection (Figure 5.5A).

In conclusion, it is suggested that, for elliptical peaks, slicing should be done symmetrically with respect to one of the semiaxes, even though for $0.8 \leq e \leq 0.9$, that is for most of the experimental 2D peaks, the bias is essentially indipendent from the selection.

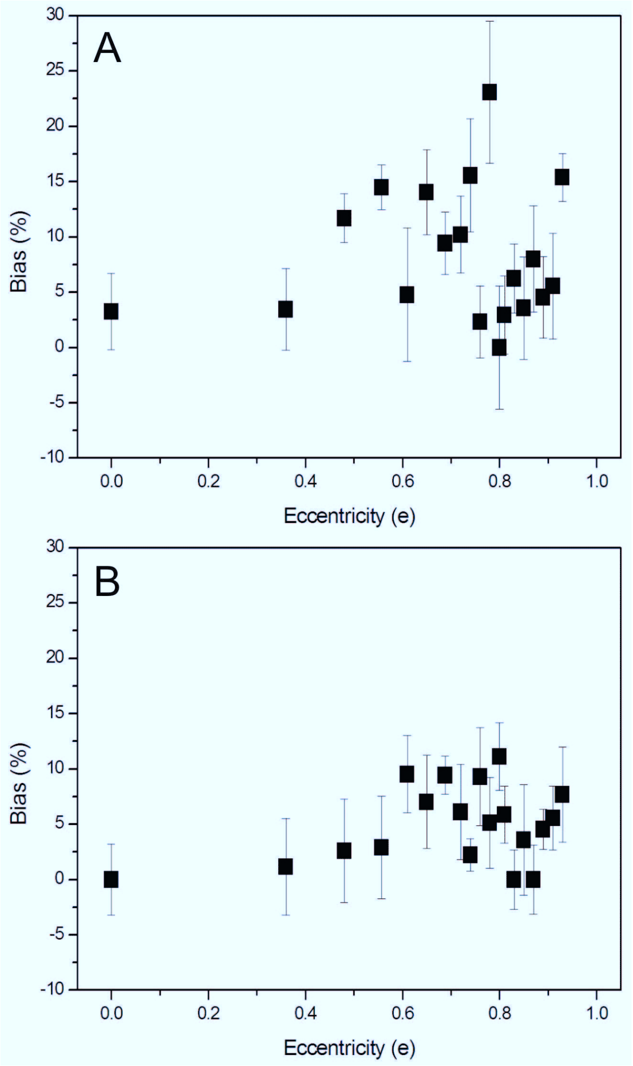


Figure 5.5: CAKE integration of simulated elliptic peaks expressed as percentage of Bias in volume estimation *vs.* Contour eccentricity (e). In (A) the fractional area was chosen in a non symmetric way with respect to the semimajor and the semiminor axes of the elliptic peak. In (B) the fractional area was chosen in a symmetric way with respect to the semimajor and semiminor axes of the elliptic peak. In both cases the $SNR = 69.5 \pm 3.2$.

5.2 Experimental test

To test the efficacy of the new integration method, after simulations, CAKE was applied to 2D-NMR spectra of a sample containing two tripeptides in known concentrations; we compared peak volume estimations obtained by CAKE with those obtained by standard integrations.

5.2.1 NMR data collection

The sample, a mixture of the tripeptides Ala-Phe-Ala (AFA) and pyroGlu-His-Pro (thyrotropin-releasing hormone, TRH), was prepared by dissolving appropriate amounts in 0.5 ml of $^1H_2$O/$^2H_2$O (90/10 v/v) to yield for each peptide a concentration of 0.10 mM. $^1H-$NMR spectra, recorded at 295 K and pH 7.4, were acquired on a Bruker DRX-600 spectrometer operating at 600 MHz, equipped with a TCI cryoprobe^{TM} fitted with a gradient along the Z-axis. Spectra were referenced to sodium 3-(trimethylsilyl)-[2,2,3,3-$^2H_4$]propionate. Homonuclear 2D clean TOCSY spectra [122] were recorded by standard techniques and incorporating the excitation sculpting sequence [95] for water suppression. We used a pulsed-field gradient double echo with a soft square pulse of 4 ms at the water resonance frequency, with the gradient pulses of 1 ms each. 512 equally spaced evolution time-period t_1 values were acquired, averaging 4 transients of 2048 points, with 6024 Hz of spectral width. Time-domain data matrices were all zero-filled to 4096 in both dimensions, yielding a digital resolution of 2.94 Hz/pt. Prior to Fourier transformation, time-domain filtering was applied with a Lorentz-Gauss window to both t_1 and t_2 dimensions. The TOCSY experiment was recorded with a spin-lock period of 64 ms, achieved with the MLEV-17 pulse sequence [98].

5.2.2 Software

NMR data processing and baseline correction were obtained with the program XWINNMR (Bruker, Biospin GmbH, Ettlingen, 2003). Standard peak integration was carried out with the programs XWINNMR and MestRe-C [123], in which integrated volumes are computed as the sum of all digital intensities within a rectangular box and a tunable ellipse bounding a peak, respectively. CAKE software was written in MATLAB language and was implemented in the graphical environment of MATLAB 7.1.

5.2.3 Experimental Results

The power of the CAKE approach was tested on a TOCSY spectrum of a mixture of two tripeptides, AFA and TRH (Figure 5.6).

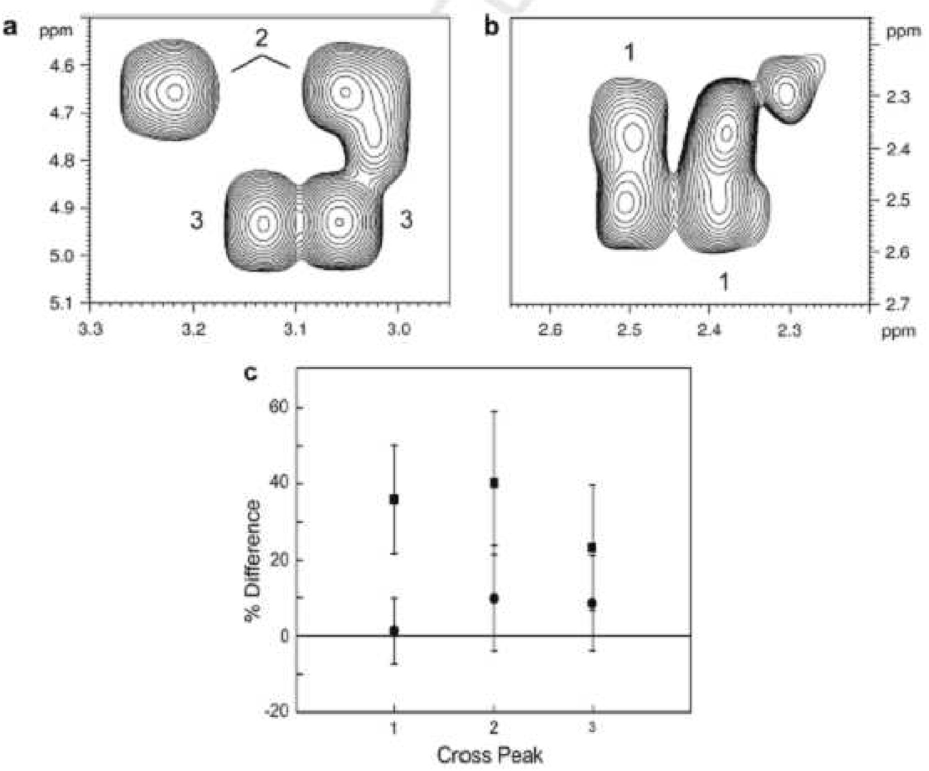


Figure 5.6: (a)TOCSY spectrum of the AFA and THR tripeptides aliphatic region, acquired at $300K$ with 64 $msec$ mixing time. Expansions (b) and (c) report peaks originating from γCH_2 protons of the TRH pyroGlu [labeled 1 in (b)], and α and β protons of AFA Phe^2 [labeled 2 in (c)], and TRH His^2 [labeled 3 in (c)].

In order to have an internal reference we selected pairs of peaks, each of them stemming from a single spin system, such that they have similar intensity within each pair but one peak overlaps with others. In particular we chose pairs that exemplify the correlations between the γCH_2 (labeled 1 in Figure 5.6b), and between α and β protons of AFA Phe^2 (labeled 2 in Figure 5.6a), and TRH His^2 labeled 3 in Figure 5.6a). The magnitude of a given TOCSY peak [governed by mixing coefficients $a_{lk}(\tau_m)$ for transfer of magnetization through the spin system from spin I_l to spin I_k] depends on the topology of the spin system, the coupling constants between pairs of spins, the efficiency of the isotropic mixing sequence employed, and the relaxation rate during the mixing pulse. Although the robustness of the integration method does not depend upon the experiment type or the intensity of the chosen peak, we looked for pairs in which the peaks are expected to have similar intensity but one of them overlaps with others. Accordingly, we selected the AMX spin system of the two aromatic residues (Figure 5.6a) in AFA and TRH. From relaxation measurements (not shown) at two different spectrometer frequencies, we estimated for both peptides similar correlation times and relaxation rates; furthermore, the measured $^3J_{\alpha\beta}$ and $^3J_{\alpha\beta'}$ values in each spin system were identical, therefore excluding differences in the peak intensity due to different coupling constants; finally, the single $^2J_{\gamma\gamma'}$ value for the γCH_2 protons of the $TRHpyroGlu$ warrants a similar intensity for the two peaks within each pair.

The selected peaks were integrated with standard and with CAKE methods and the results are reported in Figure 5.6c as the Difference percentage of volume for each cross-peak pair. For the CAKE integration we selected the most internal level belonging to a single peak, which had elliptical symmetry with eccentricity $e > 0.75$. The values obtained with CAKE for the three peak pairs are all within 10%, giving an unbiased estimation of the difference percentage of the volumes in each pair. In contrast, the standard method estimates for each peak pair values $> 35\%$ for pairs 1 and 2, and $\approx 25\%$ for pair 3. Surprisingly, the CAKE approach gives for the pair 1, which lies on the TOCSY diagonal, about zero volume difference, supporting robustness for the method, also in the presence of elliptical symmetry.

5.2.4 Bias *vs.* digital resolution

The dependence of CAKE on digital resolution was investigated by integrating the peak pair 2 (Fig. 5.6c) at different digital resolution (0.5, 1.1, 2.2, 4.3 and 8.6 Hz/pt), and integration was carried out for each value with standard and CAKE methods (Fig. 5.7). The volume of pair 2 overlapping peak (located at ω_1 =4.75 ppm and ω_2 =3:05 ppm, Fig. 5.6c) was compared to the volume of the corresponding single peak at ω_1 =4.75 ppm and ω_2 =3:05 ppm at its maximum digital resolution, taken as reference. The values obtained with CAKE are all within 2%, giving an unbiased estimation of the % Difference up to 8.6 Hz/pt. On the contrary, the standard method estimates values $>$10% already at 2.2 Hz/pt to become $\approx$ 25% at 8.6 Hz/pt. This finding can be explained by considering that a low resolution drastically reduces the number of points within an area identified by the i-th level, which, in turn, is itself poorly defined. Therefore, the sum of points done by standard methods is obviously biased. On the contrary, the Hit-or-Miss technique used in CAKE does not sum the existing points included in a level bound area, but generates random points and counts the number of "hits" (or points) that are included in the unknown area. Since a cubic interpolation (see Chapter 2) is used as a decisional mean to establish if the extracted point can be considered a "hit", a low digital resolution could, in principle, affect the peak profile. However, with CAKE we were able to correctly integrate peaks with digital resolution up to *ca.* 30 Hz/pt.

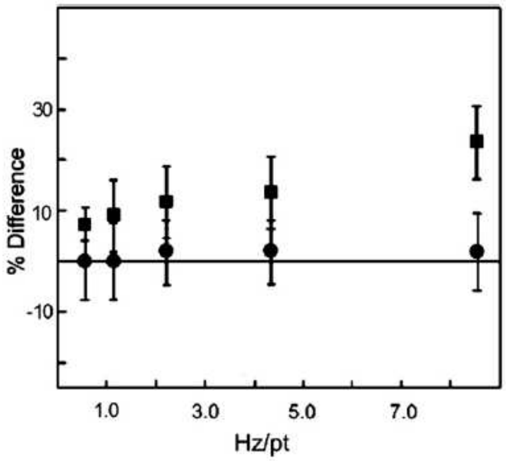


Figure 5.7: Difference percentage (%) of volume determination at different resolution for cross-peak 2, as labeled in Fig.5.6. The digital resolution was ca. 0.5, 1.1, 2.2, 4.3 and 8.6 Hz/pt. Filled squares and circles refer to the standard and CAKE integration methods, respectively.

5.3 Discussion

Quantification of NMR spectra is fundamental both in metabolomics/ metabonomics and in the structure determination of biomolecules. However, quantification of peaks is often hampered by the degeneracy of the NMR resonance frequency, a factor that aggravates with the increasing size of macromolecules and the number of metabolites. Here we have presented the CAKE approach that uses the symmetry of a single in-phase peak (a peak with a unique center corresponding to its maximum) to calculate its volume. It is obtained by multiplying the fractional volume by the R factor, a proportionality ratio between the total and the fractional volume, both evaluated with Monte Carlo techniques. Therefore, the peak volume can be estimated by integrating a known fraction of the peak, and the fractional volume can be chosen so as to minimize the effect of overlap in complex NMR spectra. Strictly speaking CAKE applies to Gaussian peaks showing cylindrical or elliptic symmetry. However, an NMR spectrum is closely approximated by Lorentzian functions, which in its 2D shape show the so-called "star effect". It can be easily removed by 2D Lorentz-to-Gauss transformation, which is routinely used for in-phase experiments, like TOCSY and NOESY. Therefore, the major assumption in this study is that the Lorentzian signal is converted into a Gaussian line by a Lorentz-to-Gauss transformation, which is routinely applied in 2D data manipulation. Integration of simulated and experimental 2D in-phase peaks with different degree of overlap shows that CAKE works well even for strongly overlapping peaks. The main advantage of CAKE is its simplicity as difficulties in its use are comparable to those presented by methods that sum all data points in a defined area. In fact, the user only has to select a peak slice not overlapping with other peaks therefore avoiding the guess of the total contour shape of the peak. Furthermore, CAKE does not require any time-consuming fitting of the peaks to functional forms, and therefore it can be easily incorporated as a subroutine in any NMR processing software. Tests on tripeptides have shown that CAKE is a powerful method for volume integration. The substantial independence of CAKE on digital resolution and SNR warrants that it can be safely used for peak integration in three-dimensional spectra. Because of its inherent simplicity the software can be extended to automated integration of three- and possibly higher-dimensionality NMR spectra.

Chapter 6
^{1}H-^{15}N SO-FAST-HMQC measurements

Contents

This chapter is based on the paper: A. Motta, D. Paris, G. Andreotti, D. Melck. *Monitoring real-time metabolism of living cells by fast two-dimensional NMR spectroscopy.* Submitted to Analitical Chemistry.

6.1 Materials and methods

6.1.1 Cell culturing

Axenic cultures of *T. rotula* cells were prepared as described in Miralto and coworkers protocols [124]. Briefly, diatoms were grown in Guillard's (F/2) Marine Enrichment Basal Salt Mixture Powder medium, containing standard and different salinities (20, 35 and 45 ‰) and unlabeled or ^{15}N-labeled $NaNO_3$, on a 12 h light/12 h dark cycle, and a light intensity of 20.9 J mol^{-1} $\mu m^{-2}s^{-1}$. Cells were kept in a 10 L carboy for 1 week and then harvested in the early stationary phase by centrifugation at 1200g in a swing-out rotor. Prior to extraction, diatom cultures were allowed to settle overnight and the supernatant was gently removed by suction with a water pump.

6.1.2 Extracts manipulation

Combined extraction of polar and lipophilic metabolites from unlabeled and ^{15}N-labeled diatoms cells was carried out by using the methanol/chloroform procedure [92] Pelleted cells were resuspended in methanol (4 ml/g pellet)-water (0.85 ml/g pellet), and sonicated for 2 min. Then 4 ml/g pellet of chloroform were added and the homogenate was gently stirred and mixed on ice for 10 min using an orbital shaker (the solution must be mono-phasic). Other 4 ml/g pellet of chloroform and 4 ml/g pellet of water were then added, and the final mixture was shaken well and centrifuged at 12000g for 15 min at 4 °C. This procedure separates a water/methanol phase at the top (aqueous phase, with the polar metabolites), a phase of denatured proteins and cellular debris in the middle, and a chloroform phase at the bottom (lipid phase, with lipophilic compounds). The upper layer of each sample was transferred into glass vials, and, after solvent removal under a stream of dry nitrogen, was stored at -80 °C until required. For 1D and 2D NMR experiments the polar extracts were resuspended in 700 μl H_2O-D_2O (90%-10%), and then transferred into an NMR tube.

6.1.3 Gel electrophoresis for protein detection

To eventually exclude the detections of small proteins from the SOFAST-HMQC *in vivo* spectra acquisition of *T. rotula*, we performed SDS-PAGE electrophoresis. SDS-PAGE on slab gel containing 12 and 15% acrylamide, in order to reach the lower limit of 3 kDa, was performed by using the standard procedure (12). Proteins were located on the gels using Comassie Brillant Blue staining. For 12% acrylamide we used Phosphorylase b (97.4 kDa), bovine serum albumine (66.2 kDa), ovalbumin (45.0 kDa), carbonic anhydrase (31.0 kDa), trypsin inhibitor (21.5, kDa), and lysozyme (14.4 kDa), all from BIO-RAD. For 15% acrylamide we used chymotrypsinogen A (24 kDa), cytochrome c (13 kDa), bovine pancreatic tripsin inhibitor (BPTI, 6.6 kDa), insulin B-chain (3.5 kDa), all from Sigma. Size-exclusion chromatography was carried out at room temperature, using a 1.5 × 50 cm Sephadex G-50 Fine column and a flow rate of 0.2 ml/min. Separate chromatography experiments of standard amino acids were performed in 50 mM sodium phosphate, at pH 6.7, using a 55 μM peptide concentration. Salmon calcitonin (3.4 kDa), bacitracin (1.4 kDa), standard amino acids all from Sigma, and sodium 3-(trimethylsilyl)-(2,2,3,3-$^2H_4$)propionate (TSP, 172 Da), from Aldrich, were used as molecular mass standards.

6.1.4 NMR experiments

All NMR experiments were carried out on a Bruker DRX-600 spectrometer, equipped with a TCI CryoProbe^{TM} fitted with a gradient along the Z-axis.

T. rotula ^{1}H and TOCSY spectra

^{1}H-NMR spectra were recorded at 600 MHz and were referenced to internal TSP. Clean total correlation spectroscopy (TOCSY)[97] spectra of cells and extracts were recorded by using the time-proportional phase incrementation of the first pulse, and incorporating the excitation sculpting sequence [95] for water suppression. We used a double-pulsed field gradient echo, with a soft square pulse of 4 ms at the water resonance frequency, with the gradient pulses of 1 ms each in duration. In general, 256 equally spaced evolution-time period t_1 values were acquired, averaging 2 (for diatoms) and 8 (for extracts) transients of 2048 points, with 6024 Hz of spectral width. Time-domain data matrices were all zero-filled to 4K in both dimensions, applying, prior to Fourier transformation, a Lorentz-Gauss window with different parameters for both t_1 and t_2 dimensions in all the experiments.

T. rotula ^{1}H-^{15}N SO-FAST-HMQC parameters set-up

The ^{1}H-^{15}N SOFAST-HMQC pulse sequence follows the scheme proposed by Shanda and co-workers [2] (Figure 6.1). First, ^{1}H pulses are applied band-selectively [77]; second, the first ^{1}H pulse has an adjustable flip angle α that allows further optimization of the sensitivity of the experiment for a chosen (short) scan time [78]. In practice, the flip angle is chosen to ensure that part of the proton magnetization is restored along the z-axis by the following 180° pulse; third, the small number of radio-frequency pulses reduces signal loss due to pulse imperfections and B_1 field inhomogeneities, and limits the effects of sample and probe heating. We used polychromatic PC9 pulse shape for adjustable flip-angle band-selective excitation [125] which yields quite uniform excitation over the desired bandwidth for flip angles in the range $0° < \alpha < 130°$. As a refocusing pulse on the ^{1}H channel we tested the r-SNOB [82] and RE-BURP [83] profiles. Because of a signal increase of *ca.* 35%, we used RE-BURP instead of r-SNOB, confirming the finding of Schanda et al. for proteins [3]. The acquisition parameters were as follows: α=120°, $\Delta(1/2JHX)$ = 6.7-5.4 ms, δ= 1.8 ms, t_1^{max}=20 ms, t_2^{max}=40 ms, and t_{rel}=1 ms. Forty complex data points were acquired in the t_1 dimension, adding 4 dummy scans (n = 80 + 4). The band-selective ^{1}H excitation (PC9, 3.0 ms) and refocusing (RE-BURP, 2.03 ms) pulses were centered at 8.0 ppm covering 4.0 ppm.

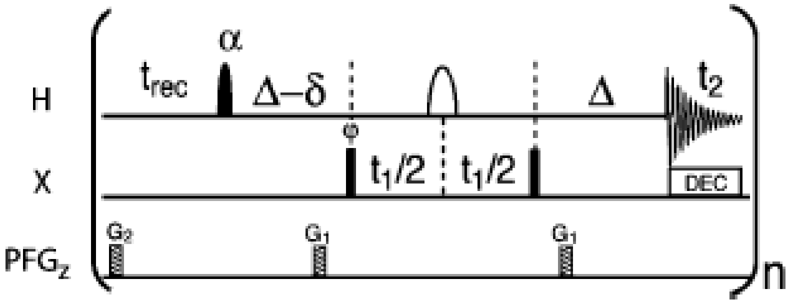


Figure 6.1: SOFAST-HMQC experiment to record ^{1}H-X (X=^{15}N or ^{13}C) correlation spectra of proteins. Filled and open pulse symbols indicate 90° and 180° rf pulses, except for the first ^{1}H excitation pulse applied with flip angle α. As described in the next section, the variable flip-angle pulse has a polychromatic PC9 shape, and band-selective ^{1}H refocusing is realized using an r-SNOB profile. The transfer delay Δ is set to $1/(2J_{HX})$, the delay δ accounts for spin evolution during the PC9 pulse, and t_{rec} is the recycle delay between scans.

^{15}N was decoupled with GARP-4 [126], with a 90° pulse length of 600 μs. ^{15}N chemical shifts are relative to external $^{15}NH_4NO_3$ (5 M in 2 M HNO_3).

6.2 Results

In the cell, metabolites experience a viscosity of *ca.* 2-3 times that of water [127, 128] and interact with other components. As such, restriction of the rotational freedom may be predicted [127]. However, their low molecular weight is likely to counterbalance the viscosity effect, and an increase of the average effective T_1 of in-cell metabolites can be expected. Therefore, a balance of intrinsic and extrinsic properties will affect metabolite relaxation. We firstly checked if high viscosity is a prerequisite for application of SOFAST-HMQC to low-molecular weight metabolites by using a sample of ^{15}N-labeled Leu (5 mM, pH 1.4, 300 K) in the presence of SDS, with a viscosity of 9 relative to water (0.894 cP). The results of the application of the SOFAST pulse sequence to such a sample are reported in Figure 6.2A, in which a ^{1}H-^{15}N correlation peak, centered at 8.01 and 172 ppm, is observed.

The influence of the viscosity on the volume of the cross-peak in Figure 6.2A was investigated by lowering the SDS concentration, and therefore the relative viscosity from 9 to 1 (no SDS). In the 9-3 range we observed that the cross-peak volume remained constant, to significantly decrease upon a reduction of the relative viscosity from 3 to 1 (Figure 6.2B). We estimated that in the absence of SDS (relative viscosity of 1) the cross-peak volume halves.

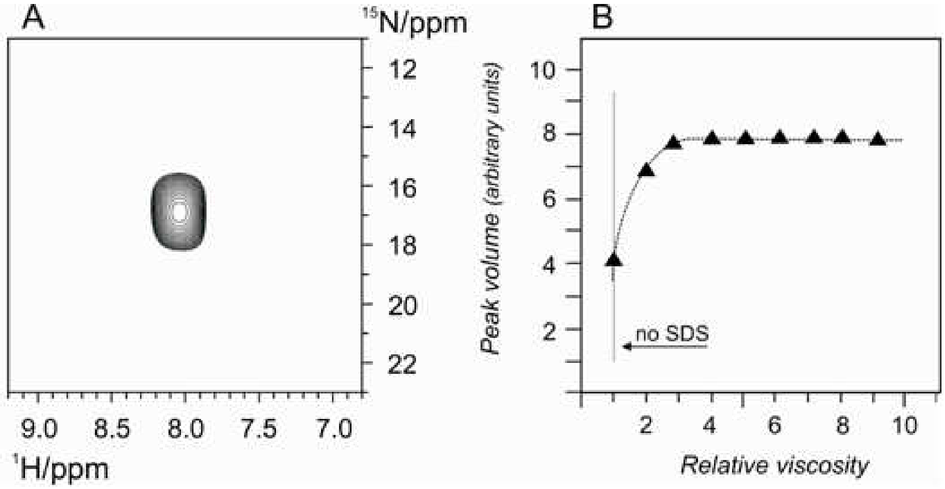


Figure 6.2: (A) ^{1}H-^{15}N SOFAST-HMQC spectrum of ^{15}N-labeled Leu (5 mM, pH 1.4, 300 K) in the presence of SDS, with an acquisition time of 14 s. The $\Delta(1/2JHX)$ value was set to 6.7 ms since J_{HX} = 74.6 Hz; for the remaining acquisition parameters see the Materials and Methods Section. (B) Dependence of the cross-peak volume on the viscosity of the medium, relative to water.

Therefore, for a molecule as small as Leu (MW 132.17 Da), a viscosity of *ca.* 3 times that of water, corresponding to the viscosity inside a living cell [127], maximizes the intensity of the ^{1}H-^{15}N SOFAST-HMQC peak. However, the efficient ^{1}H-^{15}N dipolar interaction is also important, since a well-defined cross peak, although with an intensity 1/2 of the maximum, is observed in the experiment without SDS.

6.2.1 *T. rotula* ^{1}H and TOCSY spectra

Due to intracellular viscosity, a molecule in a cellular environment displays broad NMR line widths as a consequence of the reduced tumbling rate, and overlapped, poor quality spectra are the likely result. In our case, a further complication comes from the presence of high salt concentration in the sea water culture medium, used to suspend the cells in the NMR tube. The final result is that the 1D spectrum obtained for a ^{15}N-labeled *T. rotula* sample containing ca. 50-million cells will show an unresolved "bumpy" distribution of the resonances, as shown in Figure 6.3.

In order to better resolve signals from *T. rotula*, we acquired ^{1}H (Figure 6.4) and TOCSY spectra (Figure 6.5) of *T. rotula* polar extracts (see Materials and Methods Section).

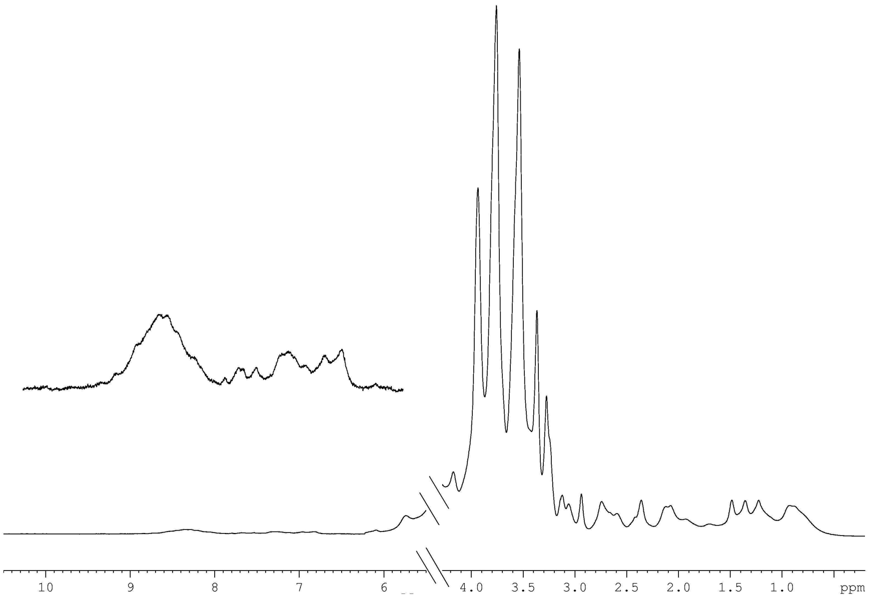


Figure 6.3: ^{1}H spectrum of *in vivo* ^{15}N-labeled *T.rotula* (50×10^{6} cells).

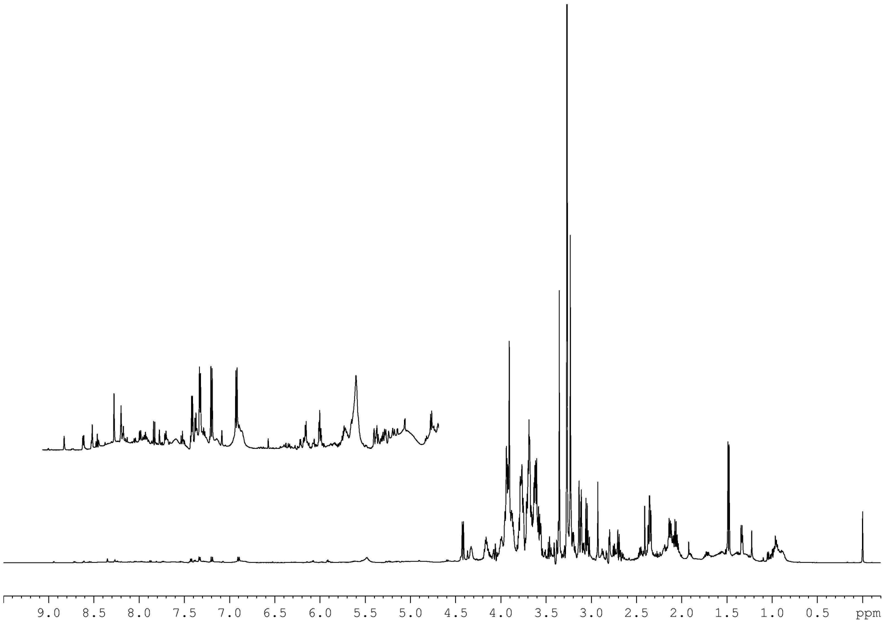


Figure 6.4: ^{1}H spectrum of ^{15}N-labeled *T.rotula* polar extracts (400×10^{6} cells).

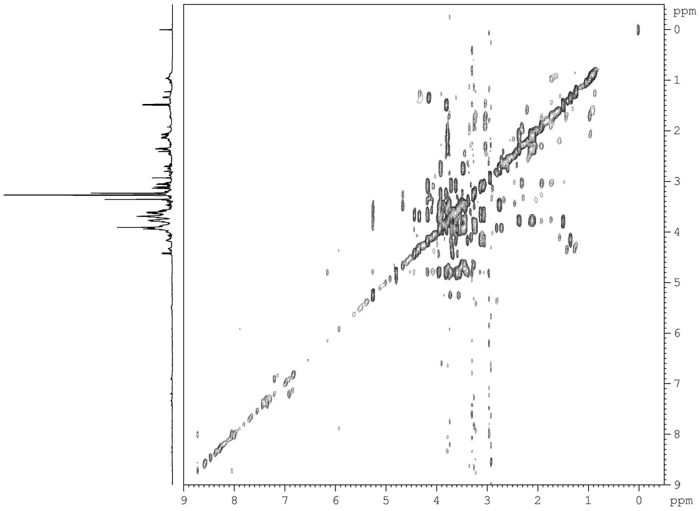


Figure 6.5: TOCSY spectrum of ^{15}N-labeled *T.rotula* polar extracts (400×10^6 cells).

6.2.2 *T. rotula* ^{1}H-^{15}N SOFAST-HMQC spectra

The ^{1}H-^{15}N SOFAST-HMQC correlation spectrum of a 50-million *T. rotula* cells is reported in Figure 6.6: it was directly acquired in the culture medium in an overall experimental time of 12 s.

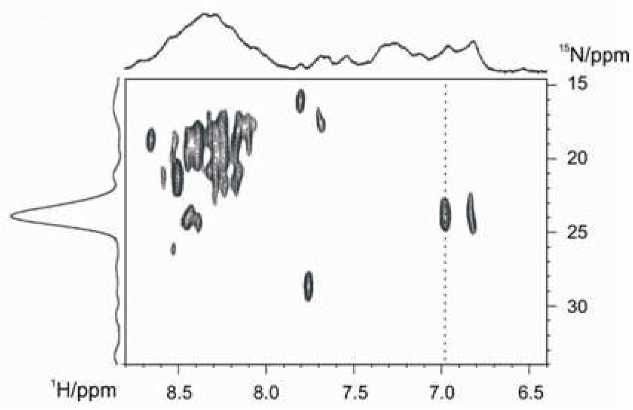


Figure 6.6: ^{1}H-^{15}N correlation spectrum (central part) of a sample of 50-million ^{15}N-labeled diatom cells (in sea water culture medium, 300 K) recorded in 12 s. 1D traces correspond to the proton spectrum (top), and (left) to a column extracted along the ^{15}N dimension at the ^{1}H frequency indicated by the dashed vertical line in the 2D spectrum.

In such a short acquisition time, the NMR experiment certainly does not kill the cells, and in fact the number of colony-forming units/OD is the same before and after the 12-s SOFAST-HMQC experiment (data not shown). Furthermore, compared with 1D, the 2D experiment presents a higher S/N, as it can be appreciated from the trace on the left side of Figure 6.6, extracted along the ^{15}N dimension (vertical broken line in Figure 6.6).

The robustness of in-cell SOFAST NMR spectroscopy was investigated by controlling several aspects [86]. Firstly, because of the high S/N ratio, we reduced the number of cells from 50 millions down to 10 millions, which, as shown in all the experiments below, appear to be sufficient for fast acquisition and high S/N spectra. Figure 6.7A reports the ^{1}H-^{15}N SOFAST-HMQC spectrum of a 10-million cells sample of ^{15}N-labeled *T. rotula*, taken directly in the culture medium. It reproduces the spectral pattern of the more concentrated sample of Figure 6.6, and shows a high S/N ratio with well resolved resonances. Secondly, when dealing with living cells it is important to consider that molecules outside the cell tumble faster and, therefore, exhibit sharper lines than internal metabolites in a more viscous environment. Consequently, a small fraction of extracellular molecules could contribute disproportionately to, or even dominate, the spectrum. This was investigated after removal of the cells from the sample by centrifugation and filtration, and analyzing the supernatant. It contained no detectable extracellular metabolites as its corresponding SOFAST-HSQC spectrum, acquired with the same parameters as the *in-vivo* spectrum 6.7A, showed no signals (Figure 6.7B), therefore ruling out any interference from the extracellular metabolites in Figure 6.7A. This was confirmed by the following step. The pellet separated from the supernatant was resuspended in fresh standard culture medium giving a spectrum (Figure 6.7C) identical to that observed when in vivo (spectrum 6.7A). It is concluded that the cross-peaks we observed in the SOFAST-HSQC experiments of Figures 6.6 and 6.7A stem from molecules within the cell, and that the amount of the released molecules, if present, are beyond detection.

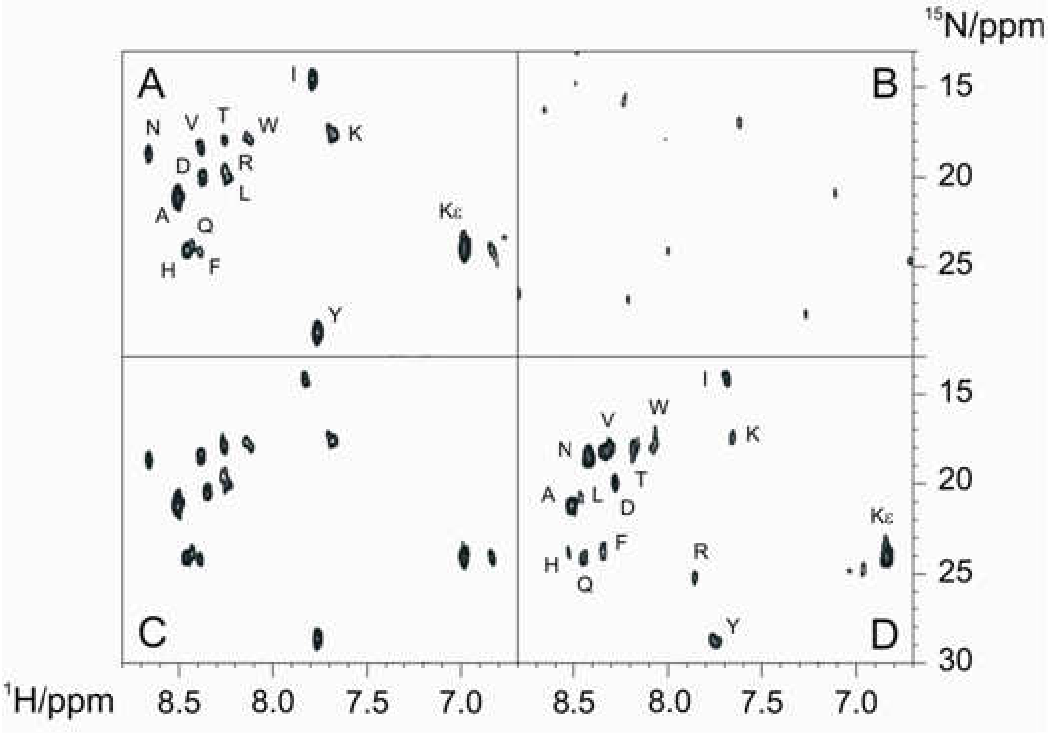


Figure 6.7: ^{1}H-^{15}N SOFAST-HMQC spectrum of ^{15}N-labeled *T. rotula* in varying conditions: (A) in vivo spectrum of 10-million cells directly in the culture medium acquired in 12 s; (B) supernatant of the sample used in (A) after removal of all cells by centrifugation and filtration (vertical scale × 8); (C) pellet after resuspension in fresh culture medium; (D) polar extract obtained with the methanol/chloroform protocol to remove proteins (see text). Peaks are labeled with the single-letter code for amino acids; the asterisk marks a yet unidentified peak.

6.2.3 Gel electrophoresis results

When investigating intracellular ^{15}N-labeled metabolites in vivo by NMR, care must be taken to avoid detection of resonances originating from low-molecular weight proteins within the cell, which might become labeled because of the unspecific labeling process. This was examined by analyzing the polar extracts of the diatom cells by using the methanol/chloroform protocol. The used procedure separates the polar metabolites in the water/methanol phase at the top, a phase of denatured proteins and cellular debris in the middle, and a chloroform phase at the bottom, with lipophilic compounds [92]. As a proof to rule out the presence of signals originating from polypeptides/proteins in the above SOFAST-HSQC spectra, we carried out SDS-PAGE gels of the polar extracts obtained from 10- and 50-million cells. Figure 6.8 reports a 12%-

acrylamide gel (6.8A), and a 15% acrylamide gel (6.8B). In both, the absence of bands in lanes 1 and 2 (reporting 10-million cell extract ran in duplicate) and lanes 3 and 4 (50-million cell extract ran in duplicate) confirmed the total absence of polypeptides/proteins down to a molecular weight of 3 kDa.

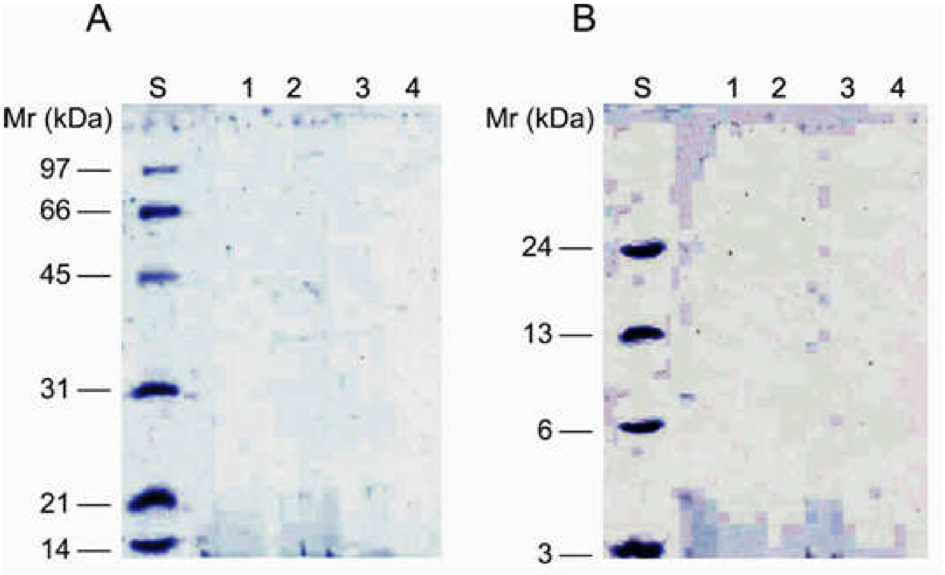


Figure 6.8: SDS polyacrylamide gel electrophoresis of ^{15}N-labeled *T. rotula* polar extracts: (A) 12% acrylamide, and (B) 15% acrylamide. In both, Lane S reports prestained protein standards with molecular weight indicated on the left side; lanes 1 and 2, 10-million cells ran in duplicate; lanes 3 and 4, 50-million cells ran in duplicate. Comassie Brillant Blue staining was used to visualize proteins.

For lower molecular weight we resorted to size-exclusion chromatography under the experimental conditions used for NMR analysis. At pH 6.7, all the molecules present in the polar extract eluted with an apparent molecular mass comparable to that of TSP (172 Da). The experiments described above confirm that the cross-peaks we observed are associated with metabolites within the cells, and that the presence of polypeptides/proteins in the spectra can be safely excluded. The SOFAST-HMQC spectrum of the polar extract (Figure 6.7D) well compares with the in vivo (6.7A) and the resuspended pellet (6.7C) data, showing only small differences in chemical shift, possibly reflecting differences in salt composition of the in-vitro NMR buffer and the cytoplasm. Identification of the cross-peaks was achieved upon a careful titration of the solution with standard amino acids, and the signals are labeled with the one-letter code in spectra 6.7A and 6.7D. It is important to notice that the spectral position of free amino acids corresponds to that observed within the cell, and a similar behavior is observed for proteins inside and outside the cell [86]. However, as for proteins, the great advantage of the observation of in-cell metabolites by fast NMR spectroscopy does not lie in the structural investigation, but on the possibility to examine the behavior of metabolites directly in the cellular compartments, and follow their fate upon a change of the physiological state of the cell as well as in the possible interaction with unlabeled/labeled proteins.

6.3 Discussion

Our simple application had shown that 2D ^{1}H-^{15}N correlation spectra of ^{15}N-labeled metabolites can be recorded in living cells in only 10-15 s of data acquisition using the SOFAST-HMQC sequence that provides high sensitivity. To the best of our knowledge, this is the first time that high-quality 2D correlation spectra of metabolites have been directly recorded in living cells on a time scale of seconds of experimental time and high S/N. Obviously, these are preliminary results and more experimental investigations are needed to explore the potentiality of SOFAST experiments for metabolic detection purpose since, in the future, it is desirable to extend the investigation to eukaryotic cell systems. Potential applications include in-cell investigation under physiological or stressing conditions, high-throughput characterization of cell lines by NMR, testing potential drugs by fast measures of in-cell metabolic changes, as well as investigation of the primary nitrogen metabolism in plant cells.

Conclusions

The results here presented confirm that high resolution NMR spectroscopy is particularly suited for biomarkers discovery. We applied recent NMR avances and developed new tools in order to improve analysis of biological samples for biomarkers characterization in metabolomic strategies.

Application of NMR spectroscopy, coupled with pattern recognition methods, to two biological issues is reported: a) the progressive liver alterations during tumorigenesis and b) the exhaled breath condensate of patients with airway diseases.

In our first application, we investigated the metabolite composition of human hepatic tissue extracts of 17 patients affected by hepatocellular carcinoma HCV-related (HCC), and 9 patients affected by liver metastases from colorectal carcinoma (MET-CRC); as a control, we used cirrhotic liver tissues of HCC patients (CIR) and normal liver tissue of MET-CRC patients (NT), respectively. PCA, together with OPLS-DA analysis, allowed spectral classes clustering and classification. All spectra were visualized by scores and loadings plots, which also highlighted the "evolution" and relationship of the different pathological liver conditions represented by the four NMR data classes. The disease evolution clearly followed the increase of the lactate together with the remarkable decrease of the glucose signal, thus suggesting that such a signal pattern may act as a potential marker for assessing pathological hepatic lesions. In particular, we identified a statistical model that could be used to distinguish hepatic metastasis and human hepatocarcinoma from a "normal" (healthy) hepatic tissue. The progressive increase of lactate/glucose ratio, within the hepatic tissues, is consistent with the enhanced conversion of glucose into lactate, through the different classes that represent different tissue conditions such as hypoxia and/or "aerobic glycolisis". Although this trend is generally known, as considered the result of oncogenic alteration in glucose metabolism following malignant transformation, we reported a further information which is the extreme lactate/glucose conversion showed by MET-CRC, compared with all of the others tissue samples under investigation. Indeed, metastasis formation is the result of a multi-step cascade of events occurring to cancer cells during tumor dissemination, which brings about considerable metabolic changes. The large increase in lactate concentration as well as the decrease of intracellular glucose level was the predominant effect for the separation of metastases from HCC and NT, and the lactate/glucose ratio in MET-CRC ranges from 9 to 40 fold higher compared to HCC and NT, respectively, thus suggesting a role for the enhanced phenomenon of "aerobic glycolysis".

A further application was addressed to investigate the ^{1}H-NMR metabolite profile of exhaled breath condensate (EBC) of patients with different airway diseases. EBC, obtained by cooling exhaled air from spontaneous breathing, is a simple, noninvasive and useful tool to study the biochemical and inflammatory molecules in the airway lining fluid. Thirtysix paired EBC and saliva samples, obtained from healthy subjects, laryngectomized patients and chronic obstructive pulmonary disease (COPD) patients, were analyzed applying ^{1}H-NMR spectroscopy followed by principal component analysis. Our aim was to assess the role of pre-analytical variables (saliva and disinfectant contamination), potentially influencing EBC, to evaluate the stability and reproducibility of samples and to discriminate healthy subjects from patients with airway disease. The results show that saliva metabolic profile is significantly different from the EBC samples and that the presence of identical metabolites in EBC and saliva does not hamper discrimination. Excluding the carbohydrate signals (absent in EBC), saliva spectra clearly differ from EBC, notwithstanding the presence of some common metabolites (leucine, lactate, propionate, acetate, etc.). Furthermore, by examining EBC from laryngectomized patients, which may represent a true saliva-free material from the lower airways, we found that in those subjects all saliva spectra strictly differed from corresponding EBC samples. Importantly, all EBC and saliva collected twice within the same day (12 h apart) showed good within-day repeatability. Finally, we could state that saliva contamination may play a minor role in the interpretation of EBC by NMR-based metabonomics. Furthermore, we considered the influence of external contaminants, as the International Consensus on EBC recommends special care in the disinfection of reusable parts of condensers. Upon standard cleaning, all EBC spectra presented signals corresponding to unknown inactive substances of the disinfectant, that completely disappeared only after washing the reusable parts with 96% ethanol. Afterwards, by selecting specific non-contaminated regions of EBC spectra for statistical analysis, an efficient discrimination of EBC subjects with airway diseases (COPD) from subjects without respiratory diseases, was obtained. Some NMR signals appear to differentiate "respiratory" (COPD) from "non-respiratory" (HS and laryngectomized) subjects, by showing both quantitative (signal intensity) and qualitative (signal absence/presence) differences; among all pyruvate, succinate, glutamine, TMAO, choline and phosphorylcholine.

As a further enhanced tool for high thoughput NMR analysis, we developed a new integration method for 2D NMR spectra quantification, which is fundamental both in metabonomics and in the structure determination of biomolecules. Quantitative information from multidimensional NMR experiments can be obtained by peak volume integration. However, the standard procedure of selecting a region around the chosen peak and addition of all values is often biased by poor peak definition and/or the degeneracy of the NMR resonance frequency, a factor that aggravates with the increasing size of macromolecules and the number of metabolites. In this thesis, we developed and tested a simple method, called CAKE, for volume integration of moderately-to-strongly overlapping peaks, using the Monte Carlo Hit-or-Miss techniques, relying upon the peak line shapes in two-dimensional NMR. The CAKE approach uses the symmetry of a single in-phase peak (a peak with a unique center corresponding to its maximum) to calculate its volume. It is obtained by multiplying the fractional volume by the R factor, a proportionality ratio between the total and the fractional volume, both evaluated with Monte Carlo techniques. Therefore, the peak volume can be estimated by integrating a known fraction of the peak, and the fractional volume can be chosen so as to minimize the effect of overlap in complex NMR spectra. All integration of simulated and experimental 2D in-phase peaks, with different degree of overlap, showed the CAKE efficacy in estimating umbiased peak volume, even for strongly overlapping peaks. Moreover, it is substantially independent on digital resolution and SNR.

Finally, we successfully investigated the possibility of exploiting enhanced NMR pulse sequences for fast spectra acquisition. In particular, we applied the so-called SOFAST-HMQC pulse scheme to detect in-cell metabolism. Created and designed for protein observation, the pulse sequence is based upon very short experimental recycle delays, which, of course, rely on short T_1 relaxations time. Even if metabolites are often characterized by T_1 relaxations time longer than those of proteins, we have applied the SOFAST experiment to ^{15}N-labeled *Thalassiosira rotula* diatom cells obtaining, to the best of our knowledge, the first application of fast NMR spectroscopy. We collected spectra in 10-15 s of acquisition time, pinpointing the *T. rotula* ^{1}H-^{15}N metabolic profiling directly in living cells. Our results, definitively show that the application of SOFAST experiments provides an instantaneous picture of the metabolic pathways occurring in a well-defined physiological state, therefore avoiding the observation of an "average" metabolism obtainable with acquisition time of hours. With this approach, biochemical processes, taking place during metabolic modifications, can be followed by real-time multidimensional NMR methods, where spectral changes are monitored during a very short temporal period. In the past, the long acquisition times associated with 2D NMR have limited the application of real-time 2D NMR to slow kinetic processes with characteristic time constants of minutes to hours. The introduction of fast 2D data acquisition schemes, such as the SOFAST experiments, could extend the time window accessible to real-time 2D NMR to the range of seconds, thus representing a further advantaging tool for metabonomic and biomarkers investigations. Obviously, it would be extremely advantageous to extend the described investigations to eukaryotic cell systems, where potential applications include in-cell investigation under physiological or stressing conditions, induced by external toxicants or potential drugs, NMR metabolic characterization of cell lines, as well as investigation of the metabolism in plant cells. In general, extensive application in the fields of metabolomics and metabonomics can be predicted, and many of the above applications are in progress in our laboratory.

Acknowledgements

I would like to thank Prof. Rosa Lanzetta ("Federico II" University of Napoli), my tutor, for her many suggestions and constant support during this research.

I am also thankful to Dr. Andrea Motta (Institute of Biomolecular Chemistry, ICB-CNR, Pozzuoli), my co-tutor, for his guidance and support through these years of work.

Thanks to Prof. Livio Paolillo ("Federico II" University of Napoli), who expressed his interest in my work and supported me with interesting scientific discussions and useful suggestions.

A special thank goes to Dominique Melck (ICB-CNR, Pozzuoli), who followed me since my first laboratory activities being crucial to the successful completion of this project.

Dr. Matteo Stocchero (S-IN, Soluzioni Informatiche, Vicenza), who gave me a better perspective on my own results by sharing with me his knowledge and competence on multivariate data analysis and providing many useful references.

I am also thankful to Dr. Rocco Romano (University of Salerno), for his scientific support and collaboration for the joint works.

I would like to thank Dr. Pina Andreotti, (ICB-CNR, Pozzzuoli) for her many suggestions and scientific support.

I had the pleasure of work with Emilio Castelluccio, Alessandro Esposito, Enzo Mirra (ICB-CNR, Pozzuoli) who supplied me with their precious technical support.

Napoli, Italy
December 2009

Debora Paris

Bibliography

[1] J.K. Nicholson, J. Connelly, J.C. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Rev. Drug Disc.*, 1:153–161, 2002. vii, 2, 14, 15

[2] P. Shanda and B. Brutscher. Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. *J. Am. Chem. Soc.*, 127:8014–8015, 2005. viii, 26, 27, 30, 34, 87

[3] P. Shanda, Ē. Kupče, and B. Brutscher. SOFAST-HMQC experiments for recording two-dimensional heteronuclear correlation spectra of proteins within a few seconds. *J. Biomol. NMR*, 33:199–211, 2005. viii, 26, 27, 30, 34, 87

[4] J.K. Nicholson and I.D. Wilson. High resolution proton magnetic resonance spectroscopy of biological fluids. *Prog. Nucl. Magn. Reson. Spectrosc.*, 21:449–501, 1989. 2

[5] J.C. Miller and J.N. Miller. *Statistics and chemometrics for analytical chemistry.* Prentice Hall, Harlow, UK, 4th edition, 2000. 3

[6] M. Otto. *Chemometrics: statistics and computer application in analytical chemistry.* Wiley-VCH, Weinheim, Germany, 1998. 3

[7] K.R. Beebe, R.J. Pell, and M.B. Seasholtz. *Chemometrics: a practical guide.* Wiley-VCH, Wiley, New York, 1998. 3

[8] R. Kramer. *Basic chemometrics: a practical introduction to quantitative analysis.* Wiley-VCH, Wiley, New York, 1995. 3

[9] A. Staib, B. Dolenko, D.J. Fink, J. Fruh, A.E. Nikulin, M. Otto, M.S. Pessin-Minsley, O. Quarder, R. Somorjai, U. Thienel, G. Werner, and W. Petrich. Disease pattern recognition testing for rheumatoid arthritis using infrared spectra of human serum. *Clin. Chim. Acta*, 308:79–89, 2001. 4

[10] W. Petrich, A.Staib, M. Otto, and R.L. Somorjai. Correlation between the state of health of blood donors and the corresponding mid-infrared spectra of the serum. *Vib. Spectrosc.*, 28:117–129, 2002. 4

[11] M. Barker and W. Rayens. Partial least squares for discrimination. *J. Chemom.*, 17:166–173, 1993. 7

[12] D.M. Haaland, H.D.T. Jones, and E.V. Thomas. Multivariate classification of the infrared spectra of cell and tissue samples. *Appl. pectrosc.*, 51:340–345, 1997. 8

[13] J. Trygg and S. Wold. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.*, 17:53–64, 2003. 10, 15

[14] S. Wiklund, E. Johansson, L. Sjöström, E.J. Mellerowicz, U. Edlund, J.P. Shockcor, J. Gottfries, T. Moritz, and J. Trygg. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal. Chem.*, 80:115–122, 2008. 10, 15, 51

[15] M. Thomas and A. Zhu. Hepatocellular carcinoma: the need for progress. *J. Clin. Oncol.*, 23:2892–2899, 2005. 13, 57

[16] S.F. Altekruse, K.A. McGlynn, and M.E. Reichman. Hepatocellular carcinoma incidence, survival trends in the United States. *J. Clin. Oncol.*, 27:1485–1491, 2009. 13

[17] M.P. Curado, B. Edwards, H.R. Shin, H. Storm, J. Ferlay, M. Heanue, and P. Boyle. Cancer incidence in five continents. In *International Agency for Research on Cancer (IARC) Scientific Publications*, volume IX of *IARC*, pages 203–210, Lyon, France, 2007. Scientific Publications. 13

[18] G.C. Mueller, H.K. Hussein, R.C. Carlos, H.V. Nghiem, and I.R. Francis. Effectiveness of MR imaging in characterizing small hepatic lesions: routine versus export interpretation. *Am. J. Roentgenol.*, 180:673–680, 2003. 13

[19] S.A. Patterson, H.I. Khalil, and D.M. Panicek. MRI evaluation of small hepatic lesions in women with breast cancer. *Am. J. Roentgenol.*, 187:307–312, 2006. 13

[20] D.C. Chieng. Fine needle aspiration biopsy of liver - an update. *World J Surg Oncol*, 2, 2004. 13, 14

[21] G. Atalay, L. Biganzoli, F. Renard, R. Paridaens, T. Cufer, R. Coleman, A. H. Calvert, T. Gamucci, A. Minisini, P. Therasse, and M. J.

Piccart. Clinical outcome of breast cancer patients with liver metastases alone in the anthracycline-taxane era: a retrospective analysis of two prospective, randomized metastatic breast cancer trials. *Eur. J. Cancer.*, 39:2439–2449, 2003. 14

[22] Y. Yang, C. Li, X. Nie, X. Feng, W. Chen, Y. Yue, H. Tang, and F. Deng. Metabonomic studies of human hepatocellular carcinoma using high-resolution magic-angle spinning ^{1}H NMR spectroscopy in conjunction with multivariate data analysis. *J. Proteome Res.*, 6:2605–2614, 2007. 14

[23] J.K. Nicholson, J.C. Lindon, and E. Holmes. 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29:1181–1189, 1999. 14, 15

[24] J.C. Lindon, E. Holmes, and J.K. Nicholson. Pattern recognition methods and applications in biomedical magnetic resonance. *Prog. Nucl. Mag. Res. Sp.*, 39:1–40, 2001. 14

[25] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold. *Multi- and Megavariate Data Analysis, Basic principles and applications*. Umetrics AB, Umeå, 2006. 15

[26] C.E. Mountford, S. Doran, C.L. Lean, and P. Russell. Proton MRS can determine the pathology of human cancers with a high level of accuracy. *Chem. Rev.*, 104:3677–3704, 2004. 15

[27] S.G. Cho, M.Y. Kim, H.J. Kim, Y.S. Kim, W. Choi, S.H. Shin, K.C. Hong, Y.B. Kim, J.H. Lee, and C.H. Suh. Chronic hepatitis: in vivo proton MR spectroscopic evaluation of the liver and correlation with histopathologic findings. *Radiology*, 226:288–289, 2003. 15

[28] R. Soper, R.U. Himmelreich, D. Painter, R.L. Somorjai, C.L. Lean, B. Dolenko, C.E. Mountford, and P. Russell. Pathology of hepatocellular carcinoma and its precursors using proton magnetic resonance spectroscopy and a statistical classification strategy. *Pathology*, 34:417–422, 2002. 15

[29] L.M. Foley, R.A. Towner, and D.M. Painter. *In vivo* image-guided ^{1}H-magnetic resonance spectroscopy of the serial development of hepatocarcinogenesis in an experimental animal model. *Biochem. Biophys. Acta*, 1526:230–236, 2001. 15

[30] J.P. Usenius, R.A. Kauppinen, P.A. Vainio, J.A. Hernesniemi, M.P. Vapalahti, L.A. Paljärvi, and S. Soimakallio. Quantitative metabolite patterns of human brain tumors: detection by ^{1}H NMR spectroscopy *in vivo* and *in vitro*. *J. Comput. Assist. Tomogr.*, 18:705–713, 1994. 15

[31] S.A. Kharitonov and P.J. Barnes. Exhaled markers of inflammation. *Curr. Opin. Allergy Clin. Immunol.*, 1:217–221, 2001. 15, 70

[32] I. Horváth, J. Hunt, and P.J. Barnes. Exhaled breath condensate: methodological recommendations and unresolved questions. *Eur. Respir. J.*, 26:523–548, 2005. 15, 70

[33] M. Maniscalco, G. de Laurentiis, and C. Pentella. Exhaled breath condensate as matrix for toluene detection: a preliminary study. *Eur. Respir. J.*, 11:233–240, 2006. 15

[34] J.C. Lindon, E. Holmes, and J.K. Nicholson. Metabonomics in pharmaceutical R & D. *FEBS J.*, 274:1140–1151, 2007. 15, 17

[35] O. Beckonert, H.C. Keun, T.M.D. Ebbels, J. Bundy, E. Holmes, J.C. Lindon, and J.K. Nicholson. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.*, 2:2692–2703, 2007. 15, 36

[36] J.D. Bell, J.C. Brown, and P.J. Sadler. NMR studies of body fluids. *NMR Biomed.*, 2:246–256, 1989. 15

[37] E. Holmes, P.J. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, and M. Spraul. Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal. Biochem.*, 220:284–296, 1994. 15, 61

[38] C.J. Silwood, E. Lynch, A.W. Claxson, and M.C. Grootveld. ^{1}H and ^{13}C NMR spectroscopic analysis of human saliva. *J. Dent. Res.*, 81:422–427, 2002. 15, 61, 63

[39] M. Grootveld and C.J. Silwood. ^{1}H NMR analysis as a diagnostic probe for human saliva. *Biochem. Biophys. Res. Commun.*, 329:1–5, 2005. 15, 62

[40] S. Carraro, S. Rezzi, and F. Reniero. Metabolomics applied to exhaled breath condensate in childhood asthma. *Am. J. Respir. Crit. Care Med.*, 175:986–990, 2007. 16, 60, 67, 71

[41] P. Latzin, J. Beck, A. Bartenstein, and M. Griese. Comparison of exhaled breath condensate from nasal and oral collection. *Eur. J. Med. Res.*, 8:505–510, 2003. 16

[42] J. Chladkova, I. Krcmova, J. Chladek, P. Cap, S. Micuda, and Y. Hanzalkova. Validation of nitrite and nitrate measurements in exhaled breath condensate. *Respiration*, 73:173–179, 2006. 16

[43] J.L. Griffin. The Cinderella story of metabolic profiling: does metabolomics get to go to the functional genomics ball? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 361:147–161, 2006. 17

[44] J. Cavanagh, W.J. Fairbrother, A.G. Palmer 3rd, N. J. Skelton, and M. Rance. *Protein NMR Spectroscopy: Principles and Practice*. Elsevier Academic Press, Burlington, MA, USA, 2nd edition, 2007. 17, 29

[45] J.J. Led and H. Gesmar. Quantitative information from complicated nuclear magnetic resonance spectra of biological macromolecules. *Methods Enzymol.*, 239:318–345, 1994. 17, 18, 20, 21

[46] G.H. Weiss, J.E. Kiefer, and J.A. Ferretti. Accuracy and precision in the estimation of peak areas: The effects of apodization. *Chemom. Intell. Lab. Sys.*, 4:223–229, 1988. 18

[47] C. Rischel. Fundamentals of peak integration. *J. Magn. Reson. A*, 116:255–258, 1995. 18

[48] V. Stoven, A. Mikou, D. Piveteau, E. Guitettet, and J. Y. Lallemand. Paris, a Program for Automatic Recognition and Integration of 2D NMR Signals. *J. Magn. Reson.*, 82:163–169, 1989. 18

[49] H. Shen and F.M. Poulsen. Toward automated determination of build-up rates of Nuclear Overhauser Effects in proteins using symmetry projection operators. *J. Magn. Reson.*, 89:585–588, 1990. 18

[50] S. Glaser and H.R. Kalbitzer. Automated recognition and assessment of cross peaks in two-dimensional NMR spectra of macromolecules. *J. Magn. Reson.*, 74:450–463, 1987. 18

[51] K.P. Neidig and H.R. Kalbitzer. Improved representation of two-dimensional NMR spectra by local rescaling. *J. Magn. Reson.*, 88:155–161, 1990. 18

[52] M. Geyer, K.P. Neidig, and H.R. Kalbitzer. Automated peak integration in multidimensional NMR spectra by an optimized iterative segmentation procedure. *J. Magn. Reson. B*, 109:31–38, 1995. 18

[53] W. Denk, R. Baumann, and G. Wagner. Quantitative evaluation of cross-peak intensities by projection of two-dimensional NOE spectra on a linear space spanned by a set of reference resonance lines. *J. Magn. Reson.*, 67:386–390, 1986. 18

[54] T.A. Holak, J. N. Scarsdale, and J.H. Prestegard. A simple method for quantitative evaluation of cross-peak intensities in two-dimensional NOE spectra. *J. Magn. Reson.*, 74:546–549, 1987. 18

[55] C. Eccles, P. Guntert, M. Billeter, and K. Wüthrich. Efficient analysis of protein 2D NMR spectra using the software package EASY. *J. Biomol. NMR*, 1:111–118, 1991. 18

[56] H. Gesmar, P. F. Nielsen, and J.J. Led. Simple least-squares estimation of intensities of overlapping signals in 2D NMR spectra. *J. Magn. Reson. B*, 103:10–18, 1994. 18

[57] R.R. Ernst, G. Bodenhausen, and A. Wokaun. *Principles of nuclear magnetic resonance in one and two dimensions.* Clarendon Press, Oxford, 1987. 18, 19, 20, 21, 27, 29, 34

[58] R. Koradi, M. Billeter, M. Engeli, P. Güntert, and K. Wüthrich. Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. *J. Magn. Reson.*, 135:288–297, 1998. 18

[59] J.J. Led and H. Gesmar. Application of the linear prediction method to NMR spectroscopy. *Chem. Rev.*, 91:1413–1426, 1991. 20, 21

[60] H. Gesmar, J.J. Led, and F. Abildgaard. Improved methods for quantitative spectral analysis of NMR data. *Prog. NMR Spectrosc.*, 22:255–288, 1990. 20, 21

[61] J.C. Lindon and A.G. Ferrige. Digitisation and data processing in Fourier transform NMR. *Prog. NMR Spectrosc.*, 14:27–66, 1980. 21

[62] G.A. Pearson. Optimization of gaussian resolution enhancement. *J. Magn. Reson.*, 74:541–545, 1987. 21

[63] R.Y. Rubinstein. *Simulation and the Monte Carlo Method.* Wiley, New York, NY, 1981. 23

[64] M.H. Kalos and P.A. Whitlock. *Monte Carlo Methods, Volume 1: Basics.* Wiley, New York, NY, 1986. 23

[65] R.R. Ernst and W.A. Anderson. Sensitivity enhancement in magnetic resonance. II. Investigation of intermediate passage conditions. *Rev.Sci.Instr.*, 37:93, 1966. 25

[66] J. Jeener. *J Ampere International School*, 1971. 25

[67] W.P. Aue, E. Bartholdi, and R.R. Ernst. Two-dimensional spectroscopy. application to nuclear magnetic resonance. *J.Chem.Phys.*, 64:2229–2235, 1976. 25

[68] J.C. Hoch and A.S. Stern. Maximum entropy reconstruction, spectrum analysis and deconvolution in multidimensional nuclear magnetic resonance. *Nucl. Mag. Reson. Biol. Macromol.*, 338:159–178, 2001. 26

[69] V.A. Mandelshtam. The multidimensional filter diagonalization method: I. Theory and numerical implementation. *J. Magn. Reson.*, 144:343–356, 2000. 26

[70] T. Szyperski, G. Wider, J.H. Bushweller, and K. Wuthrich. Reduced dimensionality in triple-resonance NMR experiments. *J. Am. Chem. Soc.*, 115:9307–9308, 1993. 26

[71] B. Brutscher, J.P. Simorre, M.S. Caffrey, and D. Marion. Design of a complete set of two-dimensional triple-resonance experiments for assigning labeled proteins. *J. Magn. Reson. B*, 105:77–82, 1994. 26

[72] S. Kim and T. Szyperski. GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. *J. Am. Chem. Soc.*, 125:1385–1393, 2003. 26

[73] Ē. Kupče and R. Freeman. Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. *J. Am. Chem. Soc.*, 126:6429–6440, 2004. 26

[74] Ē. Kupče, T. Nishida, and R. Freeman. Hadamard NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.*, 42:95–122, 2003. 26

[75] B. Brutscher. Combined frequency- and time-domain NMR spectroscopy. Application to fast protein resonance assignment. *J. Biomol. NMR*, 29:57–64, 2004. 26

[76] L. Frydman, T. Scherf, and A. Lupulescu. The acquisition of multidimensional NMR spectra within a single scan. *Proc. Natl. Acad. Sci. USA*, 99:15858–15862, 2002. 27, 34

[77] K. Pervushin, B. Vogeli, and A. Eletsky. Longitudinal ^{1}H relaxation optimization in TROSY NMR spectroscopy. *J.Am.Chem.Soc.*, 124:12898–12902, 2002. 27, 28, 30, 32, 34, 87

[78] A. Ross, M. Salzmann, and H. Senn. Fast-HMQC using Ernst angle pulses: An efficient tool for screening of ligand binding to target proteins. *J. Biomol. NMR*, 10:389–396, 1997. 27, 34, 87

[79] H.S. Atreya and T. Szyperski. G-matrix fourier transform NMR spectroscopy for complete protein resonance assignment. *Proc. Natl. Acad. Sci. USA*, 101:9642–9647, 2004. 28

[80] D.D. Traficante. Optimum tip angle and relaxation delay for quantitative analysis. *Conc. Magn. Reson.*, 4:153–160, 1992. 29

[81] M. Piotto, V. Saudek, and V. Sklenář. Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR*, 2:661–665, 1992. 30

[82] Ē. Kupče, J. Boyd, and I.D. Campbell. Short selective pulses for biochemical applications. *J. Magn. Reson. B*, 106:300–303, 1995. 30, 31, 87

[83] H. Geen and R. Freeman. Band-selective radiofrequency pulses. *J. Magn. Reson. A*, 93:93–141, 1991. 31, 87

[84] L. Emsley and G. Bodenhausen. Optimization of shaped selective pulses for NMR using a quaternion description of their overall propagators. *J. Magn. Reson.*, 97:135–148, 1992. 31

[85] Ē. Kupče and R. Freeman. Wideband excitation with polychromatic pulses. *J. Magn. Reson. A*, 108:268–273, 1994. 31

[86] Z. Serber, P. Selenko, R. Hänsel, S. Reckel, F. Löhr, J.E.Jr. Ferrell, G. Wagner, and V. Dötsch. Investigating macromolecules inside cultured and injected cells by in-cell NMR spectroscopy. *Nat. Protoc.*, 1:2701–2709, 2006. 34, 92, 95

[87] R. Freeman and Ē. Kupče. New methods for fast multidimensional NMR. *J. Biomol. NMR*, 27:101–113, 2003. 34

[88] Shrot Y. and L. Frydman. Single-scan NMR spectroscopy at arbitrary dimensions. *J. Am. Chem. Soc.*, 125:11385–11396, 2003. 34

[89] P. Shanda and B. Brutscher. Hadamard frequency-encoded SOFAST-HMQC for ultrafast two-dimensional protein NMR. *J. Magn. Reson.*, 178:334–339, 2006. 34

[90] P. Shanda, H. Van Melckbeke, and B. Brutscher. Speeding up three-dimensional protein NMR experiments to a few minutes. *J. Am. Chem. Soc.*, 128:9042–9043, 2006. 34

[91] J.K. Nicholson and I.D. Wilson. Understanding 'global' systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.*, 2:668–677, 2003. 36

[92] J.C. Lindon, J.K. Nicholson, E. Holmes, H.C. Keun, A. Craig, J.T. Pearce, S.J. Bruce, N. Hardy, S.A. Sansone, H. Antti, P. Jonsson, C. Daykin, M. Navarange, R.D. Beger, E.R. Verheij, A. Amberg, D. Baunsgaard, G.H. Cantor, L. Lehman-McKeeman, M. Earll, S. Wold, E. Johansson, J.N. Haselden, K. Kramer, C. Thomas, J. Lindberg, I. Schuppe-Koistinen, I.D. Wilson, M.D. Reily, D.G. Robertson, H. Senn, A. Krotzky, S. Kochhar, J. Powell, F. van der Ouderaa, R. Plumb, H. Schaefer, and M. Spraul. Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol.*, 23:833–838, 2005. 38, 86, 93

[93] C.Y. Lin, H. Wu, R.S. Tjeerdema, and M.R. Viant. Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics. *Metabolomics*, 3:55–67, 2007. 38

[94] J.E. Le Belle, N.G. Harris, S.R. Williams, and K.K. Bhakoo. A comparison of cell and tissue extraction techniques using high-resolution ^{1}H-NMR spectroscopy. *NMR Biomed.*, 15:37–44, 2002. 38

[95] T.-L. Hwang and A.J. Shaka. Water suppression that works: excitation sculpting using arbitrary waveforms and pulse field gradients. *J. Magn. Reson.*, 112:275–279, 1995. 39, 79, 87

[96] L. Braunschweiler and R.R. Ernst. Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. *J. Magn. Reson.*, 53:521–528, 1983. 39

[97] C. Griesinger, G. Otting, K. Wüthrich, and R.R. Ernst. Clean TOCSY for proton spin system identification in macromolecules. *J. Am. Chem. Soc.*, 110:7870–7872, 1988. 39, 87

[98] A. Bax and D. Davis. MLEV-17 based two-dimensional homonuclear magnetization transfer spectroscopy. *J. Magn. Reson.*, 65:355–360, 1985. 39, 79

[99] A.G. Palmer III, J. Cavanagh, P.E. Wright, and M. Rance. Sensitivity improvement in proton detected heteronuclear correlation experiments. *J. Magn. Reson.*, 93:151–170, 1991. 39

[100] L.E. Kay, P. Keifer, and T. Saarinen. Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity. *J. Am. Chem. Soc.*, 114:10663–10665, 1992. 39

[101] T.W.M. Fan. Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures. *Prog. Nucl. Mag. Res. Sp.*, 28:161–219, 1996. 41, 63

[102] R.E. London. ^{13}C labelling in studies of metabolic regulation. *Prog. Nucl. Mag. Res. Sp.*, 20:337–383, 1988. 41

[103] M. Hockel and P. Vaupel. Tumor hypoxia: definitions and current clinical, biologic, and molecular aspects. *J. Natl. Cancer. Inst.*, 93:266–276, 2001. 56

[104] S. Walenta, T. Schroeder, and W. Mueller-Klieser. Lactate in solid malignant tumors: potential basis of a metabolic classification in clinical oncology. *Curr. Med. Chem.*, 11:2195–2204, 2004. 56

[105] O. Warburg. On the origin of cancer cells. *Science*, 123:309–314, 1956. 56

[106] R.A. Gatenby and R.J. Gillies. Why do cancers have high aerobic glycolysis? *Nat. Rev. Cancer.*, 4:891–899, 2004. 56

[107] X.L. Zu and M. Guppy. Cancer metabolism: facts, fantasy, and fiction. *Biochem. Biophys. Res. Commun.*, 81:130–135, 2004. 56

[108] V. Quennet, A. Yaromina, D. Zips, A. Rosner, S. Walenta, and M. Baumann. Tumor lactate content predicts for response to fractionated irradiation of human squamous cell carcinomas in nude mice. *Radiother. Oncol.*, 81:130–135, 2006. 56

[109] G. Christofori. New signals from the invasive front. *Nature*, 441:444–450, 2006. 58

[110] V. Seenu, M. Kumar, U. Sharma, S. Datta Gupta, S.N. Metha, and N.R. Jagannathan. Potential of magnetic resonance spectroscopy to detect metastasis in axillary lymph nodes in breast cancer. *Magn. Reson. Imaging*, 23:1005–1010, 2005. 58

[111] L.L. Cheng, I.W. Chang, B.L. Smith, and R.G. Gonzalez. Evaluating human breast ductal carcinomas with high-resolution magic-angle spinning proton magnetic resonance spectroscopy. *J. Magn. Reson.*, 135:194–102, 1998. 58

[112] M.R. Viant. Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Comm.*, 310:943–948, 2003. 58

[113] K.F. Rabe, S. Hurd, and A. Anzueto *et al.* Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. *Am. J. Respir. Crit. Care Med.*, 176:532–555, 2007. 59

[114] R.M. Effros. Exhaled breath condensate ph. *Am. J. Respir. Crit. Care Med.*, 173:1047–1048, 2006. 61

[115] R.M. Effros, K.W. Hoagland, and M. Bosbous *et al.* Dilution of respiratory solutes in exhaled condensates. *Am. J. Respir. Crit. Care Med.*, 165:663–669, 2002. 61

[116] C.L. Whittle, S. Fakharzadeh, and J. Eades ad G. Preti. Human breath odors and their use in diagnosis. *Ann. N. Y. Acad. Sci.*, 1098:252–266, 2007. 61, 70

[117] R.M. Effros. Metabolomics in exhaled breath condensates. *Am. J. Respir. Crit. Care Med.*, 177:236, 2008. 61

[118] K. Wthrich. *NMR of Protein and Nucleic Acids.* Wiley and Sons, New York, 1986. 61

[119] P. Montuschi. Exhaled breath condensate analysis in patients with copd. *Clin. Chim. Acta*, 356:22–34, 2005. 71

[120] P. Montuschi, S.A. Kharitonov, G. Ciabattoni, and P.J. Barnes. Exhaled leukotrienes and prostaglandins in copd. *Thorax*, 58:585–588, 2003. 71

[121] S. Heili-Frades, G. Peces-Barba, and M.J. Rodriguez-Nieto *et al.* Metabonomic approach for non invasive diagnosis of inflammatory lung diseases through nuclear magnetic resonance analysis of exhaled breath condensate. *Eur. Respir. J.*, 30:38s, 2007. 71

[122] C. Griesinger, G. Otting, K. Wüthrich, and R.R. Ernst. Clean TOCSY for proton spin system identification in macromolecules. *J. Am. Chem. Soc.*, 110:7870–7872, 1988. 79

[123] J.C. Cobas and F.J. Sardina. Nuclear magnetic resonance data processing. MestRe-C: A software package for desktop computers. *Concepts Magn. Reson.*, 19A:80–96, 2003. 79

[124] A. Miralto, G. Barone, G. Romano, S.A. Poulet, A. Ianora, G.L. Russo, I. Buttino, G. Mazzarella, M. Laabir, M. Cabrini, and M.G. Giacobbe. The insidious effect of diatoms on copepod reproduction. *Nature*, 402:173–176, 1999. 85

[125] Ē. Kupče and R. Freeman. Polychromatic selective pulses. *J. Magn. Reson. A*, 102:122–126, 1993. 87

[126] A.J. Shaka, P.B. Barker, and R. Freeman. Computer-optimized decoupling scheme for wideband applications and low-level operation. *J. Magn. Reson.*, 64:547–552, 1985. 88

[127] K. Kanaori, T.L. Legerton, R.L. Weiss, and J.D. Roberts. Nitrogen-15 spin-lattice relaxation times of amino acids in *Neurospora crassa* as a probe of intracellular environment. *Biochemistry*, 21:4916–4920, 1982. 88, 89

[128] S.P. Williams, P.M. Haggie, and K.M. Brindle. ^{19}F NMR measurements of the rotational mobility of proteins in vivo. *Biophys.*, 72:490–498, 1997. 88