

# TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Ökologische Chemie und Umweltanalytik

## Datamining metabolomics: the convergence point of non-target approach and statistical investigation

Marianna Lucio

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. W. Huber

Prüfer der Dissertation:

1. Priv.-Doz. Dr. Dr. Ph. Schmitt-Kopplin
2. Univ.-Prof. Dr. Dr. h. c. H. Parlar
3. Univ.-Prof. Dr. K. Suhre, Ludwig-Maximilians-Universität München

Die Dissertation wurde am 18.09.2008 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 14.01.2009 angenommen.



to Eugenio



## Acknowledgements

I am grateful to my supervisor Philippe Schmitt Kopplin for giving me this opportunity and sharing enthusiasm and expertise.

I would like to thank all my colleagues of “BioGeomics” group for their support and for the synergy that makes it so special to work here.

Special thanks to Andras and Mourad for their incomparable collaboration and friendship.

I would like to express my gratitude to Istvan and Agi for their precious and continuous support.

The dissertation was prepared from July 1st 2005 to July 31st 2008 at the Institute of Ecological Chemistry at the Helmholtz Zentrum München National Research Center for Environment and Health in the Helmholtz Association in Neuherberg.



## PUBLICATIONS

### *Articles*

M. Lucio, P. Schmitt Kopplin (2006). Modelling the binding of triazine herbicides to humic substances using capillary electrophoresis. *Environ Chem. Letter*; 4:15-21

*Presented in the third Chapter*

R. Rosselló-Mora, M. Lucio, A. Pena, J. Brito-Echeverría, A. López López, M. Valens-Vadell, M. Frommberger, J. Antón and P. Schmitt-Kopplin (2008). Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. *Nature ISME Journal*; 2, 242-253.

*Presented in the fourth Chapter*

Gougeon R., M. Lucio, M. Frommberger, D. Peyron, D. Chassagne, Hervé Alexandre, F. Feuillat, P. Cayot, A. Voilley, I. Gebefügi, Norbert Hertkorn & Philippe Schmitt-Kopplin (2009). Wine insights: the chemodiversity of wines from data driven Oenomics “From Oenology to Systems Oenology”. In review *PNAS*.

*Presented in the fifth chapter*

M. Lucio, Gougeon R., A. De Boel, M. Frommberger, N. Hertkorn, D. Peyron, D. Chassagne, F. Feuillat, A. Voilley, I. Gebefügi, Ph. Schmitt-Kopplin (2008). Expressing forest origins in the chemical composition of oak wood barrels and corresponding wines by FTICR-MS. *Chemistry an European Journal*; 15, 600-611.

*Presented in the sixth chapter*

M. Harir, A. Gaspar, M. Frommberger, M. Lucio, M. El Azzouzi, D. Martens, A. Kettrup, and Ph. Schmitt-Kopplin (2007): Photolysis Pathway of Imazapic in Aqueous Solution: Ultrahigh Resolution Mass Spectrometry Analysis of Intermediates. *J. Agric. Food Chem.*; 10.1021/jf0720279

P. Schmitt-Kopplin, N. Hertkorn, M. Frommberger, M. Lucio, M. Englmann, A. Fekete, and I. Gebefügi (2007). Ion Cyclotron Resonance Fourier Transform Mass

Spectrometry for non-target metabolomics of molecular interactions in the Rhizosphere. *Soil Biology*; Vol. 11. *Advanced Techniques in Soil Microbiology*.

Gaspar A., M. Harir, M. Lucio, Ph. Schmitt-Kopplin (2008). Targeted borate complex formation as followed with electrospray Fourier transform ion cyclotron mass spectrometry: monomolecular model system and polyborate formation. *Electrophoresis*; 27, 66-79.

*Posters:*

M. Lucio, M. Frommberger, N. Hertkorn, I. Gebefugi, U. Schurr, K. Wieland, P. Schmitt-Kopplin (2006). An Interdisciplinary platform for plant and Rhizosphere metabolomics. Poster at: 4<sup>th</sup> International conference on Plant Metabolomic, Reading (England). October 5<sup>th</sup> to 7<sup>th</sup>

M. Lucio, S. Thaller, E. Holzmann, N. Hertkorn (2006). <sup>13</sup>C NMR relaxation analysis of Suwannee river fulvic acid. Poster at: 3<sup>rd</sup> European Symposium on NMR Spectroscopy, Freising (Germany).

R. Gougeon, A. De Boel, M. Lucio, M. Frommberger, D. Peyron, D. Chassagne, F. Feuillat, A. Voilley, P. Schmitt-Kopplin (2006). FTICR-MS analysis of cooperage OAK wood extracts. Poster at: 1<sup>st</sup> International Symposium on Ultrahigh Resolution Mass Spectrometry for the Molecular Level Analysis of Complex (BioGeo) Systems, 6/7 November 2006, GSF Neuherberg (Germany).

M. Frommberger, N. Hertkorn, M. Lucio, M. Englmann, I. Gebefugi, A. Hartmann, P. Schmitt-Kopplin (2006). Towards a better understanding of bacterial communication: Analysis of Quorum Sensing by capillary separation techniques coupled to mass spectrometry and by ultrahigh resolution FTICR mass spectrometry. Poster at: The International Workshop "Development and control of functional biodiversity at micro- and macro-scales" October 5<sup>th</sup> to 7<sup>th</sup> Neuherberg, (Germany).



M. Lucio, Ramon Rosselló-Mora, Josefa Antón Philippe Schmitt-kopplin (2008). Use of ICR-FT/MS to study the metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter* rubber. Poster at: 56<sup>th</sup> ASMS Conference on Mass Spectrometry, May 31 and June 1, Denver (Colorado).

M. Lucio, A. Fekete, P.Schmitt-Kopplin (2008). Data driven system biology based on a time dependant Metabolomic approach using ultrahigh resolution mass spectrometry (ICR-FT/MS). Poster at: EU/USA Microbiology Taskforce Workshop "Metabolomics and Environmental Biotechnology" (Invited Guest) - 16-17 June - Mallorca (Spain).

## Abbreviations and symbols

amu	Atomic mass unit
AV	Alveolar
AW	Bronchoalveolar
CE-MS	Capillary electrophoresis-mass spectrometry
CV	Cross validation
DModX	Distance to the model in X space (row residual SD), after A components
DModY	Distance to the model in Y space (row residual SD), after A components
EBC	Exhaled breath condensates
FT	Fourier Transform
GC-MS	Gas chromatography-mass spectrometry
HCA	Hierarchical cluster analysis
HPLC	High-performance liquid chromatography
ICR-FT/MS	Cyclotron Resonance Fourier Transform Ion Mass spectrometry
ISI	Insulin Sensitivity Indices
kDa	Kilo Dalton
KM	Kendrick Mass
LC-MS	Liquid chromatography-mass spectrometry
m/z	Mass-to-charge ratio
MS	Mass spectrometry
msec	Milli seconds
MSI	Metabolomics Standards Initiative
MW	Megaword (million bits in size)
NMR	Nuclear Magnetic Resonance
OCS	Orthogonal Signal Correction
OPLS	Orthogonal PLS
OPLS-DA	Orthogonal PLS discriminant analysis
p	Loadings
PC	Principal component
PCA	Principal Component Analysis

PLS	Partial least squares
PLS-DA	Partial least squares for discriminant analysis
ppb	Part per billion
ppm	Part per million
PRESS	Predictive residual sum of squares
psi	Pound per square inch
$Q^2$	Fraction of total variation that can be predicted
$R^2(Y)$	Fraction of sum of squares of Y explained by each component
RMSECV	Root mean square error of cross validation
RMSEP	Root mean square error of prediction
S/N	Signal to Noise
SD	Standard deviation
SS	Sum of squares
t	Scores
tPS	Predicted scores
UV	Unite Variance
VK	Van Krevelen
VIP	Variable importance in the projection
$w^*$	Weights that combine X variables to form the scores



## Table of contents

<b>1</b>	<b>Introduction .....</b>	<b>15</b>
<b>2</b>	<b>Metabolomics.....</b>	<b>23</b>
2.1	Metabolomics overview .....	23
2.1.1	One or more crucial metabolites: Biomarkers.....	25
2.2	Different technologies in metabolomics data .....	26
2.2.1	The novelty in the non-target analysis with ICR-FT/MS .....	29
2.3	Chemometrics .....	35
2.3.1	Exploratory data analysis .....	35
2.3.2	Classification modeling .....	36
<b>3</b>	<b>Strategy for large dataset.....</b>	<b>37</b>
3.1	Instruments for the analysis.....	37
3.2	Organizing the data.....	40
3.2.1	Formula calculation .....	44
3.3	Data transformation and normalization .....	47
3.4	Similarities and distances between the data .....	48
3.5	Statistical analysis .....	50
3.5.1	Multivariate analysis.....	52
3.5.1.1	Unsupervised analysis.....	53
3.5.1.2	Supervised analysis.....	59
3.6	Data validation .....	62
3.7	Data evaluation and visualization .....	65
3.7.1	Van Krevelen diagram .....	65
3.7.2	From masses to database .....	69
<b>4</b>	<b>Metabolic evidence for biogeographic isolation of the extremophilic bacterium Salinibacter rubber .....</b>	<b>71</b>
4.1	Introduction .....	71
4.2	Materials and methods .....	74

4.2.1	ICR-FT/MS procedure .....	75
4.3	Targeted approach .....	76
4.4	Statistical analysis .....	77
4.5	Inference on the biogeographic isolation .....	85
4.6	Conclusions .....	95
<b>5</b>	<b>Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines by ICR-FT/MS.....</b>	<b>97</b>
5.1	Introduction .....	97
5.2	Wood samples collection .....	101
5.2.1	Barrels and wine elaboration.....	102
5.2.2	Wood and wine samples preparation.....	102
5.3	ICR-FT/MS analysis.....	103
5.4	NMR Spectroscopy .....	103
5.4.1	Analysis of NMR spectra.....	104
5.5	Statistical analysis .....	104
5.6	Result and discussion.....	105
5.6.1	Wood differentiation .....	105
5.6.2	The species effect .....	108
5.6.3	The forest effect.....	115
5.6.4	Wood-wine correlations.....	121
5.7	Conclusion .....	124
<b>6</b>	<b>The chemodiversity of wines: From Oenology to “Systems Oenology”.....</b>	<b>127</b>
6.1	Introduction .....	127
6.1.1	<i>Oenolomics</i> : describing the chemical spaces of wine .....	130
6.1.2	From <i>oenolomics</i> to oenonomics and systems oenology .....	135
6.2	When systems oenology witnesses to the story that a wine tells.....	137
6.3	Methods .....	140
6.3.1	Statistical analyses .....	141

6.4	Discussion and conclusions .....	143
<b>7</b>	<b>Metabolomics approach in health.....</b>	<b>145</b>
7.1	Introduction .....	145
7.2	Metabolomic analysis of exhaled breath condensate for smokers, no-smokers COPD patients with ICR-FT/MS .....	146
7.2.1	Statistical elaboration.....	148
7.3	Metabolomic analysis of plasma of pre-diabetic patients with various insulin resistant index values .....	154
7.3.1	Pre-Diabetic state definition.....	155
7.4	Data analysis.....	156
7.5	Conclusion .....	165
<b>8</b>	<b>Conclusions and outlook .....</b>	<b>167</b>
<b>9</b>	<b>Bibliography .....</b>	<b>171</b>





# Chapter 1

## 1 INTRODUCTION

The fascinating world of metabolomics is enabling new and important discoveries, managing to be at the forefront in the life science like genomics together with proteomics. In the midst of this ongoing development, the awareness of the importance of metabolomics is being accomplished. Nowadays biology, medicine and the environmental sciences for studying living organisms are very likely to be studied with metabolomic approaches; because it offers a rapid, non targeted and effective way to diagnose illness and to monitor patient therapy. Moreover from an environmental point of view it is possible to determine the air quality measuring the pollution level, to derive the health state of the environment and furthermore to estimate the food quality. While reading this thesis, it will be possible to understand and assess the power of the tool of

metabolomics applied to different branches of science, opening the possibility to address this approach to several problematics without any limits or preconceptions, reaching different and new information. Metabolomics is growing very rapidly and integrates the knowledge of earlier developed omics-branches such as *genomics*, *proteomics* and *transcriptomics*.

From a traditional definition, in the field of human health, metabolomic measures the concentrations of the large number of naturally occurring small molecules (called metabolites), that are produced as intermediates and end-products of all metabolic processes (Bhalla, et al., 2005). They are measured from biological samples such as urine, saliva, blood plasma, tissue sample and even the simple breath can carry the information about the state of health.

In environmental issues the same approach can be followed looking holistically to all small molecules detectable in a given system in various scale, integrating thus metabolites from living organisms and all their biotic/abiotic transformation products.

The total number of different metabolites is still unknown; some estimation ranges from 200,000 (Ott, et al., 2006) to about 1,000,000 (Wink, 1988), but even this latter estimate may be conservative. Including plant and bacterial metabolites that are not necessary to keep the organism alive, also referred to as secondary metabolites, the number is enormously larger (Ott, et al., 2006). The probable number of metabolites is also considerably larger than the number of corresponding genes (Green, et al., 2008), so it seems that the currently available databases cover at best 2% of the total number of existing metabolites (Green, et al., 2008).

Low-molecular-weight compounds (<800 amu) are particularly interesting in the study of metabolomics, because they serve as substrates and products in the various metabolic pathways. These small molecules include compounds like sugars, lipids, amino acids, which provide important hints for the state of health.

Whereas genes and proteins set the stage for what can happen and what makes it happen in the cells (see figure 1.1) many of the actual activities, regulated by the metabolites, are at the metabolic level, like:

- cell multiplication
- energy transfer
- cell to cell communication

The information about the actual cellular environment can be retrieved looking more closely to the metabolite behavior. The environment of the cell is connected

to many exogenous factors like nutrition, drug, pollutant and many others. Thus, they can be used as tracers for human health predictions. In a similar manner, it is possible to analyze the effects of environmental stress (such as pollution and climate changes) in environment and *biogeography* metabolomic (Green, et al., 2008) studies, allowing for example mapping of metabolites as a result of organism adaption strategies to particular environments (pristine or polluted).

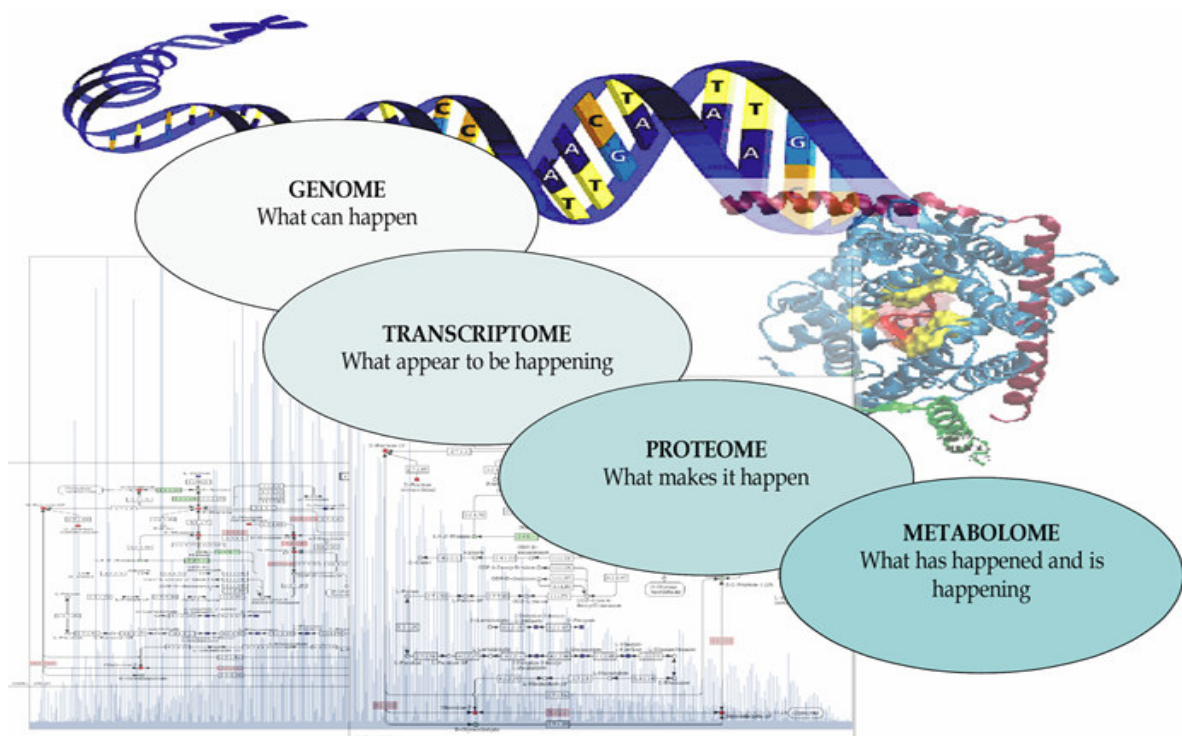


Figure 1.1: This figure, adapted from (Dettmer, et al., 2007), represents the technology's flow from genome to metabolome by time, space and under various stresses; (figures adapted from the US Department of Energy).

Figure 1.2 represents the fields in which the ICR-FT/MS is applied and the future possible depths, giving the possibility to create an “ICR-FT chain” in environment and health. At this point it is anticipated that the application of the *-omic* technologies and especially the study of what is happening will contribute to improve molecular diagnostics and will provide ‘deep’ insights into the pathogenic alterations in diseases or mechanisms of pharmacological

interventions (Bilello, 2005).

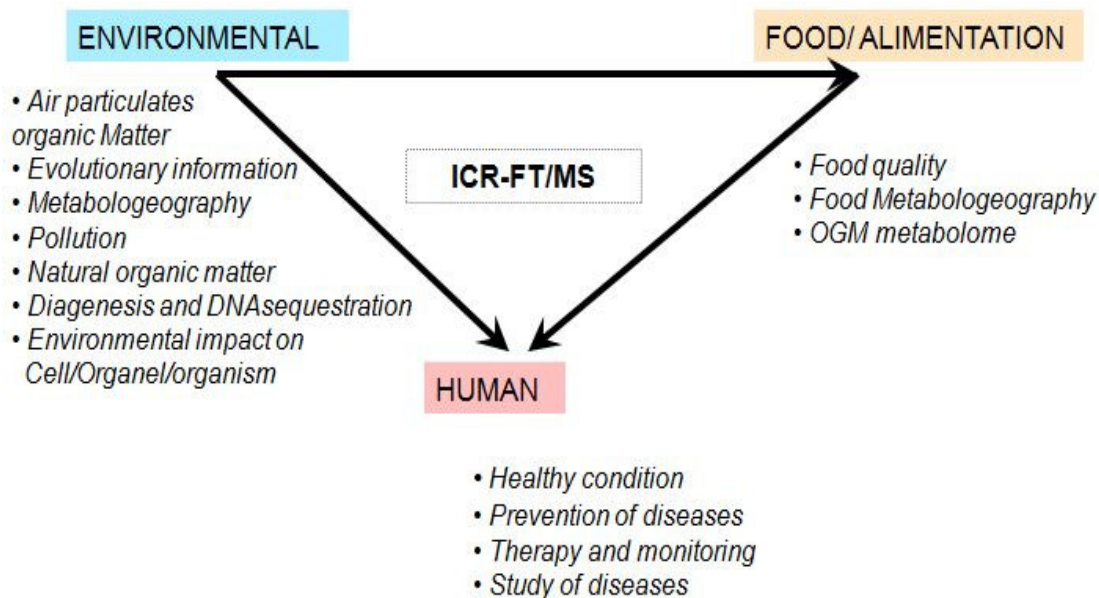


Figure 1.2: The different fields in which the Metabolomic data analysis with ICR-FT/MS is applied.

Of course all this is possible not only for a simple awareness, but also for the availability of always more sophisticated technology.

Indeed, recent advances in high and ultra-high techniques and instrumentations have fundamentally changed how metabolic processes are studied. Previously, most analytical methods were targeted to a narrow group of compounds (metabolites), usually on the basis of separation technology for a specific chemical class of compounds. However, the advance of non-targeted analytical methods solves this constraint, and now many different metabolites of different metabolic origins and chemical properties can be measured simultaneously from a single sample extract. Consequently, the amount of data generated in metabolomics studies is very huge and the integration of large multi-variant type data can be very complicated to explain. This change in how analytical approaches are conducted has eliminated one limitation and is

opening a new era, regarding the global metabolic and metabolite studies. The major bottleneck is the acquisition and the processing of complex data sets to uncover meaningful biological interpretation. Especially when the different -omics sub disciplines are integrated, as advocated in the 'systems biology' approach (Davidov, et al., 2003), it can provide an extensive, more holistic view on disease and environment.

The latest technology instrument signs the way of the modern metabolomics study. ICR-FT/MS (Ion Cyclotron Resonance Fourier Transform Mass Spectrometry) at high magnetic field is a new generation of mass spectrometer, with ultra-high resolution and mass accuracy. At present fifteen Tesla is one of the highest field strength magnets commercially available in an ICR-FT/MS. If the non target approach represents a challenge with the normal difficulties, ICR-FT/MS (12 Tesla) represents an absolute novelty (see paragraph 2.2.1).

In June 2005, former GSF, now Helmholtz Zentrum München (German Research Center for Environmental Health- Munich) has installed the first 12 Tesla ICR-FT/MS mass spectrometer in Europe (ICR-FT/MS, Fourier transform ion cyclotron resonance mass spectrometer; at present one out of eight of its magnet size class worldwide) and the instrument was operating already at the end of the year. During its first year of measurements, the instrument was equipped with APOLLO I electrospray (ESI), which is characterized by low sensitivity. Due to the installation of an APOLLO II electrospray (ESI) device, the sensitivity power could be increased. A further step was done when adopting the 4 Mega Words (MW, see chapter 2.1 for details) processing size quadrupling resolution to the current configuration defined as ultra-high resolution. No software was available for handling large datasets and the goal was to create and find new approaches for the FT data in metabolomic field. Figure 1.3 represents the schema of the thesis structure, where in the last chapters are practical examples in which the technology and the analysis are applied. The ability was to standardize the process of handling data and to apply this to several issues like:

- environmental-*biogeography* (the study was done on the extremophilic bacterium *Salinibacter rubber*)
- food chemistry (the study is focused on the high complexity of the wine)
- biomedical and diseases diagnosis (pulmonary disease and the study of pre-diabetic with a high risk to develop type 2 diabetes).

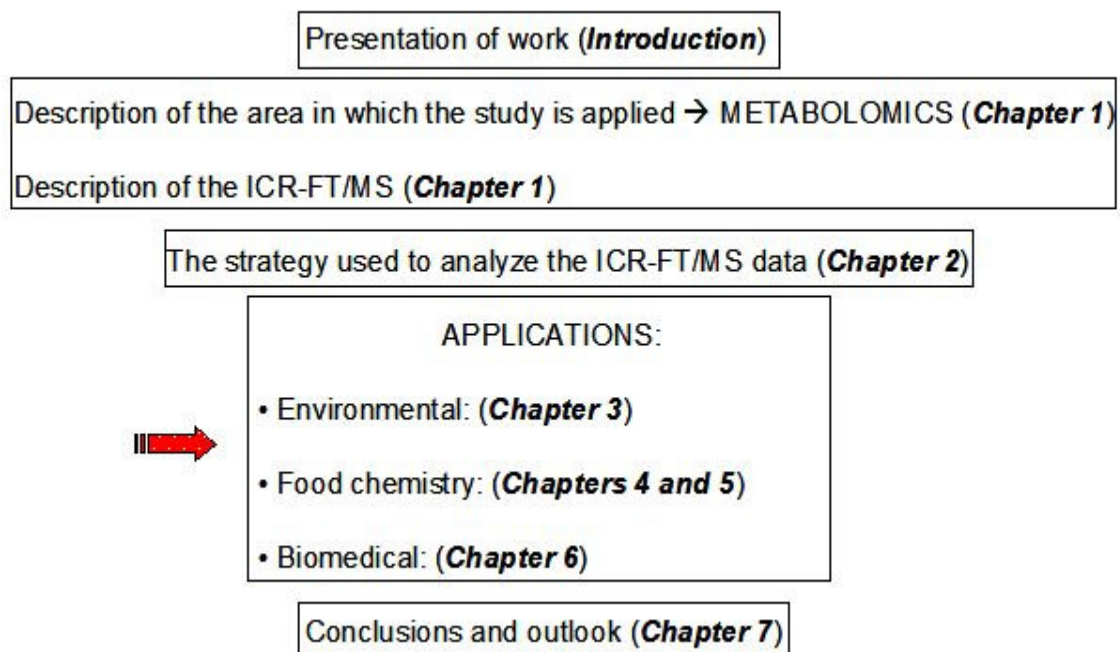


Figure 1.3: Schema of the thesis structure and the various applications of the method of analysis.

All analysis presented in this work were made with an ICR-FT/MS 12 Tesla. This thesis contains for the first time methods and tools developed exclusively for this ultra-high resolution instrument in order to take advantage of the enormous potential of this unique instrument. These represent an absolute novelty because developed in-house. In target analysis the elements to be studied are known to be detected in advance. This predetermines the methods and consequently the data analysis. In the non target analysis the numbers of metabolites detected are limited only by the instruments resolving power; thus we aim to find self consistent relation between these quantities. At the time when the ICR-FT/MS had its first light at the GSF in 2005, only one publication focusing on non-target analysis with ultra-high resolution was available. Therein (“Non-targeted metabolome analyses by use of Fourier transform ion cyclotron mass spectrometry”, (Aharoni, et al., 2002)) the authors separate

the metabolites via ultra-high mass resolution and the exact chemical composition was possible due to accurate mass determination.

Our method foresees two fundamental steps: the first one consisting in the data reduction followed by data analysis and visualization and the second one to find the putative structures and the metabolite profiles. The measurements of a sample generate an ASCII file containing the mass to charge ( $m/z$ ) and their respective intensity, in two distinct columns. These row data are first subject to spectral alignment (see chapter 2). With this algorithm the standard data configuration is traced back, suitable for statistical investigation. It is used in biology as well as in many other branches of science and technology where all the measurements have to be arranged in a data table. Since these data are highly multivariate in nature, it is necessary to use analytical techniques to cope with challenges regarding the data amount, notably noise and collinearity. Such collinearity problems can sometimes lead to serious stability problems when statistical analysis methods are applied (Weisberg, 1985), (Martens, et al., 1990).

After the usual cleaning and preprocessing (in many analysis it is an obligatory step) of the data, it was necessary to design a strategy to describe the hidden information and present them in an intuitive visualization. Several statistical techniques and visualizations procedure were studied and applied to the data, integrated with software which assigns automatically the formulae.

At the end, the identification of certain metabolite in their given biological context was strategic. The use of different databases made it possible to obtain detailed information about small molecule metabolites.

The results derived from this study will be presented here with different examples and articles.





# Chapter 2

## 2 METABOLOMICS

### 2.1 Metabolomics overview

Several terms have been derived from metabolite. A decade ago it was used for the first time the word metabolome, in order to refer to all low molecular mass compounds produced and modified by a living organism (Oliver, 1998), (Nicholson, et al., 1999). Metabolomics means “understanding biochemical mechanism, identifying biomarkers, quantitatively analysing concentration and fluxes, probing molecular dynamics and interactions”. A summary of all possible terms inherent to metabolomic field are given in Table 2.1. Also the possible types of investigation are reported. These investigations are imposed by the purposes and the goals and/or by the type of available instrumentation present in the laboratory (see paragraph 2.2).

<b>METABOLITE</b>		
Small molecules (low-molecular-weight (<~1500 Da)) that participate in general metabolic reactions and that are required for the maintenance, growth and normal function of a cell.		
<b>METABOLOME</b>	<b>METABONOME</b>	
The total sums of metabolites of a given biological system under particular physiological conditions. The metabolome is divided into xometabolome (metabolites outside the cell) and endometabolome (intracellular metabolites) (Villas-Bôas, et al., 2007a)	<i>“The sums, products and interactions of all the individual compartments/metabolomes (including extra-genomic sources) dispersed in a complex organism; the ‘Global’ System”</i> (J. Nicholson; Imperial College-London*).	
<b>METABOLOMICS</b>	<b>METABONOMICS</b>	
Identification and quantification of all metabolites in a specified cellular, biofluid or tissue section.	The quantitative measurement of the time related multi-parametric metabolic response of living system to pathophysiological stimuli or genetic modification (Nicholson, et al., 1999). It evaluates tissue and biological fluids for changes in endogenous metabolite levels effects of a disease or a therapeutic treatment.	
<b>METABOLIC PROFILING</b>	<b>METABOLIC FINGERPRINTING</b>	<b>METABOLIC FOOTPRINTING</b>
Quantitative analysis of set of metabolites or derivative products (identify or unknown) in a selected biochemical pathway or a specific class of compounds. This includes target analysis, the analysis of a very limited number of metabolites, e.g. single analytes as precursors or products of specific biochemical reactions.	Unbiased, global screening approach to classify samples based on metabolite patterns or “fingerprints” that change in response to disease, environmental or genetic perturbations with the ultimate goal to identify discriminating metabolites. Quantification and metabolic identification are generally not involved.	Called also exometabolome, it is the observation of what a cell or system excretes under controlled conditions (Kell, et al., 2005).

Table 2.1: Glossary related to metabolomics definitions and to the different approaches applied in this field. (\*) Jeremy Nicholson was

among the first to apply the tool of metabolomics analysis to NMR (and now to MS) to the assessment of metabolite changes in biofluids over time. He is who has coined in 1996 (together with his colleagues) the word “metabonomics”.

Nowadays metabolomics analysis are of interest in a variety of areas such as human and animal nutrition (Whitefield, et al., 2004), (Gibney, et al., 2005), (Rist, et al., 2006), cancer diagnosis and therapy (Hartmann, et al., 2006), (Malhi, et al., 2006), biomarker discovery (Goodacre, 2005), (Schlotterbeck, et al., 2006), toxicology (Robertson, 2005), (Gerner, et al., 2006), obesity studies (Hochberg, 2006), enzyme discovery, (Saito, et al., 2006), (Villas-Bôas, et al., 2006), drug discovery (Harrigan, 2006), transplantation (Wishart, 2005), agriculture (Bender, 2005), (Dixon, et al., 2006) and bioremediation (Singh, 2006). It has also the claim to speed up the functional analysis of genes with unknown function (Villas-Bôas, et al., 2007b); in this optic the changes in the metabolite biochemical composition could be used to correlate the mutation of a small number of metabolites or establish the part of metabolism affected by comparison with profiles of mutants of genes of known function (Raamsdonk, et al., 2001).

### 2.1.1 One or more crucial metabolites: Biomarkers

Although the term biomarker historically refers to analytes in biological samples, any measurement that predicts an individual’s disease state or response to a drug can be called a biomarker (Baker, 2005). A biomarker is defined as “*a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic intervention*” (Atkinson, et al., 2001). It is distinguished from a clinical end point, which is defined as “a characteristic or variable that reflects how a patients feels, functions or survives” or a surrogate end point defined as a biomarker that is intended as a substitute for a clinical end point (Atkinson, et al., 2001). A surrogate end point is expected to predict clinical benefit (or harm, or lack of benefit) based on epidemiological, therapeutic, path physiological or other scientific evidence (Atkinson, et al., 2001). To expedite the clinical drug evaluation process, there is a high demand for biomarkers that adequately, and with great specificity, indicate the presence or absence of the desired pharmacological response (Lewin, et al., 2004). It has now become evident that a broader array of ‘knowledge-based’ (relating to the known mechanism of action),

combinatorial biomarkers (Koop, 2003) (or biomarker profiles) can be used for better decision-making, i.e. to stop the development of nonviable drug candidates as early as possible and transferring the available resources to potentially more successful ones (Baker, 2005), (Rolan, et al., 2003). In the long run, scientists are looking to metabolomics to fill important gaps in systems biology, a research paradigm focused on all the interconnected molecular pathways in cells and organisms. Short-term clinical goals for the field are more affected by the search for biomarkers, or molecular indicators of pathology. Individual metabolites have already been used as disease biomarkers for years, for example: elevated glucose is an indicator of diabetes mellitus and cholesterol is a metabolite long conjoined with heart problem and stroke. Metabolomics enables the identification of biomarkers based on entire groupings of metabolites that are up or down regulated in concordant under specific conditions.

## 2.2 Different technologies in metabolomics data

The different types of platforms have developed different approaches in the metabolomics study (a general schema is reported in figure 2.1).

Generally we can summarize the approaches on metabolic study into two prevalent strands: separation/mass spectrometry (sep-MS) and NMR methodologies.

The traditional technologies of measuring metabolomics data are Liquid chromatography-mass spectrometry (LC-MS), Capillary electrophoresis-mass spectrometry (CE-MS) and Gas chromatography-mass spectrometry (GC-MS). These are prevalent characteristics for low throughput. They are applied essentially in the target analysis for quantitative metabolite profiling. Target analysis is restricted to the substrate and/or the direct product of a specific metabolic step (Bhalla, et al., 2005). GC-MS technologies allow the identification and robust quantification of a few hundred metabolites within a single extract (Roessner, et al., 2001), (Halket, et al., 2003). Compared to the gas chromatography technologies LC-MS offers several distinct advantages because it is adapted to a wider array of molecules including a range of second metabolite (alkaloids, flavonoids, isoprenes, glucosinolates, oxylipins, phenylpropanoids, pigments and saponins (Aharoni, et al., 2003), (Matuszewski, et al., 2003)).

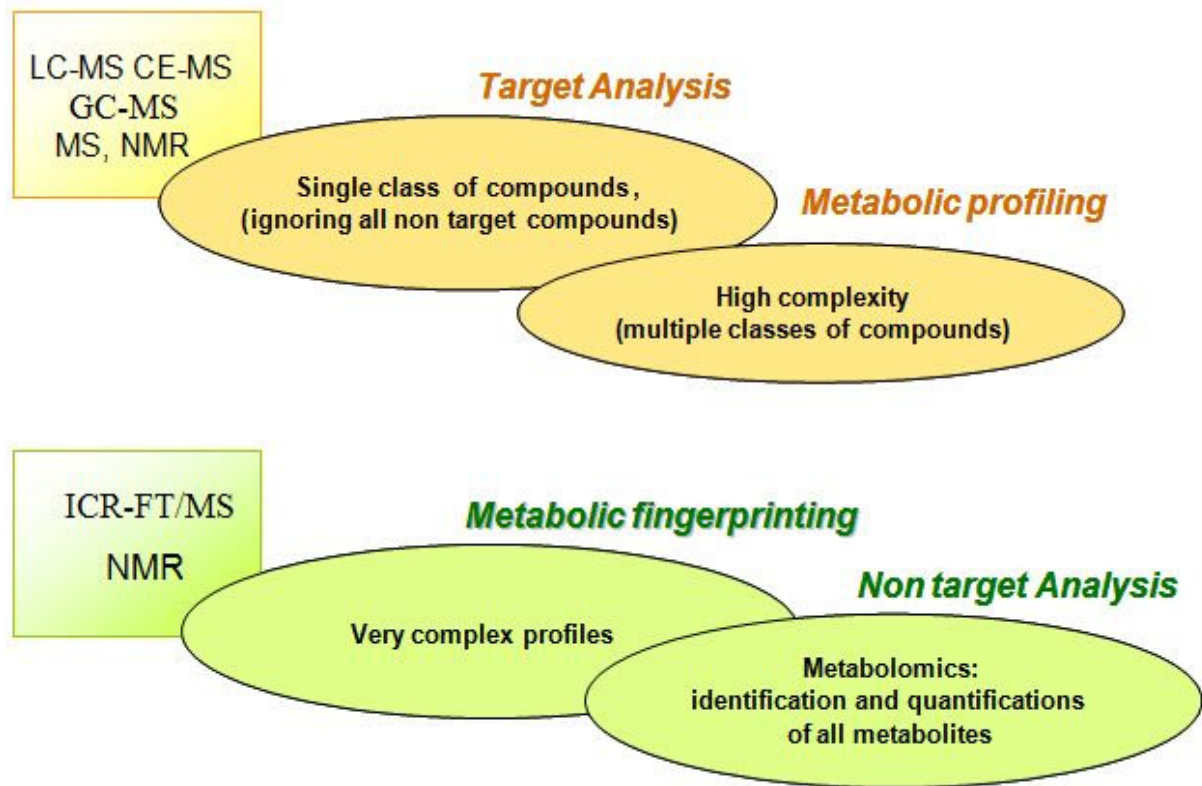


Figure 2.1: The figure shows the different types of instrumentations associated with their possible type of analysis. Usually GC-MS, LC-MS, CE-MS and NMR are applied in the target analysis and in the metabolic profiling.

A second approach is the metabolic profiling, where analysis is restricted to the identification and the quantification of a selected number of pre-defined metabolites in a biological sample (Bhalla, et al., 2005). Here the targeted components are more in number and focused on classes of metabolites (i.e. lipids, sugars, peptides, proteins, etc), the information is of a structural basis with further possible target quantification (Schmitt Kopplin, et al., 2007).

	NMR		Mass Spec
<b>LOGISTIC</b>			
Capital cost		No advantage	
Routine operating costs		No advantage	
Maintenance	Advantage		
Per sample cost	No advantage	No advantage	
Footprint		No advantage	
Required technical skills <sup>a</sup>			Advantage
Instrument “up-time”	Advantage		
Instrument life-span		No advantage	
<b>ANALYTICAL CONSIDERATION</b>			
Sensitivity			Big Advantage
Reproducibility (within lab)	Advantage		
Reproducibility (across labs)	Big Advantage		
Quantitation	Big Advantage		
Average run speed		No advantage	
Capacity (samples/day)		No advantage	
Sample preparation requirements	Advantage		
Sample analysis automation	Advantage		
Versatility <sup>b</sup>	Advantage		
Selectivity <sup>c</sup>			Advantage
Nonselectivity <sup>c</sup>	Advantage		
<b>METABONOMICS</b>			
Resolvable metabolites			Big Advantage
Identification of unknowns			Advantage
Potential for sample bias <sup>d</sup>	Big Advantage		
Data analysis automation			Advantage

Table 2.2: Comparison of NMR and mass spectrometry for metabolomics study, adapted from (Robertson, 2005). <sup>(a)</sup> Pool of qualified analysts is much smaller for NMR than MS. <sup>(b)</sup> Generally any NMR instrument can be configured for most applications while different MS instrumentation may be required for specific applications. <sup>(c)</sup> MS excels at selective identification of a molecular entity, while NMR excels at identification of all protons containing species in a sample. Therefore, selectivity can be an advantage or disadvantage depending on the nature of the application. <sup>(d)</sup> Potential for misleading, incomplete or no reproducible data set due to bias inherent to the technology (e.g., ion suppression in MS).

Nuclear magnetic resonance (NMR) is a consolidate technique in the metabolomic field for global metabolic fingerprinting and biomarker

identification. Being non invasive and associated with high developed software it is a fundamental technique in the metabolite profiling, despite to its lower sensitivity and resolution compared to mass spectrometry methodologies. The study was initiated by Jeremy Nicholson (Imperial College, London) and with the publication: “‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data (Nicholson, et al., 1999), he laid the groundwork for the metabonomic approach.

A new up-coming strategy for a non target approach is Ion Cyclotron Resonance Fourier Transform Mass Spectrometer (ICR-FT/MS), a rapidly emerging alternative to other types of mass spectrometers capable of non-target metabolic analysis and suitable for rapid screening of similarities and dissimilarities in large collections of biological samples (Aharoni, et al., 2002).

For these two branches a comparison between advantages and disadvantages is reported in table 2.2.

### 2.2.1 The novelty in the non-target analysis with ICR-FT/MS

The concept and application of non-targeted analysis of cellular metabolites in a system-wide hypothesis-driven approach has transformed the methodological strategies in different areas of life sciences, with ever powerful vigorous. The instruments used for this approach are NMR and ICR-FT/MS (see figure 2.2); the attention will be focused in particular on the later one, that allows identification and quantification of metabolites based on their accurate mass determination.

The technique of ICR-FT/MS was first presented in the 1950's (Von Hippel, et al., 1949) where it was demonstrated using measurements of very small mass differences with very high precision. The technique remained a largely academic tool until the application of FT methods (Cooley, et al., 1965) by Alan Marshall and Melvin Comisarow in the early 1970's (Comisarow, et al., 1974). Since that the instrumental evolution continues without stopping.

The basic FT-MS foresees the ions to be produced in the source itself. By entering the cell ions are in an environment having low pressure values around to 10-11 mBar.

This is obtained by cooling processes using liquid helium and liquid nitrogen. The cell itself is embedded in a spatial uniform magnetic field. Injecting the ions in a magnetic field and according to the magnetodynamics, the charges are subject

to a force, referred to as Lorentz Force, which is perpendicular to the magnetic field vector and injection velocity vector. This is given by the formula:

$$\vec{F} = z \cdot \vec{v} \times \vec{B}$$

where  $\vec{F}$  is Lorentz Force,  $\vec{v}$  and  $z$  are respectively the velocity and charge of the ion and  $\vec{B}$  the magnetic field. The force to which the ion is subject can also be described by the simple Newtonian expression:

$$\vec{F} = m \cdot \vec{a}$$

where  $m$  is the mass of the ion and  $\vec{a}$  its acceleration.

Equalizing these two expressions it is possible to derive the angular frequency of the orbiting ion, which is given by:

$$\omega_c = \frac{z \cdot B}{2 \cdot \pi \cdot m}$$

Each  $m/z$  unit must be excited because their orbital radius is too small to be detected. Each single ion  $m/z$  packet having natural frequency  $\omega_c$  will couple with the excitation frequency reaching the resonance. Then it drops back to the ground state (natural frequency). Being at higher orbits the  $m/z$  packets induce an altering current between two detector plates. The current frequency is the same of the cyclotron frequency and its intensity is proportional to the number of ions. This results in a complex frequency vs. time spectrum produced by all the ions and containing all the signals - the FID. Deconvolution of this signal by FT methods results in the deconvoluted frequency vs. intensity spectrum which is then converted to the mass vs. intensity spectrum (the mass spectrum by the previous equation).

The ICR-FT/MS performance parameters improved in proportion to the strength of the magnet, "B" or "B<sup>2</sup>" (Marshall, et al., 1996). According to this an instrument equipped with 12 Tesla the resolving power increases linearly with increasing



magnetic field strength (Marshall, et al., 1996). Basically, ICR frequency is proportional to  $B$  ( $B$  is the magnetic field strength which is constant). As  $B$  increases, all of the ICR frequencies also magnify, as does the difference between any two ICR frequencies (see figure 2.2).

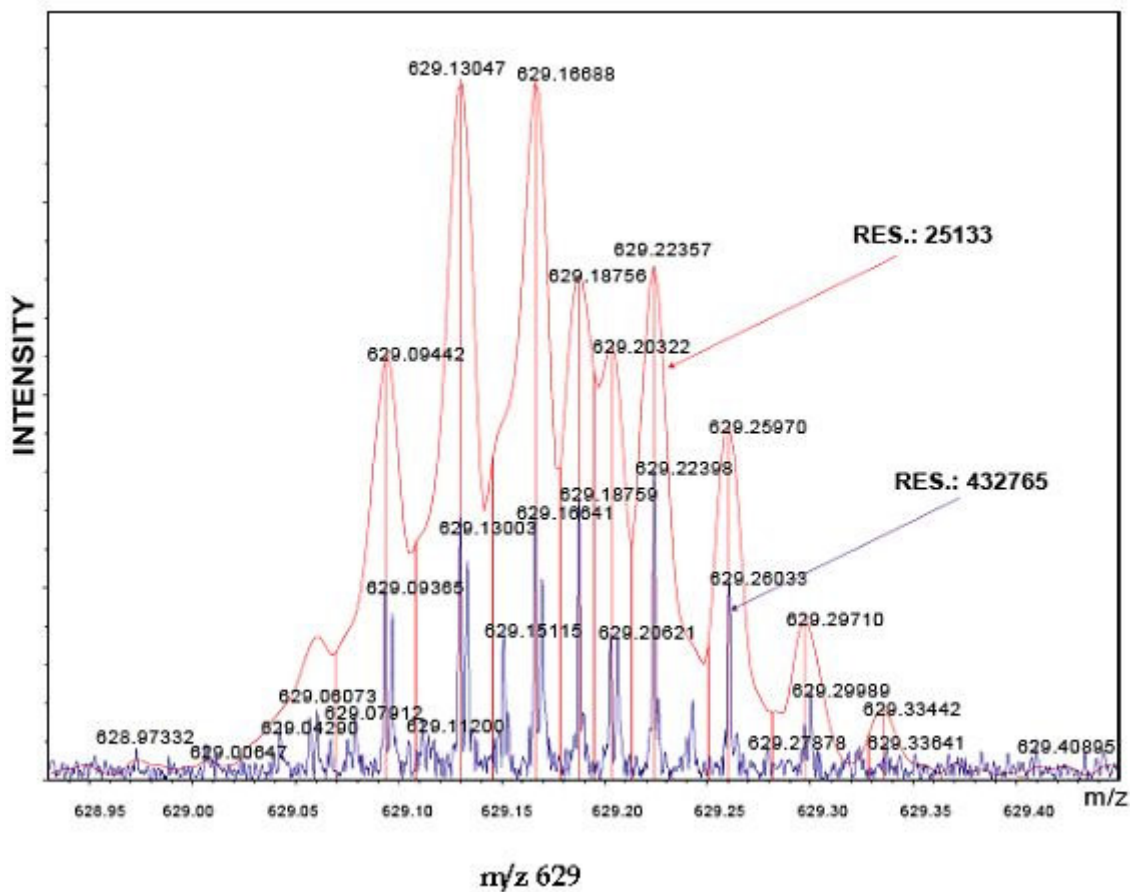


Figure 2.2: The ultra-high resolution at mass 629 for 7 Tesla compared with 12 Tesla is shown.

The major advantage of ICR-FT/MS is that it enables the assignment of thousands of elemental compositions of metabolites in a mass range from 120 to

800 kDa (Kilodalton) directly out of complex mixtures by virtue of ultra-high mass accuracy (<100 parts per billion, ppb) and ultra-high resolution (500,000 at mass 500) at high-field strength. It offers experimental mass accuracies of 0.1 part per million (ppm), which is nearly an order of magnitude better than the most advanced time-of-flight-based mass spectrometers currently available.

The principal characteristics of the ICR-FT/MS used in metabolomics are:

- Ultra high resolution (Peak Capacity, figure 2.3)
- High mass accuracy (elementary composition, figure 2.4)
- Semi-quantitative approach (relative differentiation, figure 2.5)

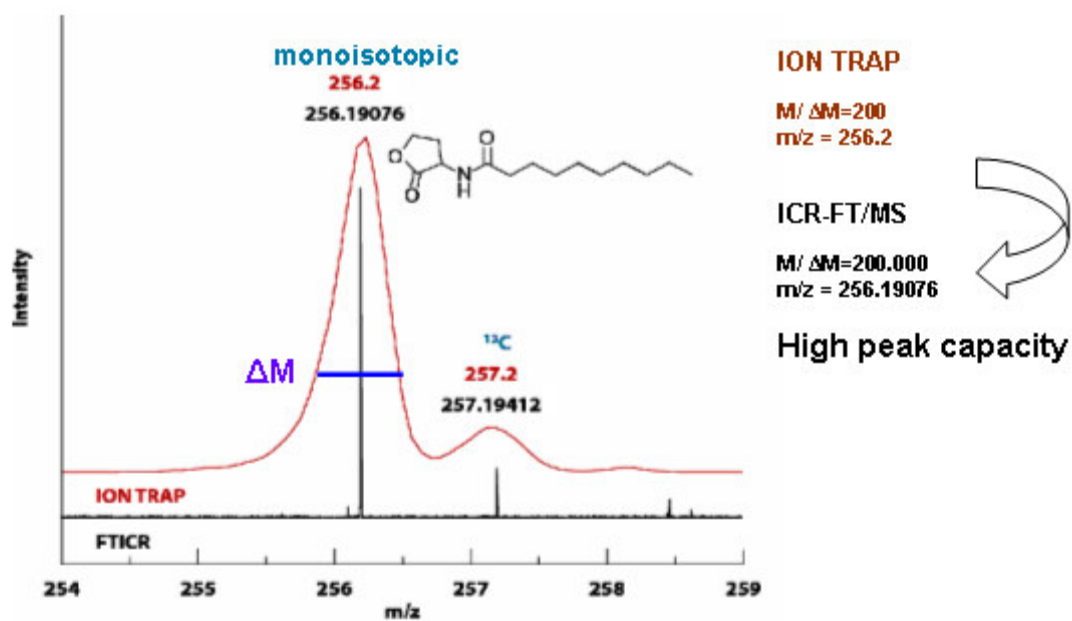


Figure 2.3: Visualization of high peak capacity of ICR-FT/MS related to ION TRAP.

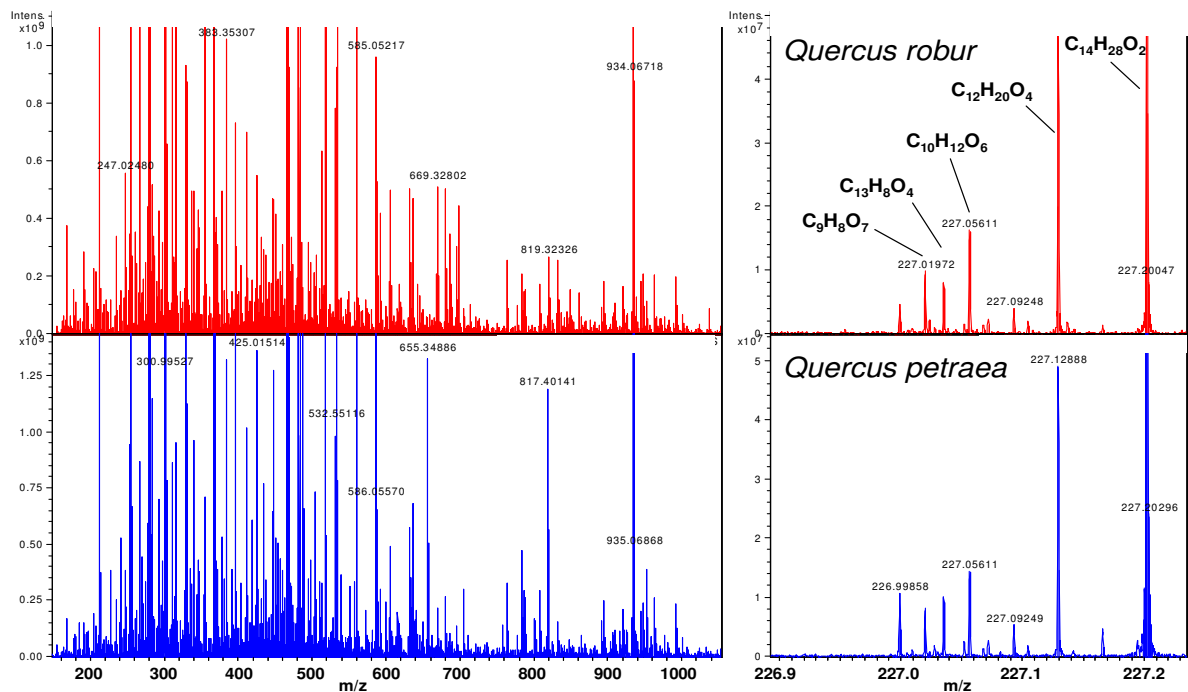


Figure 2.4: In order to point out the high resolution we show in this figure two examples; the typical negative ion mass spectrum of the sessile and pedonculate wood extract samples (extracted averaged wood sample of each species); detail on mass 226.90-227.20 with elementary composition assignment of the major intensities.

Due to ultra-high resolution and mass accuracy for one nominal mass it is possible to assign dozen of molecular formulae of different constituents derived from complex mixture.

By the use of ICR-FT/MS it is possible to achieve separations (with the ultra-high resolution) and identification of the metabolites (the accurate mass determination permits the determination of the elemental composition), and the relative quantification (achieved by comparing the absolute intensities of each masses using the internal calibration).

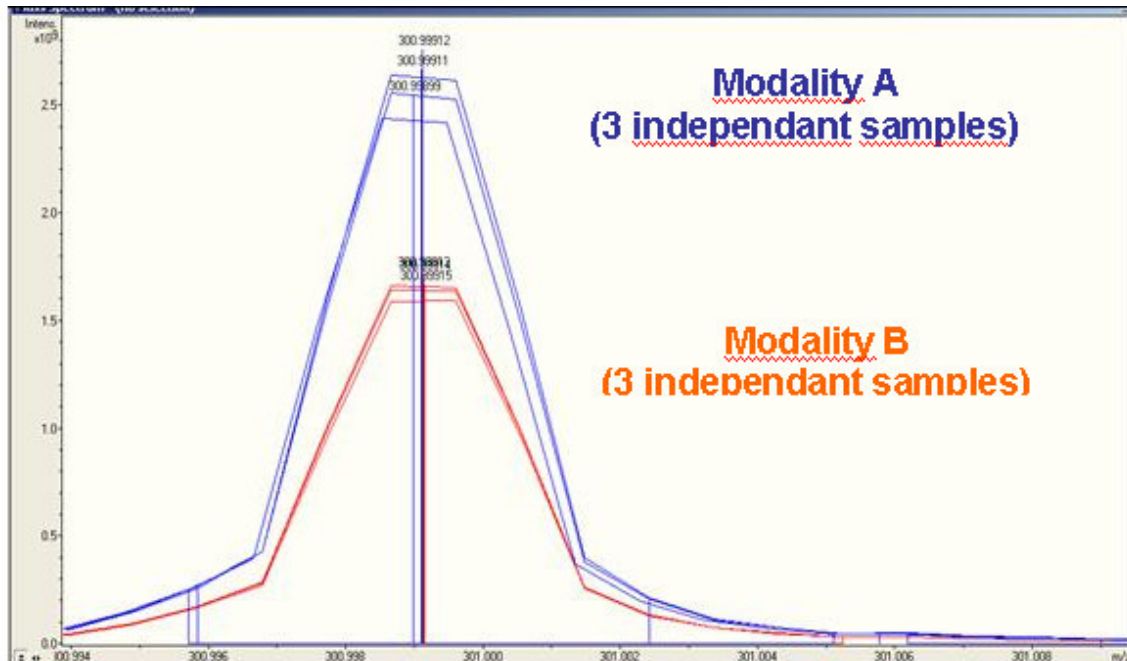


Figure 2.5: Visualization of semi quantitative approaches (relative differentiation). This figure shows the reproducibility of a measurement. Three independent samples are injected in different concentration, but the measurement is perfectly reproduced.

The most problematic things remain the lack of chromatography, which make the technique unable to differentiate between isomers, due to their identical molecular mass. This could be handled by further fragmentations, which give greater accuracy of identification.

Accurate mass measurement capability of ICR-FT/MS has been proven as a significant tool for improving confidence level in metabolomics identification in bottom-up approach. However, all of the advantages of ICR-FT/MS could not be fully demonstrated because there is a gap between the complexity of ICR-FT/MS spectra and capability of interpretation of the information by software. So it is still necessary to develop integrated software applicable for interpretation of mass spectra of metabolomics obtained from a high-end ICR-FT/MS.

## 2.3 Chemometrics

Chemometrics is the application of mathematical and statistical analysis to the chemistry field. It is closely linked with MS and NMR, and it has as output the productions of tools, in order to process the spectra. It starts from the traditional correcting baseline effects, smoothing, peak alignment, outlier detection, normalization (Deming, 1986), (Lavine, et al., 2004), (each of these steps must be evaluated if necessary to apply or not, in relation with the type of instrument available). After these preliminary steps normally it proceeds with the search for patterns, to track properties of the samples analyzed and to prepare and use multivariate classification models. At this point the Chemometrics finishes and the Bioinformatics discipline begins. The border between these two sciences is not well defined. Anyway the common goals for both are to find structures in experimental information and to describe it in an interpretable and easy way: *“The [metabolic] profile will give you knowledge and information rather than just data”* (Bruce Hammock University of California, Davis). Technology and the modern metabolomics are growing very fast, for this fact MS and NMR suffer from well documented technical limitations (Weckwert, 2003). Especially defining the molecular composition of complex mixtures is one of the most difficult tasks in metabolomics (Markley, et al., 2007). Imperative are the recommendations of the Metabolomics Standards Initiative (MSI), it suggests that metabolomics study should report all details of: the experimental design, metadata, experimental procedures, analytical, data processing and statistical analysis which are applied (Lindon, et al., 2005).

### 2.3.1 Exploratory data analysis

Patterns of association exist in many data sets, but the relationships between samples can be serious and abstruse to discover when in the data matrices are stored complex data with thousands of variables from the biology system. Exploratory data analysis can reveal and model hidden patterns in complex data by reducing the information to a more comprehensible form.

Such a Chemometrics analysis for instance can show the reasons why a variable in the model is an outlier and indicate whether there are patterns or trends in the data, underlying qualitative features (latent structures) from the multivariate spectral data. Multivariate analysis such as principal component analysis (PCA) and

hierarchical cluster analysis (HCA) are used to reduce large complex data sets into a series of optimized and interpretable objects. These views emphasize the natural groupings in the data and show which variables are strongly related to those patterns.

### 2.3.2 Classification modeling

Many applications require the samples to be assigned to predefined categories (classes), a priori information. This may involve determining whether a sample has the same chemical property of a specific group or not, or predicting an unknown sample as belonging to one of several distinct groups. A classification model is used to predict a sample's class by comparing the sample to a previously analyzed experience set, in which the classes are already studied, based on the multivariate similarity of one sample to others.

Two different classification models that are commonly used are K-Nearest Neighbor (KNN) and Soft Independent Modeling of Class Analogy (SIMCA). When these techniques are used to create a classification model, the answers provided are more reliable and include the ability to reveal unusual samples in the data. Moreover with partial least squares (PLS) technique it is possible to measure the degree of predictability. With regression models, the analyst is interested in predicting some value (rather than assigning a class designation) for an unknown sample.

# Chapter 3

## 3 STRATEGY FOR LARGE DATASET

### 3.1 Instruments for the analysis

Within this chapter we outline the general approach we applied to different types of datasets. Several exemplifications and examples are shown too.

The data analysis of ICR-FT/MS requires the development of sophisticated tools and statistical methods in order to extract useful information. For some aspects the techniques used for the data analysis are similar to NMR, gas or liquid chromatography data processing. Indeed metabolomics studies often require multivariate pattern recognition techniques to extract meaningful results. Before any statistical analysis it is necessary to prepare the dataset. To this end we have developed a data processing pipeline using different programming languages.

All measurements were performed with the Bruker Daltonics APEX Qe Fourier transform mass spectrometer (ICR-FT/MS) equipped with a 12 Tesla superconducting magnet and an Apollo II electro spray source. The spectra were zero filled to a processing size of 4 Megaword (MW).

After the instrument commissioning in 2005 the available processing size was of 1 MW which limited the mass resolution. Also the data processing was specifically designed for this computer storage capacity. With the advent of 4 MW processing size (in 2007) the data analysis needed to be completely re-designed.

The raw data are processed with DataAnalysis 3.4 (Bruker Daltonik) software that is hard-coded in the instrument. After our pipeline aligns the spectra and chemical formulae are computed by software developed in-house in Python and FORTRAN 95. Statistical analyses are done with SIMCA-P 11.5 (Umetrics, Umea, Sweden) and SAS version 9.1 (SAS Institute Inc., Cary, NC, USA). Different types of database are used to find metabolite annotation from exact mass. The most common are present in open sources databases accessible from the web (KEGG: Kyoto Encyclopedia of Genes and Genomes, KNApSack: A Comprehensive Species-Metabolite Relationship Database, METLIN: A Metabolite Mass Spectral Database). Based on collections of aerosol publications and from wine samples, we have developed our own database managed by a database management system. This was essential especially at the beginning when we wanted to annotate the type and the formula to a list of exact masses. The pathways visualization is performed with a tool available on the web only at the end of 2007. This is the reason because many elaborations are without this tool. For Van Krevelen and Kendrick mass visualization has been prepared an excel sheet.

The following paragraphs will examine in detail all processes of acquisition, processing and elaboration of the data, which are represented in a schematic way in figure 3.1.



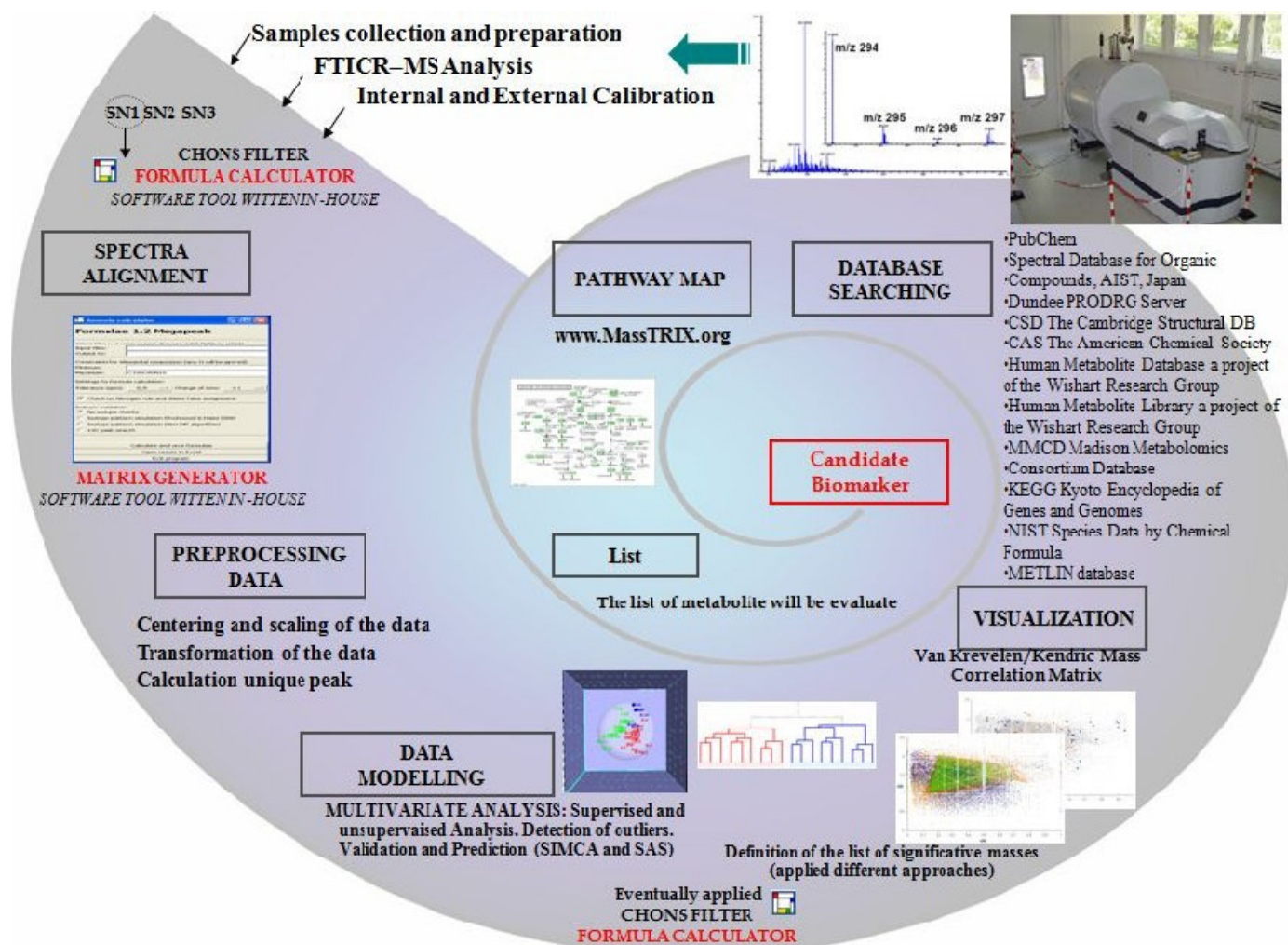


Figure 3.1: General scheme for ICR-FT/MS data processing. After the sample collection and preparation, the measurement and the calibration the process of investigation starts. It is necessary to align the spectra and store them in an ordered matrix. The spectra are exported at different S/N in order to stabilize the process and not to exclude possible information. The preprocessing phase (data centering, scaling, etc.) is followed by building up of statistical models and their validation. Once a list of possible biomarkers is drawn the chemical properties of these are investigated through graphical solutions (i.e. Van Krevelen Diagram) and/or cross correlation on existing data bases or with MassTRIX in order to submit to find a pathway maps.

## 3.2 Organizing the data

Being no tools and literature available, our data processing pipeline was used for the first time ever. For this reason all the procedures needed to be developed from the beginning, and sometimes the experience suggests possible choices.

The alignment starts with the mass spectra output of the ICR-FT/MS, processed by Data Analysis Bruker Software. The mass spectra are exported in ASCII format and they are extracted at signal to noise=1 (S/N), S/N=2 and S/N=3. Each ASCII file represents a mass spectrum having on x-axis the mass-to-charge ratio and on y-axis the measured intensities. Both of the observationally determined quantities are subject to statistical investigations.

The spectra extracted by ICR-FT/MS are analyzed as samples in the context of the population which they belong to. This requires a preliminary data structure design. For this purpose was developed the “Matrix Generator”, an in-house software coded in Python. The union of sample files, operated by Matrix Generator, is foreseen following the logical scheme visualized in the integration flux represented in figure 3.2.

A first step requires sorting the joined data according to the mass value. To this end the data belonging to the same sample are tagged. This way they are recognizable over the whole data processing pipeline. Once spectra data are tagged, joined, sorted a second step groups the different masses. According to the mass error, moving within the range 0.1 - 10 ppm (part per million), a cross-correlation between the observed mass positions and error template is performed leading to averaged masses in case of positive correlation and their abundances are reported. The next step restarts the comparison with the previous mass and it verifies if it is within the defined range.

The so obtained masses are used in spectra assigning the respective measured intensities or the value 0 (zero) if the element associated to that mass is not present. As a result we obtain a data set composed by all the masses (measured and calculated) and their respective intensities (if present, otherwise 0), then the peaks are aligned between samples so that one mass in a sample corresponds to the same compound in the next sample if present. This calculation is admittedly not justified by any analytical consideration but only serves as a first step to simplify the input lists.

In this way The Matrix Generator aims consistent comparison among different spectra and it avoids wrong comparison results between different spectra.

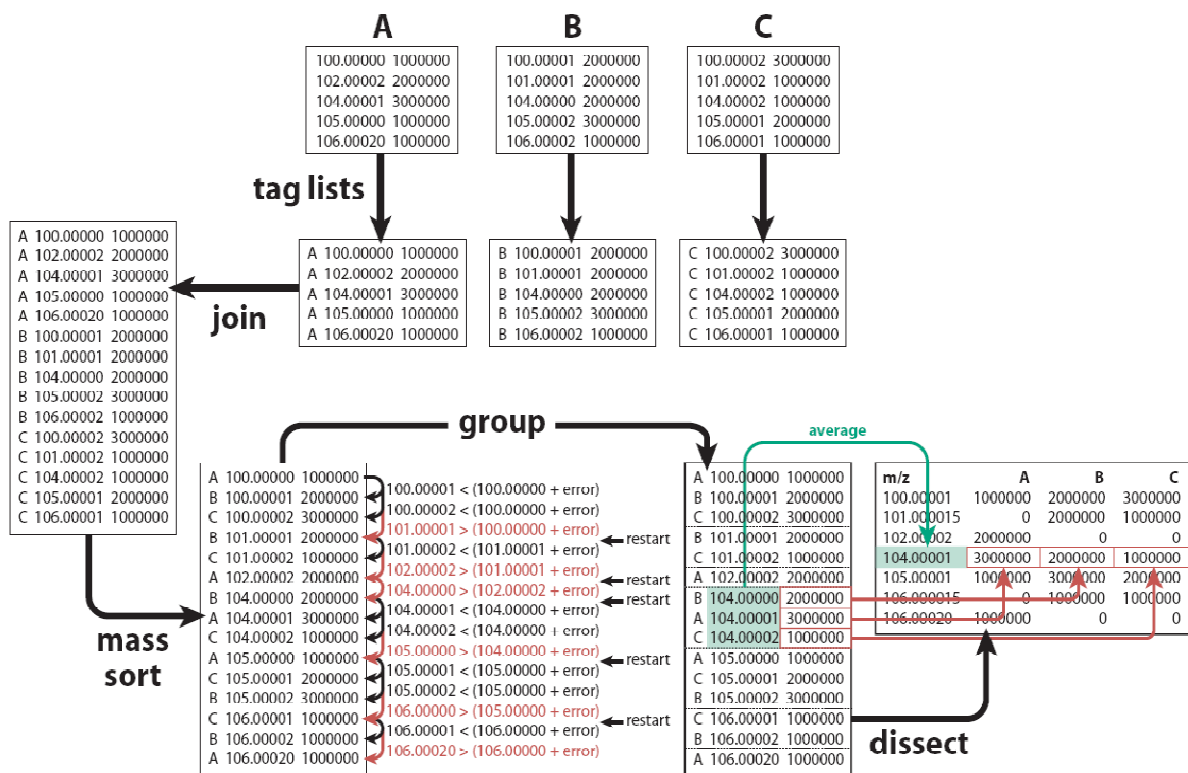


Figure 3.2: Schematic view of the spectra alignment algorithm. A, B and C represent the mass spectra. The window widths, based on it the masses will be averaged, are set in advance and it ranges from 0.1 ppm to 10 ppm.

The value of the bin (measured in ppm and chosen for the spectra alignment) must be defined in order to avoid false assignments. To define the value were created different matrices at different values of ppm (from 0.1 to 10 ppm) and counted the total masses generated. This process was also done at different value

of MW (1 and 4 MW). The ideal value came from after the inspection of the graph, an example is given by the figure 3.3a in which we tested the value for *Pseudomonas Putida* samples. This value was always chosen for different type of measurement and it brings good results, it limits false assignment and false unifications. We have observed the phenomena also for the samples measured at 1 MW. It was also shown empirically that it no produces false assignment, distinguishing different masses and put together the right “double peak events”.

The figure 3.3 shows the plot of the  $m/z$ -values as function of ppm-values in log-log scale. The data are fitted with a broken power-law model described by the equation here below:

$$F(\text{ppm}) = A \cdot \text{ppm}^{-\alpha} \quad \text{for } \text{ppm} \leq \text{ppm}_b$$

$$F(\text{ppm}) = A \cdot \text{ppm}_b^{\beta-\alpha} \cdot \text{ppm}^{-\beta} \quad \text{for } \text{ppm} > \text{ppm}_b$$

where  $A$  is normalization constant,  $\text{ppm}_b$  is the beak value and  $\alpha, \beta$  are the spectral indices respectively before and after the beak value. The two different indices show the sampled data belonging to two populations different in origin. An interesting deepening is taken from literature (Savaglio, et al., 2000). The value chosen for the ppm-window is 1 ppm at 4 MW. The best fit resulted in two spectral indices respectively of values  $\alpha = 0.48$  below and  $\beta = 0.26$  before and after the  $\text{ppm}_b$  fitted value of 1.3. The first leg of the curve (left of 1 ppm) could be connected to the type of calibration while the second is more tied to a wrong merger, but more investigations are needed for these deductions.

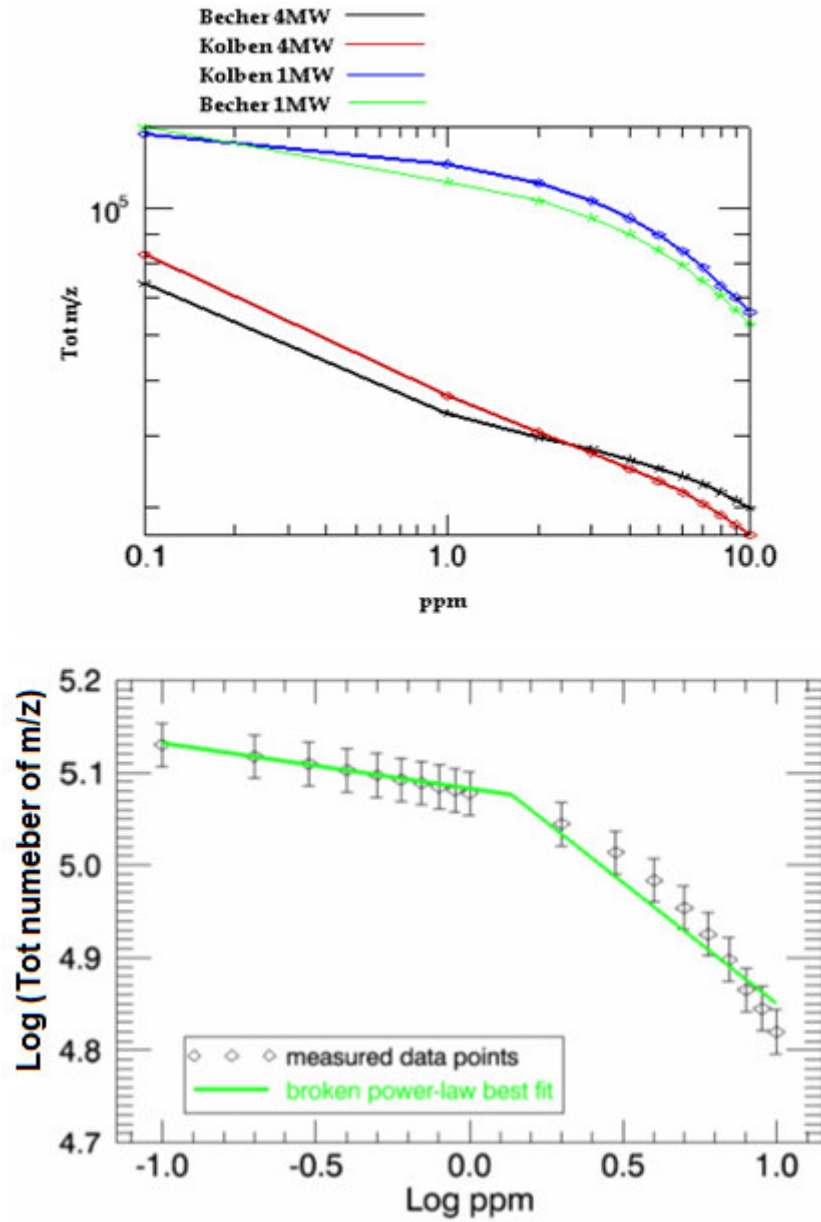


Figure 3.3: The upper panel shows the amount of masses derived at different levels of ppm (from 0.1 to 10) in comparison with 1 MW and 4 MW processing sizes. The data came from an experiment aiming to follow

the metabolome of *Pseudomonas Putida* during growth conditions. The scale in both axes is logarithmic. For the bin, in Matrix Generator was chosen the value 1 ppm. The lower panel of the figure shows the Becher 4 MW model fitted with a broken law power. The broken power law fit is shown in green, which breaks at 1.3. The value 1 ppm was defined as good value.

### 3.2.1 Formula calculation

The formula calculation is used in order to assign the chemical composition of the masses and to filter those are not assigned.

The 4 MW processing size required a hardware upgrade on the ICR-FT/MS. Since software and memory capacity are related all together also the processing data logic was updated in order to achieved the maximum information at 4 MW processing size. The Formula Calculation was applied as a filter with the 1 MW processing sizes before the advent of the 4 MW one. To validate all processes the spectra were computed according to two different significances:  $S/N=3$  and  $S/N=1$ .

Two different statistical models and processing methods were evaluated to achieve a robust result. In the case of  $S/N=1$  the formula calculation was applied before spectra alignments aiming to filter wrong assignments. These spectra were aligned; pattern defined and finally chemical properties and biomarkers derived. Simultaneously the spectra computed at significance level  $S/N=3$  were aligned with the software Matrix Generator. In this case formula calculation was applied at the end of the processing pipeline in order to filter the masses of interest. The two models were compared to stabilize the result and to validate the possible list of biomarker (a general schema is proposed in figure 3.4).

For elemental formula calculation, as a prerequisite, all mass spectra need to be calibrated to maximum accuracy, either internally or externally. In case of the samples presented here, external calibration was done with an appropriately concentrated arginine solution in positive and negative mode, while internal calibration was accomplished utilizing ubiquitous solvent impurities, either phthalate diesters (for positive ESI spectra) or fatty acids (for negative ESI spectra). For both external and internal calibration (where the further served as a positive control of proper system configuration), a maximum mass error of 0.1 ppm was accepted.

From adequately calibrated spectra, peak lists were generated which contained up to 10,000 mass/intensity pairs. A portable program written in

FORTRAN 95 with a modular graphical user interface and processing front-end written in Python/TkInter was used for rapid batch calculation of possible elemental formulae.

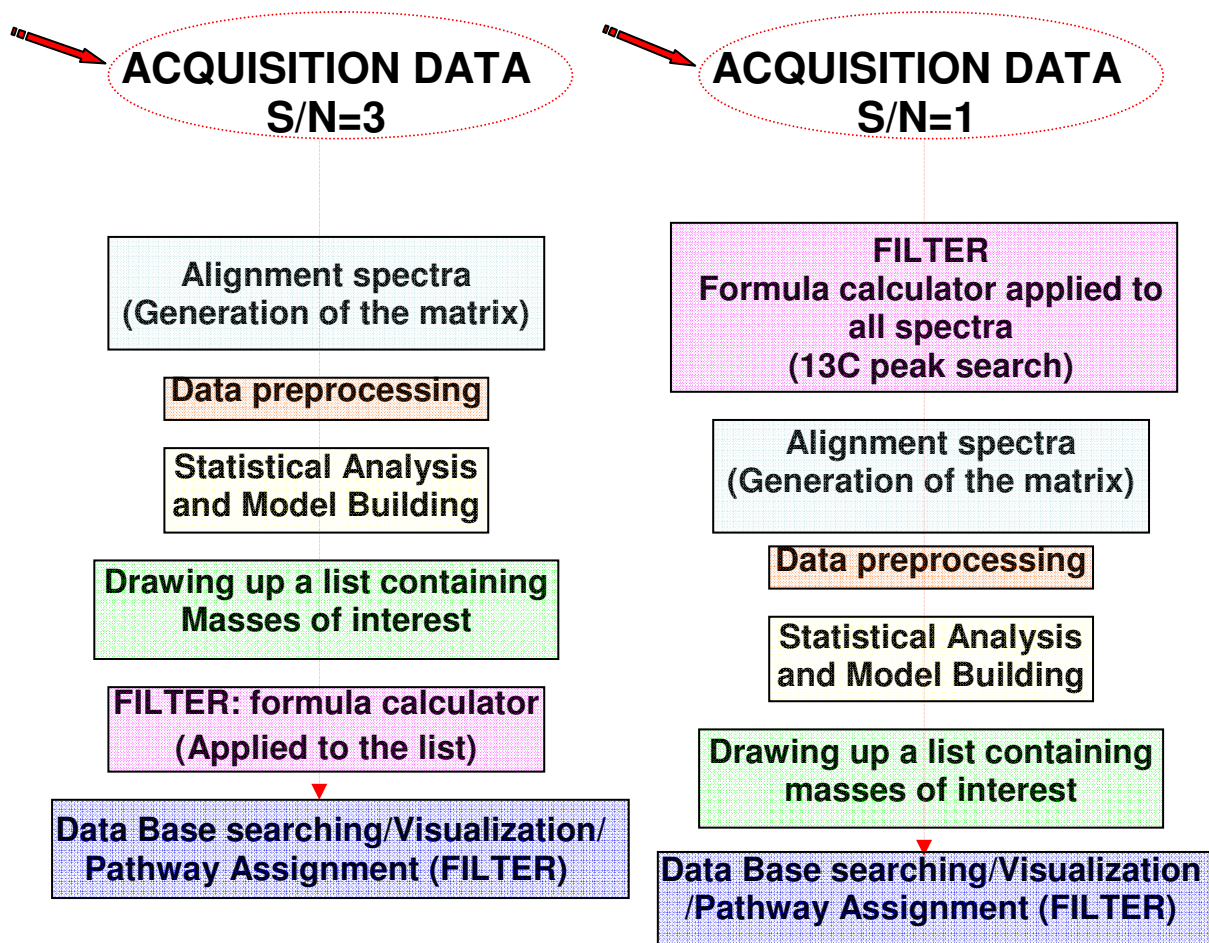


Figure 3.4: General schema to show in which level the filter is applied. In the last step Van Krevelen and/or Data Base searching are in turn equivalent to a filter.

In principle, formula calculation was performed by subtracting (for positively charged peaks) or adding (for negatively charged peaks) an electron mass (0.000549 atomic mass unit (amu)) from or to the measured  $m/z$  value. This formal neutralization was followed by a transformation of the mass value from a  $^{12}\text{C}$  based IUPAC scale to a  $^1\text{H}$  based hydrogen scale, where  $\text{H} = 1.000000$  amu,  $\text{C} = 12.000000/1.007825$  amu,  $\text{O} = 15.994915/1.007825$  amu, etc. The purpose of this transformation was to reduce the number of necessary calculation steps by a reduction of the number of possible elemental combinations that needed to be considered for each mass. In this transformed mass, only the heavier elements contribute to the fractional part of the number, but not H, which normally is the most abundant element in biomolecules. In parallel with this transformation, a database consisting of all possible combinations of elements (excluding H but analogously H-transformed) that were previously allowed, that fit in the mass window of interest, and that match some very basic chemical rules was generated (nitrogen rule, minimum and maximum O/C and H/C ratio, presence of a  $^{13}\text{C}$  peak mass or comparing to theoretical isotope patterns) and only the masses in conjunction with their automated generated theoretical isotope pattern (existence of the  $^{13}\text{C}$  isotope) were taken into consideration (Hertkorn, et al., 2007).

In the next step, the fractional part of the transformed measured mass was compared to the fractional parts of the combinations in the database. Whenever a match within a previously defined error window (typically with 4 MW 0.1 ppm) was found, the integer part of the database mass was subtracted from the integer part of the ion mass which directly resulted in the number of hydrogen, as the H mass was previously scaled to 1.000000. The formulae generated with this method were afterwards subjected to more advanced modular plausibility filtering steps, e. g. involving a check for the nitrogen rule, minimum and maximum O/C and H/C ratio, presence of a  $^{13}\text{C}$  peak mass or comparing to theoretical isotope patterns.

The validity of the abovementioned algorithm was extensively checked by comparing different known, but slower, methods of elemental formula calculation. In the algorithm was not found errors of imputation.



### 3.3 Data transformation and normalization

The choice of the data pretreatment does not depend only on the biological information to be obtained, but also on the data analysis method chosen since different data analysis methods focus on different aspects of the data (Van den Berg, et al., 2006). Normalization techniques scaling and transformation will affect the results and the validity of the analysis. It is imperative to choose appropriate one, taking care also of these “new” data.

Data transformation is the process of changing the scale of the data so that it is more comparable from high to low. The main goals are to find appropriate transformations of the data, which make the data more suitable for multivariate analysis. Moreover these transformations stabilize the variance, because for high-complex assay this tends to rise with the intensities.

Common transformations, applied to ICR-FT/MS, are the logarithm and generalized logarithm:

- Log transformation
- Generalized log-transformation

$$g(\lambda) = \ln (y^2 + \sqrt{y^2 + \lambda})$$

In the formula of Generalized log-transformation  $y$  is the original spectroscopic intensity,  $g$  the transformed intensity and  $\lambda$  is a transformation parameter, it can be estimated using a maximum likelihood method using a set of replicate measurements (Box, et al., 1964), (Rocke, et al., 2003). The generalized log transformation was used especially with data in which the low intensities values have a predominant importance.

Normalization is the process of adjusting for systematic differences from one array to another. This row operation makes the spectra directly comparable with each other (Craig, et al., 2006). In the same case the concentration of a specific metabolite can be determined by an independent means (e.g., glucose using conventional clinical chemistry) and this then provides a reference value (Anthony, et al., 1994).

One common method of normalization involves setting each observation (spectrum) to have unit total intensity by expressing each data point as a fraction of the total spectral integral. Normalization may be done before or after transformation.

Scaling is done in the columns of the data. The most convenient normalizations to use with these data are:

- Pareto: the intensity is divided by the square root of the standard deviation:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$$

- UV (Unit Variance): Variable j is centered and scaled to unit Variance, i.e. the base weight is computed as  $1/s_j$ , and with  $s_j$  is the standard deviation of variable j computed around the mean. With this normalization variable with a very small variance will obtain an equal impact on the results:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$$

### 3.4 Similarities and distances between the data

The task of detecting patterns of relations, trends, and anomalies is made considerably easier when “similar” variables are arranged contiguously and ordered in a way that simplifies the pattern of relations among variables. This is referred to as “main effect ordering”.

To have a first visualization of the data it was useful many times to design a correlation matrix. It computes the correlation coefficients of the columns of the matrix. Here it is presented (see figure 3.5) an example taken by the analysis of Supernatant Positive of Salina Bacter Ruber (correlation matrix: 37x37).

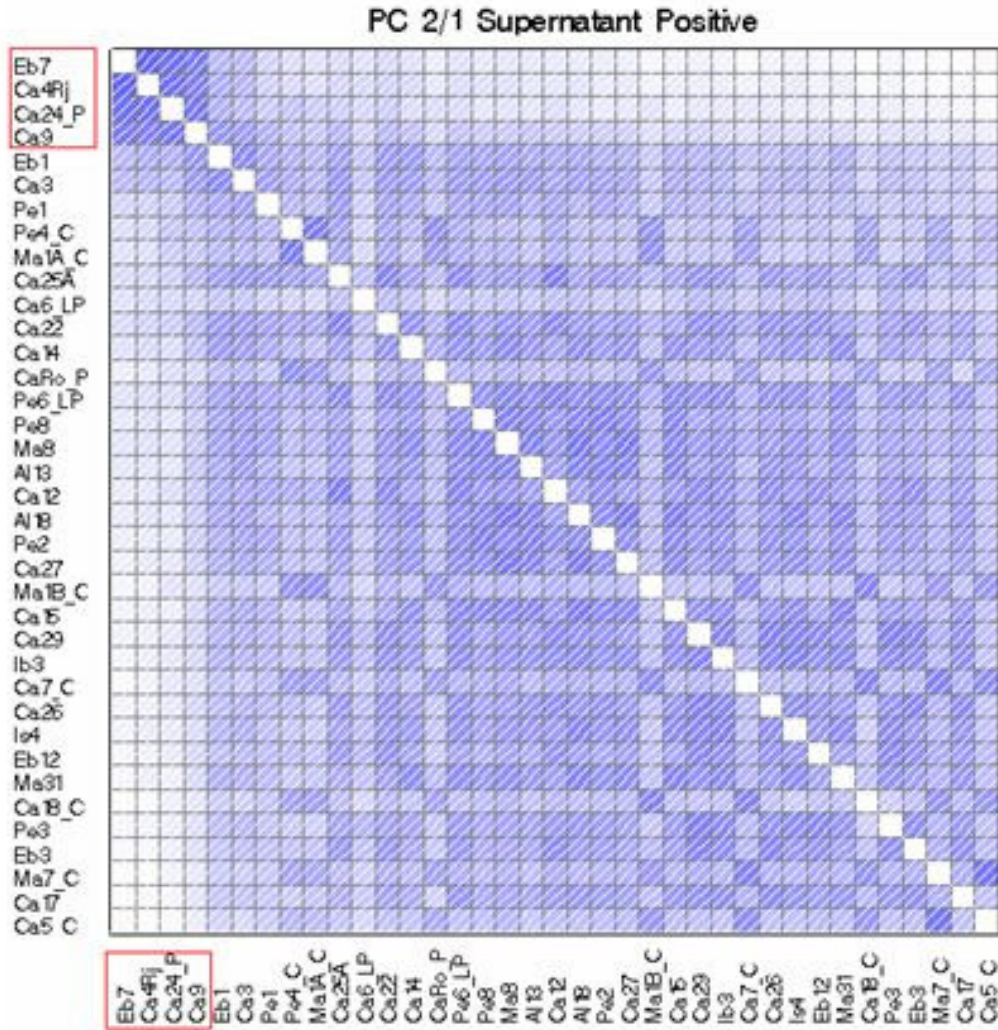


Figure 3.5: Correlation matrix of the Salina Bacter data. In the red the outlier samples are represented. They were subsequently detected also with multivariate techniques. This symmetric matrix defines possible similarities among the samples.

## 3.5 Statistical analysis

Setting up a matrix is the starting point for the creation of a statistical model. It becomes useful for summarizing and visualizing a huge amount of data, for classifying and discriminating and finally finding quantitative relationships among the variables with the possibility to make also a prevision for new samples. In many cases the final aim however is obtaining a plausible list of metabolites or biomarkers that explicate the phenomena being studied. The large numbers of peaks in the spectra that are all potential biomarkers create modeling and validation challenges (Westerhuis, et al., 2008). A multivariate analysis based on projection methods represents a number of efficient and useful methods for the analysis and modeling of the complex data. These methods include principal component analysis (PCA) and partial least -squares (PLS) (Trygg, et al., 2006). This modus operandi is robust and efficient for modeling and analysis of complicated chemical and biological data, which are characterized by noisy missing values and collinear data structure.

Singularity and collinearity: variables are said to be collinear if they are highly correlated. The problematic related with the collinearity are:

- highly correlated variables ( $\rho > .9$ ) make matrix inversion unstable and problematic and can lead to failures in calculation
- collinear variables can complicate make models difficult to interpret
- collinear predictors in a linear model can cause large standard error estimates, reducing statistical power

However it is difficult and still a matter of debate to determine when a model has become sufficiently detailed for its task and how confident one can be in its predictions. To fully realize the potential of information contained in this complexity science much more investigation and software developing is necessary.

The initial objective in metabonomic is to classify a spectrum based on identification of its inherent patterns of peaks and, second, to identify those spectral features responsible for the classification, which can be achieved via supervised and unsupervised pattern recognition technique (Westerhuis, et al., 2008).

When a statistical model is set-up and validated, it is possible to extrapolate a list of masses characteristic for the different groups' object of study. These masses are the possible biomarker. Here we give an example of the methods to

extrapolate a list of biomarkers; they are represented in figure 3.6b) the chemical property are visually represented with the Van Krevelen Diagram (see figure 3.6 (Rossello´-Mora, 2008)). The variables (single m/z) discriminative for each class were chosen according to their correlation coefficient value (Rossello´-Mora, 2008). Finally we could confirm or figure out the most important metabolite ranking them and applying simple univariate analysis of ANOVA (analysis of variance), Student t-test or non parametric equivalents can be used to ascertain if there is any statistically difference between individual metabolites (Altmaier, et al., 2008).

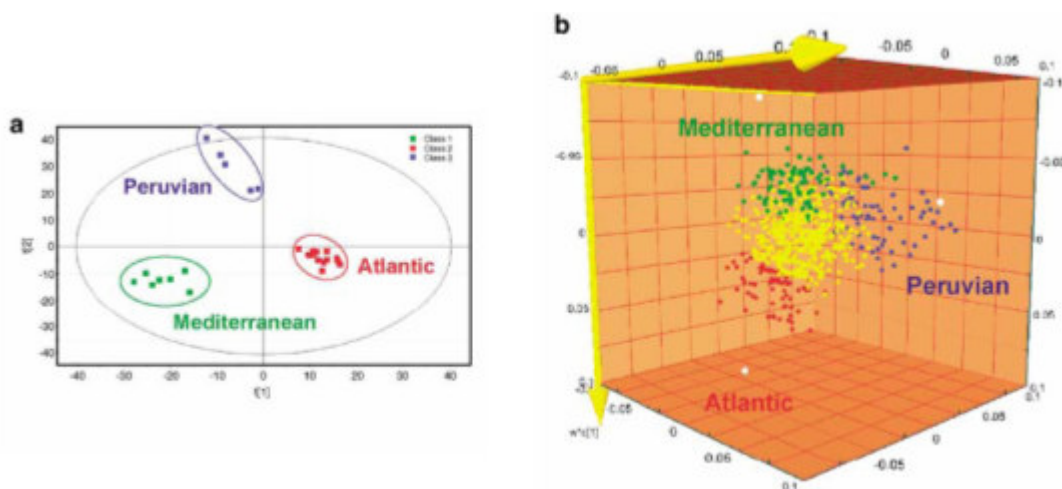



Figure 3.6: A) Score plot of the PLS-DA of all cellular insoluble fractions analyzed with electrospray-positive mode ICR-FT/MS showing the differentiation based on geographical origin of the samples; B) loading plot of the PLS-DA model correlating the 2099 m/z values of known elementary composition (C, H, O, N, S and m/z<550) to the geographical origins. The m/z values having a high correlation with geographical origin are highlighted with a corresponding color; the no discriminating masses are represented in yellow. The chemical properties of the masses have been investigated with the van Krevelen diagram and querying the general metabolome databases ([www.metabolome.jp](http://www.metabolome.jp), [www.genome.jp/kegg/](http://www.genome.jp/kegg/)).

### 3.5.1 Multivariate analysis

For meaningful interpretation, the appropriate statistical tools must be employed to manipulate the large raw data sets in order to provide a useful, understandable, and workable format. Different multidimensional and multivariate statistical analysis and pattern-recognition programs have been developed to distill the large amounts of data in an effort to interpret the complex metabolic pathway information from the measurements (Nicholson, et al., 1999), (Boutilier, et al., 2005), (Smith, et al., 2006). The multivariate analysis up to now is one of the most powerful tools able to interpret natural phenomena. This is divided into: unsupervised and supervised analysis. This classification depends on whether there is information available when the investigation starts or not. Before the statistical investigation it is necessary to transpose the data (see figure 3.7).

m/z	Spectra 1	Spectra 2	Spectra 3	...	Spectra n
m/z (1)					
m/z (2)					
m/z (3)					
m/z (4)					
...					
m/z (j)					
Property (1)					
Property (2)					
...					
Property (i)					

**Transpose** 

m/z	VAR 1 (m/z (1))	VAR 2 (m/z (2))	VAR 3 (m/z (3))	VAR 4 (m/z (4))	... VAR j (m/z (j))	Y1(Property (1))	Y2(Property (2))	... Yi(Property (i))
Spectra 1								
Spectra 2								
Spectra 3								
...								
Spectra n								

Figure 3.7: Before any statistical investigation it is necessary to transpose the matrix of data. The spectra will be the new observations; instead all the m/z and the spectra properties will be the new variables. This is the possible configuration to analyze complex system.

### 3.5.1.1 Unsupervised analysis

Principal component analysis (PCA) is a multivariate projection method used to compress information contained a data table or matrix  $X$  into a few so-called “principal components” (see figure 3.8). The objective of the compression is to explain as much of the variation in the original data set as possible. This is achieved by using the new form of latent variables, the principal components (PC). By reducing the dimensionality of the data it becomes much easier to get an overview of the variation, so that groups, trends and outliers can be identified among the observations. The reason why such a compression is possible is that variables are correlated with each other. If the variables were independent (uncorrelated), compression using PCA would not be possible. In many cases correlations between variables (e.g. metabolites) occur because they change according to some systematic underlying common factor. PCA has the ability to detect these underlying factors and compress the information based upon them. Each principal component consists of one score vector  $t$  and one loading vector  $p$ . The score can be regarded as the new variable and the loadings as the link between the original variables. Scores are linear combinations of the original variables and the influence of the original variables is represented in their loadings. By viewing the score and loading plot simultaneously it is possible to interpret the variables that influence the positions of the observations in the scores. Another very convenient tool to use with PCA-analysis is to view the loading plot and interpret which variables are positively correlated (located in the same quadrant) or negatively correlated (located in the opposite diagonal quadrant) to each other.

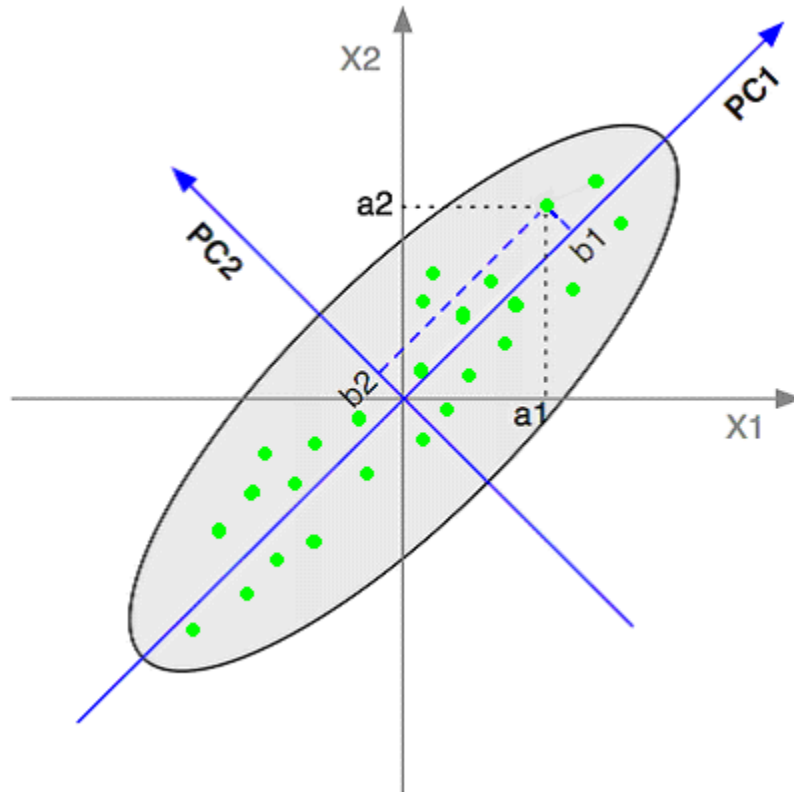


Figure 3.8: this figure shows the coordinates for a point in the original coordinate system (attributes  $X_1$  and  $X_2$ ) and in the PC coordinate system (PC1 and PC2). For the PC coordinate system, the majority of points can be distinguished from each other by just looking at the value of PC1.

The PCs define a new basis that is a model for describing the data. The projections of the observations onto the axes of the new basis define their coordinates in the model. The values of the coordinates are the scores ( $t$ ). By plotting these scores it is possible to get a visualization of the structure of the data (see figure 3.3a), in this case there are two natural groups: one follows their fulvic fractions characteristic and another one is characterized by humic fractions. This example is taken from (Lucio, et al., 2006), in which the implications



between the pesticide properties and the humic structures is described. The dataset was analyzed with PCA. The score scatter plot of the first two principal components is shown in figure 3.9a. Samples close to each other in the plot have similar properties with respect to different fractions. Fractions far from each other are dissimilar; there were denoted four zones. The first principal component (horizontal axis) mainly describes the aromaticity (in the right side) and aliphaticity (in the left). Acidity (in upper part of the graph) is negative correlated with N-containing functional groups.

The meaning of the scores, the impact of different variables on the model, is given by the orientation of the model with respect to the variables (see figure 3.9b).

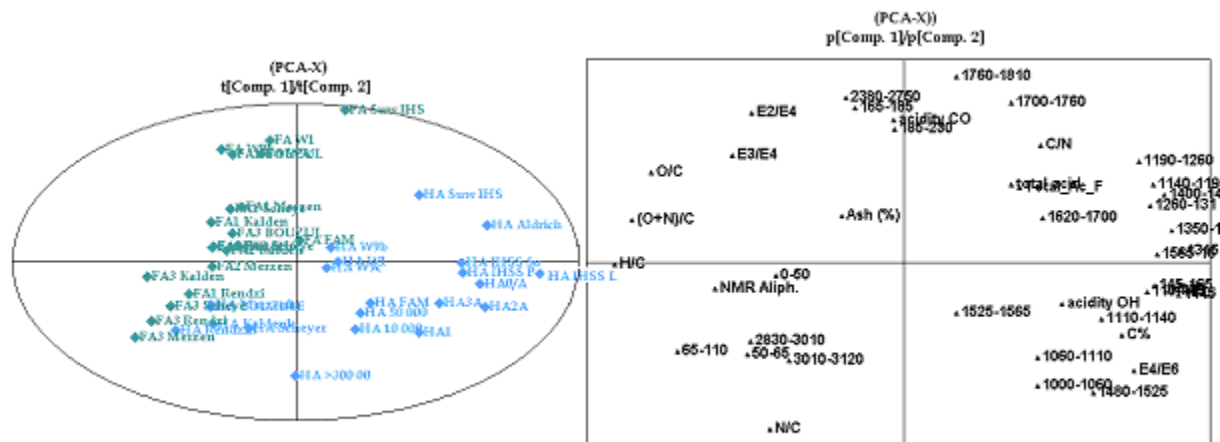


Figure 3.9: In the left panel the score scatter plot (t1/t2), and in the right panel the loading plot (p1/p2), right part B), from PCA essential grouped in two classes constituted of the Fulvic (left) and Humic fractions (right), illustrating similarities and dissimilarities among samples (Lucio, et al., 2006).

In the coordinate transformation the cosines of the angles between the old and the new coordinate system are called loadings  $p$ .

Algebraically, they can be explained as how the variables are linearly combined to form the scores.

The values of all observations projected onto the first principal component make up the vector  $t_1$ , and the scores for PC2 another vector,  $t_2$ . Similarly, the loadings calculated between the variables and PC1 constitute the vector  $p_1$ , and between the variables and PC2 the vector  $p_2$ .

The decomposition of a mean centered  $X$  matrix to the scores, loadings and residuals ( $E$ ) can be written using the following formula:

$$X = TP^T + E = t_1 p_1^t + t_2 p_2^t + \dots + t_\alpha p_\alpha^t + E$$

The data reduction is accomplished by neglecting unimportant directions where the sample variation is insignificant. This is repeated until no significant direction in the  $K$ -dimensional is left, i.e. the residual. The maximum number of components ( $a$ ) is the same as the number of variables. The number of significant PCs can be estimated by a number of methods, such as calculating the size of eigenvalues (Jackson, 1991) or cross-validation. After all significant variation in  $X$  has been described by the PCA model the remaining variation, the residual, is non-systematic and represents the distance between each point in the  $K$ -space and its point on the plane.

As previously mentioned, the plot of the scores describes the structure of the data. This plot is called a score scatter plot. Observations grouped together in a score scatter plot have similar properties, since they are described similarly by the principal components.

Similar to the principal component analysis is the hierarchical cluster analysis (HCA). In multivariate analysis HCA is a general approach to cluster analysis. Its purpose is to find relatively homogeneous cluster/groups whose members are all “close” to one other, based on measured characteristics (an example is given by the figure 3.10).

A key component of the analysis is the repetition of the calculation between objects and clusters once objects begin to be grouped into cluster.

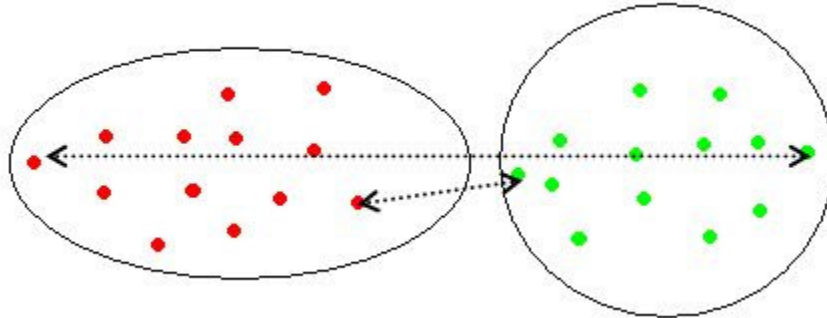


Figure 3.10: Plot of two clusters of observation (● on the left part and ● on the right part of the graph) and the different type of distance (single and complete linkage, respectively takes the smallest and the largest possible distance).

The single linkage clustering method (or the nearest neighbor method) is a method of calculating distance between clusters in hierarchical cluster analysis. The linkage function specifying the distance between two clusters is computed as the minimal object-to-object distance denoted by:

$$d(x_i, y_j)$$

where the objects  $x_i$  belong to the first cluster, objects  $y_j$  belong to the second cluster. In other words, the distance between two clusters is computed as the distance between the two closest objects in the two clusters.

Mathematically the linkage function is described by the following expression:

$$D(X, Y) = \min_{x \in X; y \in Y} d(x, y)$$

$D(X, Y)$  is the distance between objects  $x$  and  $y$  and  $X$  and  $Y$  are two sets of objects (clusters) The complete linkage clustering (or the farthest neighbor method) is a method of calculating distance between clusters in hierarchical

cluster analysis. The linkage function specifying the distance between two clusters is computed as the maximal object-to-object distance. In other words, the distance between two clusters is computed as the distance between the two farthest objects in the two clusters. Mathematically the linkage function is:

$$D(X, Y) = \max_{x \in X; y \in Y} d(x, y)$$

The output is a dendrogram; it is a tree-like plot (see figures 3.11 and 3.12), where the branches represent cluster obtained on each step of hierarchical clustering.

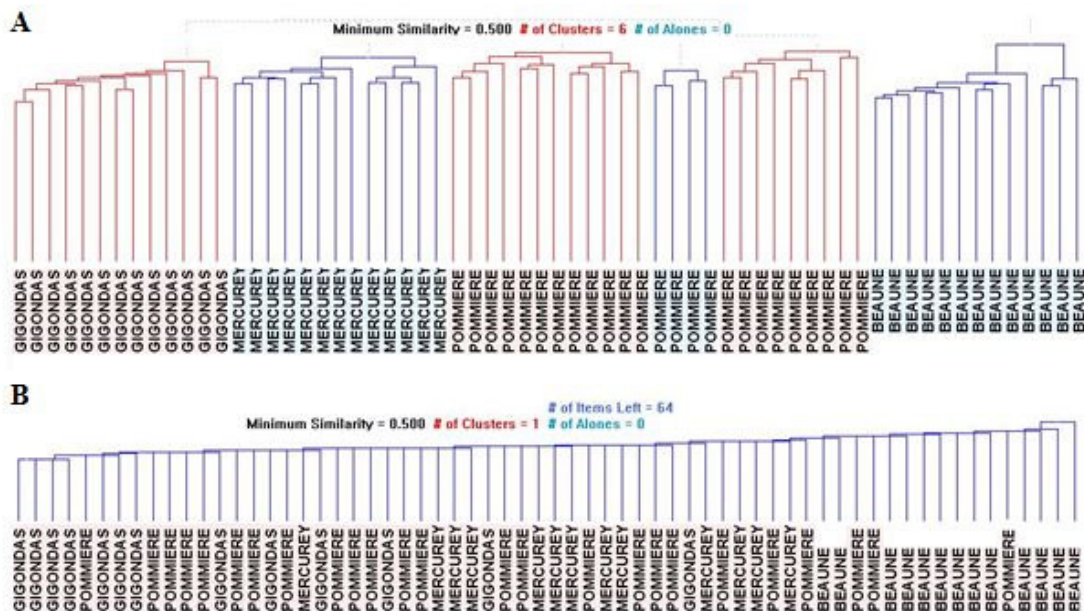


Figure 3.11: The upper panel A shows a complete linkage and B shows a single linkage. The clusters due to the analysis group perfectly according to the origin of the wine (Data from Tonnellerie 2000,

measured in negative mode with S/N=1 filtered with the formula calculator, CHONS rules).

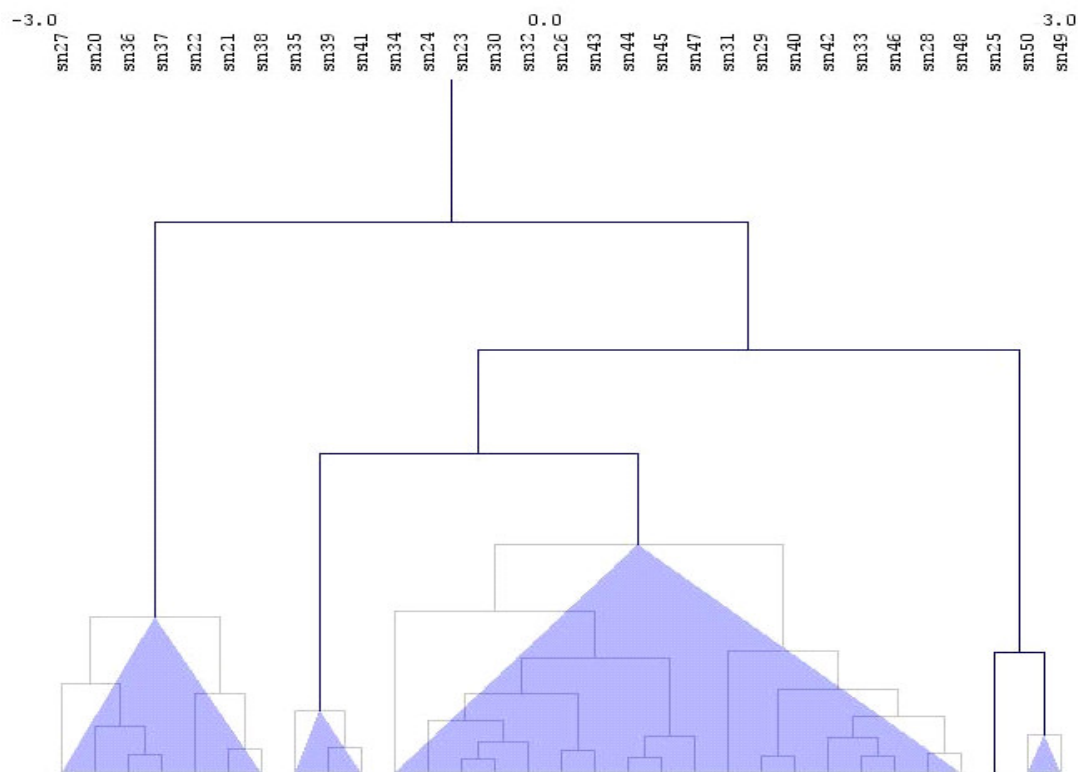


Figure 3.12: Salina Bacter Ruber, cluster analysis with complete linkage, in the right part there are the outliers.

### 3.5.1.2 Supervised analysis

The classification, based on the metabolic profile, is one of the main issues in this research. The classical method for this purpose is the partial least squares discriminant analysis (PLS-DA) (Vong, et al., 1988), (Barker, et al., 2003). It is a multivariate method used to classify and it is suitable when the number of experiments (in this occasion spectra) is small compared to the amount of variables (m/z) and when it is present multicollinearity (Geladi, et al., 1986). Unfortunately this method frequently over fits the data and rigorous

validation is necessary with the cross validation and permutation testing (see paragraph 3.6). Supervised PLS-DA analysis uses independent (expression levels: the X block) and dependent variables (classes: the Y block) for class comparisons. It is a classical partial least regression (PLS) an example is given by the figure 3.13.

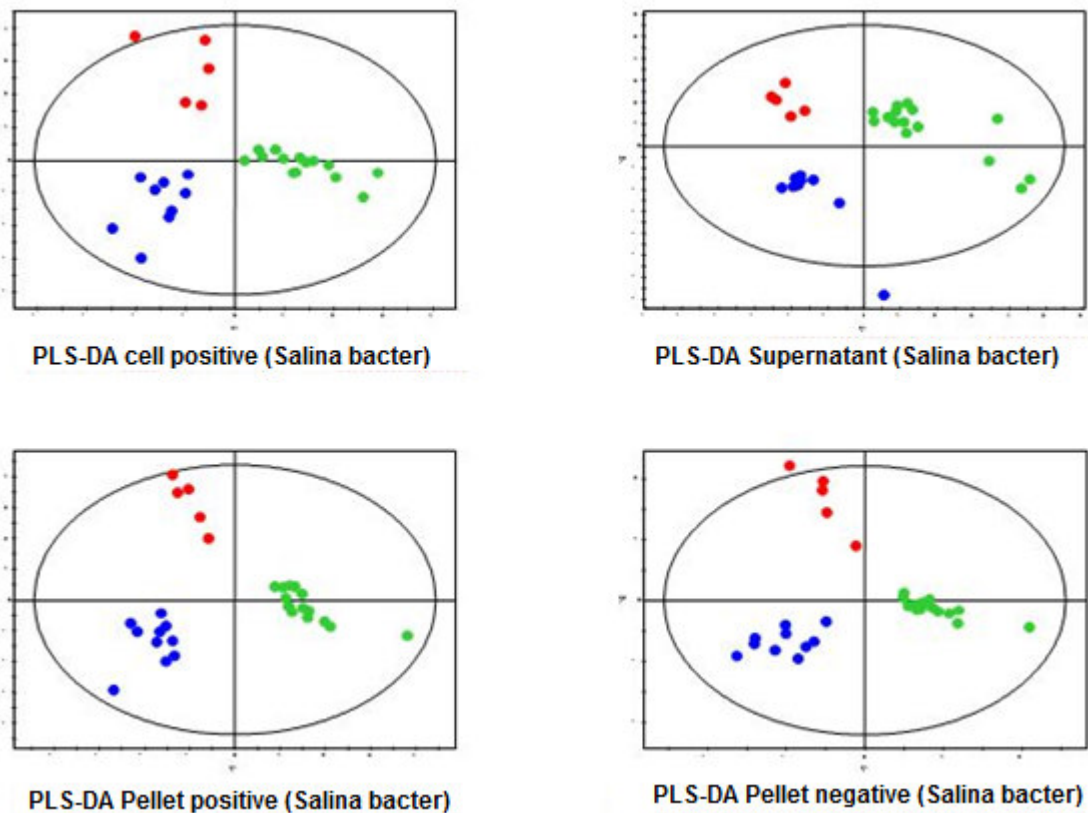


Figure 3.13: Four statistical models (PLS-DA) built up for the Salina Bacter Ruber (Supernatant/ Pellet Positive and Negative, Cell Positive). They are grouped according to their geographical origin (Atlantic ●, Peruvian ●, Mediterranean ●). Through the “glog transformation” was possible to clarify the association within the masses especially in Cell and Supernatant in which it was not clear with other transformations.

Additionally, PLS-DA provides a quantitative estimation of the discriminatory power of each descriptor by regression coefficient. The magnitude of the coefficient represents the relative importance of each data on the separation of the classes (Brindle, et al., 2002), and measures the effect of particular substances identified by m/z value (Lee, et al., 2003). The PLS-DA and Partial least squares (PLS) modeling could be used to determine the relative metabolite concentration of the metabolites of interest. Partial least squares projections to latent structures or PLS is a regression extension of PCA, which is used to find the relationship between a predictor matrix  $X$  and a response matrix  $Y$ , where the response matrix contains additional characterization of the samples in  $X$ . PLS uses the underlying or latent variables (scores) of  $X$  in order to describe the variation in both  $X$  and  $Y$ , in contrast to PCA which only models  $X$ . Therefore, the objective is also different when calculating latent variables in PLS, namely to extract the variation in  $X$  needed to predict the variation in the response  $Y$ . PLS-DA with OPLS-DA (see chapter 3.4) are often used in metabolomic field for classification (Barker, et al., 2003), (Bylesjö, et al., 2006), (Trygg, 2002), (Trygg, et al., 2002), (Rossello´-Mora, 2008). The main benefit with OPLS-DA lies in the ability to separate predictive variation from variation that is uncorrelated to it, in order to facilitate understanding of different sources of variation (Trygg, 2008).

OPLS-DA provides a solution to identify different sources of variability, both predictive and uncorrelated, and also facilitates understanding of any sampling, experimental, or preprocessing issues (Trygg, 2008).

Mathematically the  $X$  block is summarized by the  $X$ -scores  $T$ , and the variation in the response block,  $Y$ , is described by the  $Y$  scores  $U$ . Basically PLS maximizes the covariance between  $T$  and  $U$  (Trigg). For each model dimension a weight vector  $w$ , is calculated, and it reflects the partial contribution of each variables  $X$  to the modeling of  $Y$ . Therefore the matrix of the weights is  $W$ , reflects the structure in  $X$  that maximizes the covariance between  $T$  and  $U$ . The  $Y$ -weight instead is represented by the matrix  $C$  and parallel is computed also the matrix of the  $X$ -loadings,  $P$ . The  $P$  matrix is calculated in order to deflate  $X$  appropriately. The formulas for the decomposition are shown here:

$$X = TP' + E \quad Y = TC'$$

The set of PLS regression coefficients are computed according to:

$$B = W(P'W)^{-1}C''$$

The  $\hat{y}$  prediction is given by:

$$\hat{y} = x'W(P'W)^{-1}C'' = x'B$$

If it is necessary to remove from  $X$  information that is orthogonal to  $Y$ , within a PLS model, it is possible to apply Orthogonal Signal Correction (OSC). The data are transformed (mean-centered) and then OSC can be used to remove one component at a time from  $X$  using the algorithm for calculating the principal components of a data set (NIPALS algorithm: Nonlinear Iterative partial Least Squares).

In the context of classification it is important to show the value of sensitivity and specificity. Both parameters can be expressed in the context of a hypothesis test, sensitivity is an estimation of  $(1 - \alpha) \times 100$ , where  $\alpha$  is the probability of first type error (that is the *pr{to reject  $H_0/H_0$  is true}*) and the specificity is  $(1 - \beta) \times 100$ , where  $\beta$  is the *pr{reject  $H_0/H_0$  is false}*. Thus, specificity is related to the second type error, the power of the test is  $1 - \beta$ .

### 3.6 Data validation

The validation of the statistical models was crucial during the data analysis of the experiments. To this end we applied several methods described here below.

The use of a single cross validation (CV) may lead to bias and overestimation of the true error rate. For this reason it is the method always used for finding the optimal model. This procedure, now standard in the multivariate analysis, starts from the assumption to exclude from the model development a portion of the data and with the rest develop a number of parallel models and predict the omitted by the different models. Finally the predicted values are compared to the actual



ones. The squared differences between predicted and observed values are summed in the predictive residual sum of squares (PRESS), computed as:

$$PRESS = \sum \sum (x_{ik} - \hat{x}_{ik})^2$$

For every component (i), the overall PRESS/SS is computed, where SS is the residual sum of squares of the previous component, and also  $(PRESS/SS)_k$  for each Y variable (k). These values are good measures of the predictive power of the model. Normally the model is also evaluated through two different values:  $R^2$  (goodness of the fit) and  $Q^2$  (goodness of the prediction), respectively calculate as:

$$R^2 = 1 - RSS/SSX_{tot. corr}$$

$$Q^2 = 1 - PRESS/SSX_{tot. corr}$$

Where RSS are the residual sum of squares and  $SSX_{tot. corr}$  represent then total variation in the X matrix after mean -centering. They are value in the range [0,1]. An excellent model is with  $Q^2 > 0.9$  and the difference between the value of  $R^2$  and  $Q^2$  may not exceed 0.2/ 0.3 (see figure 3.14a).

Jack-Knife define the stability of the regression coefficient (Efron, 1982), (Martens, et al., 2000) and in the estimation of confidence intervals, it is used in the PLS procedure.

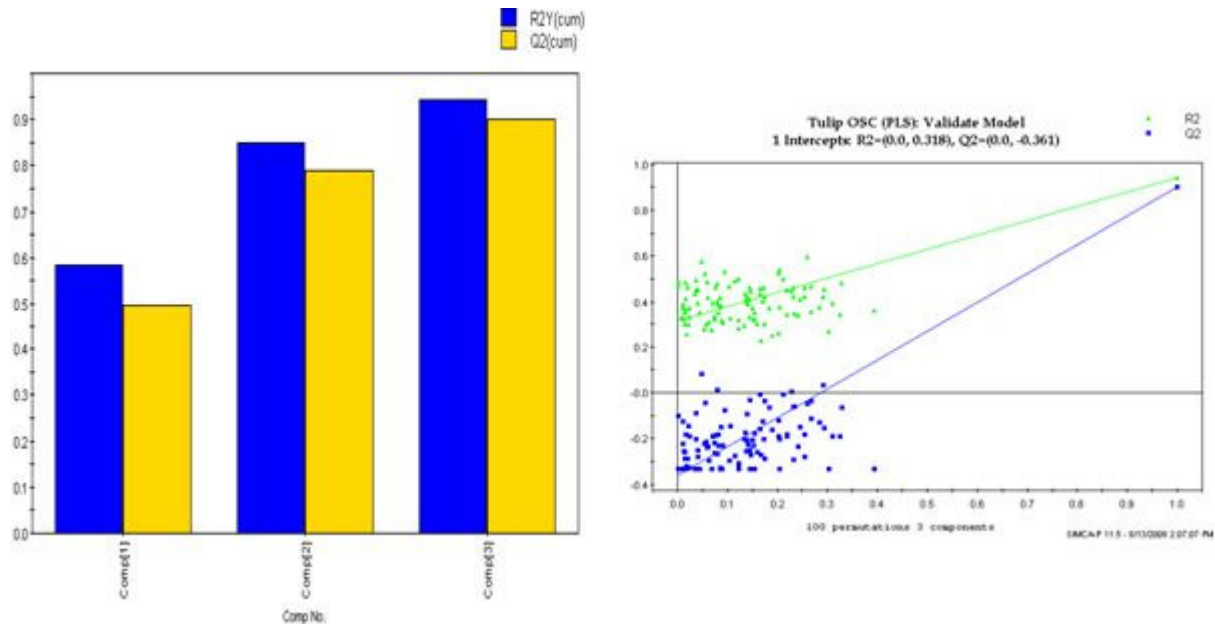


Figure 3.14: Example of validation. Data come from a study on pre-diabetic state, which precedes overt type 2 diabetes. Left panel: value of  $Q^2$  and  $R^2(Y)$ . Right panel: a random permutation test (100 permutations) was applied to assess the robustness of PLS-DA model. These data do not exhibit any characteristics of an overfit model.

In order to assess the discrimination an exact or an approximate permutation test can be used (see figure 3.14b), (Efron, et al., 1993), (Manly, 1997), (Mielke, et al., 2001). The study was performed using plasma of 47 metabolically characterized, non-diabetic individuals at high risk to develop type 2 diabetes. The spectra analyzed with ICR-FT/MS were evaluated by an OSC PLS.

Parallel to these methods, classified as “internal validation”, it is possible to evaluate the model also with an “external validation”. Normally all the dataset is divided into two different groups (they should be representative of the population and of the classes present in the model): one is called training set the other one validation set. Using the first one a model is created, and this could also predict the remaining dataset. This method has been applied when a large number of samples are available.

## 3.7 Data evaluation and visualization

The ability to collect and visualize biochemical relevant information about metabolites has only recently become available (Breitling, et al., 2006). Many visualizations can be used in metabolomic field. The most common are dendrogram from clustering results (Roessner, et al., 2001), and representations of samples in principal components analysis (Nicholson, et al., 1999), (Fiehn, et al., 2000), (Roessner, et al., 2001). With the non target approach the visualization becomes also a tool to describe and summarize the property of a class of objects (spectra). Due to the big amount of information present in a dataset it was necessary to develop a strategy to summarize the chemical property of a class of objects. This was done with different visual approaches.

So the main goal of a statistical model is to compiling a list of important masses (VIP list or regression coefficient list, etc.) which play an important rule for the purpose of statistical model. After it is possible to investigate its properties in two ways using:

- Van Krevelen and/or Kendrick mass plots
- MasSTRIX ([www.masstrix.org](http://www.masstrix.org)) annotates metabolites in high precision mass spectrometry data.

The chemical property of the list of candidate biomarker can be visualized with the Van Krevelen method (1961). It represents a graphical method where the atomic hydrogen/carbon (H/C) ratio is plotted as a function of the atomic oxygen/carbon (O/C) ratio. In this diagram each molecular formula represents a compound is shown as a point whose coordinates are determined by elemental composition [Meija, 2006].

The same list can be submitted to MasSTRIX in order to display on organism specific KEGG pathway maps, and optionally add any additional genomic or transcriptomics information by highlighting the corresponding enzyme boxes.

### 3.7.1 Van Krevelen diagram

Van Krevelen developed a graphical method to study the process in which the atomic hydrogen/carbon (H/C) ratio is plotted as a function of the atomic oxygen/carbon (O/C) ratio. Van Krevelen diagram is often used for the classification of coals and kerogens. A frequent application of the van Krevelen

diagram is to illustrate the changes in elemental compositions that occur during the alteration of organic geochemicals in a geologic environment; e.g. H/C and O/C ratios have been used to follow the effects of diagenesis on humic substances (Hue, et al., 1977), (Reuter, et al., 1984). These diagrams have also been used by various authors to illustrate compositional differences between humic acids and fulvic acids, and also to show variations in humic substances as a function of source. For example, Kuwatsuka (Kuwatsuka, et al., 1978) used a van Krevelen diagram to compare the elemental compositions of soil humic and fulvic acids, coals, plant tissues and various classes of organic compounds.

Visser (Visser, 1983) employed a van Krevelen diagram to compare fulvic and humic acids from aquatic and terrestrial sources. The magnitude of the H/C ratio has also been used to indicate the degree of aromaticity or unsaturation (a small value) or aliphaticity (a large value) of a substance. Perdue (Perdue, et al., 1983) has pointed out that the total unsaturation of a humic material cannot be obtained solely from the H/C ratio; in addition to unsaturated forms of carbon the H/C ratio is also a function of unsaturation present in functional groups, primarily carboxyl and carbonyl groups, with lesser contributions from other miscellaneous forms of unsaturation. If H/C ratios are calculated for the 21 humic material samples in the study of Perdue (Perdue, et al., 1983) and compared to the aromatic carbon contents corrected for the various forms of noncarbon unsaturation it is seen that, though the actual numbers differ, samples which exhibit a high aromatic carbon content also exhibit a small H/C ratio and vice versa. The lone exception is a spodosol fulvic acid with high total acidity (12mequiv/g), a low corrected aromatic carbon content but a moderate H/C ratio (0.85). The H/C ratio thus appears to be a qualitatively useful parameter for comparing the aromaticities of humic materials. To date, the number of humic samples plotted on a single van Krevelen diagram has been relatively small. The value of any such investigation would be enhanced by enlarging the data base and by using humic substances from a wider variety of source environments. In addition, when a large data set is employed one can justifiably apply statistical methods to quantify the relationships between the various groups of humic substances by establishing the statistically significant differences.

For example, the spectrum of a red wine from Burgundy (i.e. Vosnes Romanée, 1995) can lead up to 17,400 peaks at a signal-to-noise = 2, (115,000 at a signal-to-noise = 1), which can be unambiguously attributed to 1180 unique elemental CHONS compositions with 200 ppb tolerance and confirmation with <sup>13</sup>C-signal (3890 compositions at 500 ppb tolerance), from which only a few hundred may

correspond to masses of metabolites such as those gathered in our database (see figure 3.15), that have already been observed in model solutions or in wines with targeted analyses.

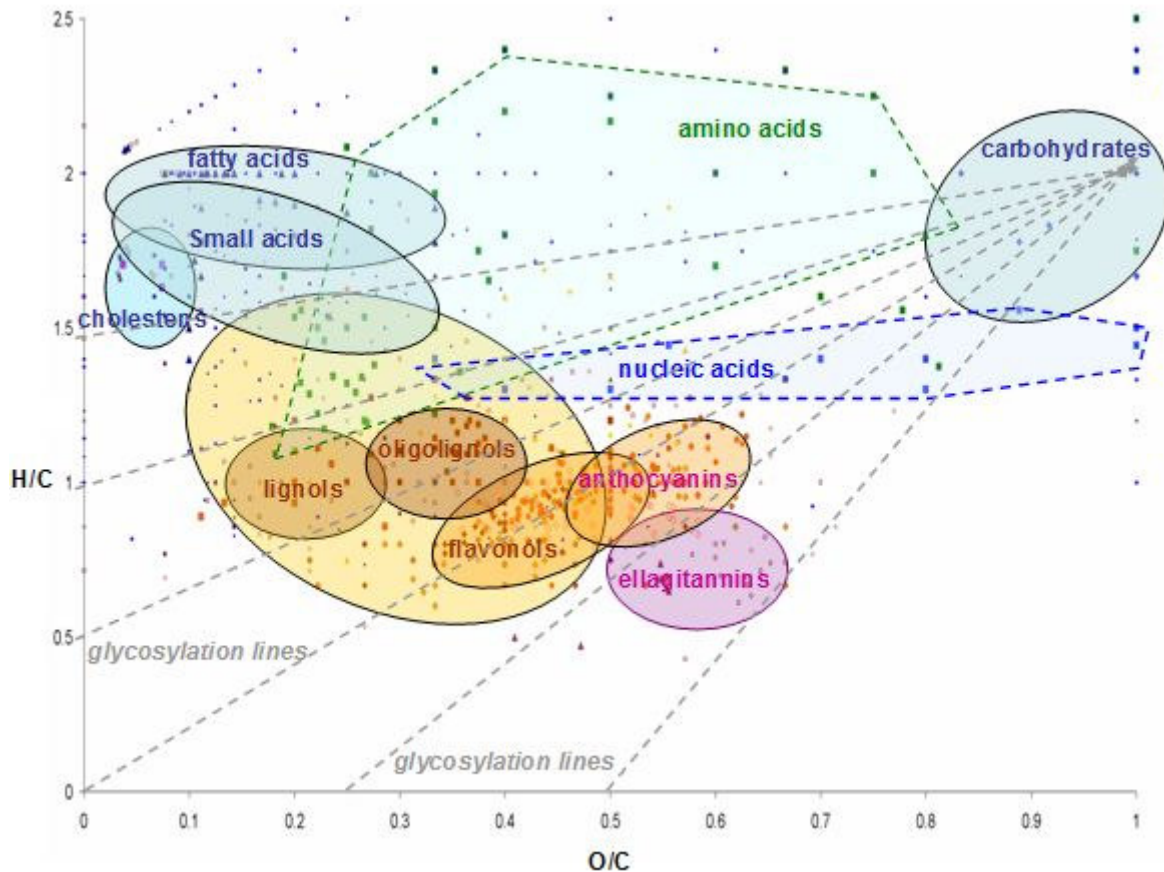


Figure 3.15: Example of van Krevelen representation based on the compounds from the database illustrated in figure 6.2 (chapter 6), here showing the positioning of various classes of molecules. The glycosylation line represents the virtual line along which the (O/C, H/C) values would move on the diagram when following successive glycosylations (for example: anthocyanins and their corresponding mono and di-glycosylated anthocyanins).

A graphical representation of the various chemical spaces (CHO, CHOS, CHON, CHONS) of wines are then obtained (see figure 3.16), which visually highlight specific cluster series of elementary compositions observed within nominal masses. Using a home-compiled database of compounds that can exist in model wine solutions or that have been actually observed in wines, allows to similarly represent the specific contributions of phenolics, peptides, polysaccharides, nucleotides and any other classes of compounds present in wines, and which can be positively or negatively ionized. It must be noted however, that many of the compounds responsible for the aroma of wines, which exhibit  $m/z$  values below 150, are not detected under our experimental conditions.

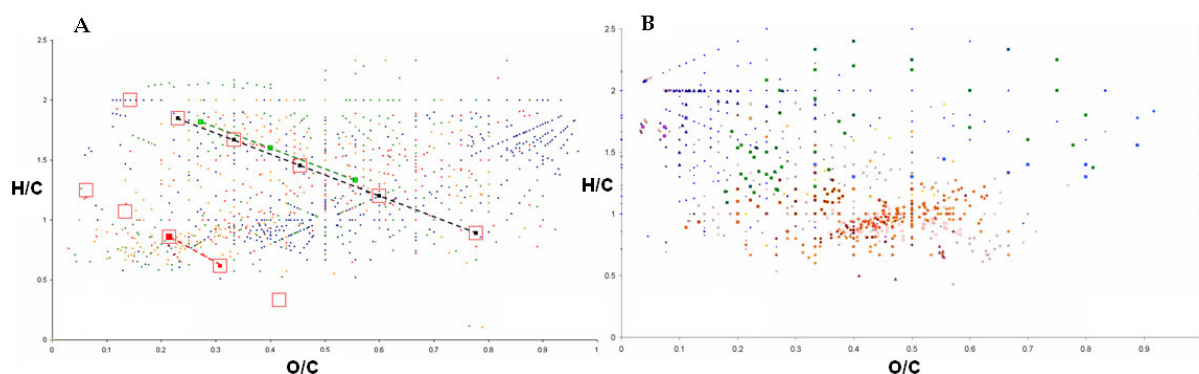


Figure 3.16: A): Example of visualisation in van Krevelen diagram (H/C versus O/C atomic ratios) of the three series of elemental compositions from figure 6.2 (chapter 6) in the light of the chemical space of over 2.000 cumulated elemental compositions found in the white Chardonnay Beaune 1998 and Pinot Noir Mercurey 1998 (colour code: CHO, CHOS, CHON, CHONS) B): van Krevelen diagram (H/C versus O/C atomic ratios) of our home-compiled database of compounds that can exist in model wine solutions or that have been actually observed in wines.

### 3.7.2 From masses to database

Once the interested masses were found it was necessary to find the putative structures and the metabolite profile. To this end an approach was identified supported by several databases accessible from the web. They give biochemical information, and they are able to combine different species and experimental condition.

At the same time we started to build our own laboratory database. This is built around the known information achieved by the experimental result, based on the literature (inherent with the aerosol) and on the wine database (exemplification in figure 3.15, it was done in collaboration with Regis Gougeon, Institut Universitaire de la Vigne et du Vin - Jules Guyot, Dijon) or based on the information available in data base as KNAP-SACK DB.

The main structure of the database is presented in figure 3.17.

ID	DB_Name	Formula	Type	Description	C	O	H	N	S	F	P	Cl	Br	K	H_C	O_C	N_C	Negativ_M	Neutral_M	Positiv_M
11475	GSF_DB	C4H7NO2	2-(Aminomethyl)-2-propenoic acid		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11476	GSF_DB	C4H7NO2	2-Amino-2-butenic acid		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11477	GSF_DB	C4H7NO2	2-Amino-3-butenic acid		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11478	GSF_DB	C4H7NO2	2-Azetidinecarboxylic acid		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11479	GSF_DB	C4H7NO2	3-Aminodihydro-2(3H)-furanone		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11480	GSF_DB	C4H7NO2	4-Hydroxy-2-pyrrolidinone		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11481	GSF_DB	C4H7NO2	5-Hydroxy-2-pyrrolidinone		4	2	7	1	0	0	0	0	0	0	1.75	0.5	0.25	100.0404	101.04768	102.05495
11482	GSF_DB	C5H11NO	3-Methylbutanoic acid; Amide		5	1	11	1	0	0	0	0	0	0	2.2	0.2	0.2	100.07679	101.08406	102.09134
11483	GSF_DB	C5H12NO	N-(2-Hydroxyethyl)aziridine; N-Me(1+)		5	1	12	1	0	0	0	0	0	0	2.4	0.2	0.2	101.08461	102.09189	103.09917
11484	GSF_DB	C4H12BNO	Trihydro(morpholine-N4)boron, 10Cl		4	1	12	1	0	0	0	0	0	0	3	0.25	0.25	89.084613	90.091889	91.099165
11485	GSF_DB	C6H15N	1-Hexylamine		6	0	15	1	0	0	0	0	0	0	2.5	0.17	0.17	100.11317	101.12045	102.12773
11486	GSF_DB	C6H15N	3-Methyl-1-butylamine; N-Me		6	0	15	1	0	0	0	0	0	0	2.5	0.17	0.17	100.11317	101.12045	102.12773
11487	GSF_DB	BrNa	Sodium bromide (NaBr)		0	0	0	0	0	0	0	0	1	0				77.911062	78.918338	79.925614
11488	GSF_DB	CHCl2F	Dichlorofluoromethane		1	0	1	0	0	1	0	2	0	0	1			100.93666	101.94393	102.95121
11489	GSF_DB	C2H2F4	1,1,1,2-Tetrafluoroethane		2	0	2	0	0	4	0	0	0	0	1			101.00199	102.00926	103.01654
11490	GSF_DB	C4H6O3	2-Oxobutanoate		4	3	6	0	0	0	0	0	0	0	1.5	0.75		101.02442	102.03169	103.03897
11491	GSF_DB	C4H6O3	4-Hydroxy-2-oxobutanal		4	3	6	0	0	0	0	0	0	0	1.5	0.75		101.02442	102.03169	103.03897
11492	GSF_DB	C4H6O3	4-Methyl-1,3-dioxolan-2-one		4	3	6	0	0	0	0	0	0	0	1.5	0.75		101.02442	102.03169	103.03897
11493	GSF_DB	C4H6O3	4-Oxobutanoic acid		4	3	6	0	0	0	0	0	0	0	1.5	0.75		101.02442	102.03169	103.03897
11494	GSF_DB	C4H6O3	Acetoacetate		4	3	6	0	0	0	0	0	0	0	1.5	0.75		101.02442	102.03169	103.03897

Figure 3.17: This is a part of the structure of the database developed in our laboratory and computing system.

Only recently was available MasSTRIX which annotates metabolites in high precision mass spectrometry data (data coming from ICR-FT/MS). The process compares a submitted mass peak list (experimental masses) against all compounds of the KEGG chemical compound database, additionally including  $^{13}\text{C}$ ,  $^{15}\text{N}$  and other isotopes, and optionally adding selected lipids with variable fatty acid chain

lengths. Raw input masses from electrospray ionization (ESI) MS can be corrected on-the-fly for the addition or the abstraction of a proton (and optionally a Na ion in positive mode). To cope with the requirement of very low measurement errors (in the sub-ppm range), exact masses of all KEGG compounds have been recomputed from the corresponding chemical formula using high-precision atomic mass data (Wapstra, et al., 2003). MassTRIX then calls the KEGG/API (<http://www.genome.jp/kegg/soap/>) to generate pathway maps, where the annotated compounds and genes are highlighted using different colors-thus differentiating between organism-specific and extra-organism items (Suhre, et al., 2008). From a list of non characterized components MassTRIX works like a filter giving back all of them that are assigned and recognized in KEGG database. But there are still many masses that play an important role in the statistical models, but they are unknown and not yet classified yet. On the other hand we can detect the formula structure of isomers but we cannot differentiate between each of them since ICR-FT is a qualitative method, so we cannot have information about their stereoisomer. Only by putting together different technologies it is possible to avoid this limit of the analytical techniques. In the case of multiple assignments the only information we can achieve is their chemical formula calculated by the software.



# Chapter 4

## 4 METABOLIC EVIDENCE FOR BIOGEOGRAPHIC ISOLATION OF THE EXTREMOPHILIC BACTERIUM *SALINIBACTER RUBBER*

### 4.1 Introduction

Biogeography constitutes a cornerstone approach for studying biodiversity patterns at different taxonomic levels in the prokaryotic world (Ramette, et al., 2006).

The biogeography of prokaryotes and the effect of geographical barriers as evolutionary constraints are currently subjected to great debate. Some clear-cut evidence for geographic isolation has been obtained by genetic methods but, in

many cases, the markers used are too coarse to reveal subtle biogeographical trends. Until today most of the studies searching for the geographical isolation of prokaryotic populations and their divergence have been directed towards genetic characters (Green, et al., 2006), (Huges-Martiny, et al., 2006), (Ramette, et al., 2006), (Whitaker, et al., 2003).

In this regard, the importance of geographic barriers influencing microbial speciation is subjected to great debate (Whitaker, et al., 2003), and the old microbiological tenet of '*everything is everywhere, but, the environment selects*' has served as a starting point for research on environmental forces that may lead to genetic and phenotypic allopatric segregation (de Wit, et al., 2006).

The difficulties in retrieving phenotypic information, which is thought to be a stepping stone for taxonomic classifications (Staley, 2006), may hamper the discovery of divergences in prokaryotic populations at the phenotypic level. Clear biogeographic differences have been observed for some prokaryotes, such as thermophilic Archaea (Whitaker, et al., 2003) and *Cyanobacteria* (Papke, et al., 2005), and for fluorescent pseudomonads (Cho, et al., 2000).

Contrary to eukaryotic microorganisms, phenotypic evidence for allopatric segregation in prokaryotes has never been found. Currently, only phenotypic differences matching biogeography have been reported for eukaryotic microorganisms (Fenchel, et al., 2006). Prokaryote taxonomy, diversity, and ecology have benefited from the developments of molecular techniques. Ribosomal RNA - based approaches (Amann, et al., 1995), genomics, and currently, metagenomics (Green-Tringe, et al., 2005) are the major sources of information for understanding the diversity of the prokaryotic biome. These approaches give information of paramount importance, but only at the genetic level. However, analyzing the expression of the genotype may lead to a better understanding of the interactions microbes have with their environment. A microorganism is not only a composite of its genome, but the multiple expressions of its genotype (Cavalier-Smith, 2007), and there is a significant part of the genome that might never be expressed (Ochman, et al., 2006). Approaches such as functional transcriptomics and proteomics may be considered as a dynamic link between the genome and the cellular phenotype (Singh, 2006), the real interaction of the organism with its environment. We recently showed that the extremely halophilic bacterium *Salinibacter ruber* (Antón, et al., 2000) can be isolated from different parts of the world in sites as diverse as Mediterranean coastal solar salterns (Peña, et al., 2005) and the remote Andean Peruvian salterns of Maras at 3,380 m above sea level (Maturrano, et al., 2006). *S. ruber* growth is constrained to relatively

small water bodies with high salt concentrations in restricted areas on Earth. The extreme conditions and geographical isolation of its environments are optimal circumstances for observing allopatric speciation, as demonstrated for the hypertermophilic archaeon *Sulfolobus* (Whitaker, et al., 2003), and thermophilic *Cyanobacteria* (Papke, et al., 2005).

Only a metabolomic approach, based on ultrahigh resolution mass spectrometry, was able to reveal phenotypic biogeographical discrimination. This procedure was skillful to demonstrate that strains of the cosmopolitan extremophilic bacterium *Salinibacter ruber*, isolated from different sites in the world, can be distinguished by means of characteristic metabolites, and that these differences can be correlated to their geographical isolation site distances. The approach allows distinct degrees of discrimination for isolates at different geographical scales. In all cases, the discriminative metabolite patterns were quantitative rather than qualitative, which may be an indication of geographically distinct transcriptional or posttranscriptional regulations. ICR-FT/MS enables the assignment of thousands of elemental compositions of metabolites in a mass range from 120 to 800 Dalton directly out of complex mixtures by virtue of ultra high mass accuracy (< 100 ppb) and ultrahigh resolution (>1,500,000 at mass 600) at high field strength. This represents the initial, but crucial, step in metabolite annotation, for instance, by use of various targeted databases (i.e. KEGG, Kyoto Encyclopaedia of Gene and Genome database). This technique is acquiring an increasingly important position in “metabolomics” (Want, et al., 2007) together with spectroscopic methods, such as nuclear magnetic resonance spectroscopy (NMR) (Nicholson, et al., 1999). However, high-field ICR-FT/MS (Marshall, 2004) showed to have the highest resolution among all spectrometric methods in revealing fine scale diversity in complex mixtures. This method may help in revealing phenotypic patterns of geographically isolated organisms at the level of the direct interaction with the environment (phenotype) that may not be clearly indicated by indirect interaction (genotype). This chapter is based on the article: “Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*” (Rossello´-Mora, 2008).

## 4.2 Materials and methods

Strain isolation and culture conditions: Brine samples were directly plated onto 25% SW agar medium supplemented with 0.1% yeast extract (Antón, et al., 2002). Plates were incubated at 37°C until growth was observed. Subsequent colonies were isolated in pure cultures, and those corresponding to *S. ruber* were studied further. Liquid 25% SW medium supplemented with 0.2% yeast extract was used to grow biomass in liquid conditions with vigorous shaking at 37°C. For the metabolomic studies, all strains were inoculated and incubated for the same time under the same conditions. Biomass was harvested by centrifugation. Table 4.1 indicates the list of strains used in this study, their origin and year of isolation. Two growth batches were prepared in order to evaluate two simultaneous independent experiments: a complete set of strains from all different locations, and a second batch made up with four to five replicates of selected Mediterranean strains (i.e. 13 and P18 from Alicante, M8 and M31 from Mallorca and IL3 from Ibiza).

Strains	Origin	Area considered	Year of isolation
M8, M31	Mallorca, Llevant salterns	Mediterranean	2000
P13, P18	Alicante, Santa Pola salterns	Mediterranean	2000
E1, E3, E7, E12	Tarragona, Trinidad salterns	Mediterranean	2001
IL3	Ibiza, Ibiza salterns	Mediterranean	2001
ES4	Israel, Eilat salterns*	Mediterranean	2001
C3, C4Rj, C6, C9, C12, C14, C15, C17, C22, C25A, C26, C27, C29	Canary Islands, La Palma salterns	Atlantic	2001
PR1, PR3, PR2, PR6, PR8	Perú, Maras salterns	Peruvian	2003

Table 4.1: List of *S. ruber* strains used in this study and their isolation origin. \* Eilat Salterns are located by the Red Sea, but we consider it as Mediterranean for proximity and climate similarities.

Metabolite extract preparation: A total of 3 ml of cell suspension grown on liquid media were collected by centrifugation. Two milliliters of cell-free supernatant were stored for further chromatographic extraction. Supernatant was acidified by the addition of 50  $\mu\text{l}$  of 98-100% formic acid (MERCK KGaA, Darmstadt, Germany). Pelleted biomass was then suspended in 1 ml of bidistilled water, and sonicated to obtain a clear lysate extract. The lysate was then acidified by the addition of 50  $\mu\text{l}$  of 98-100% formic acid. After the acidification, the clear lysate formed insoluble aggregates that could be separated from the soluble fraction by centrifugation. The clear supernatant was stored for further fractionation, and the insoluble pellet was resuspended in 500  $\mu\text{l}$  of methanol. Sample preparation resulted in three complementary fractions: the extracellular, cellular soluble and cellular insoluble fractions. Solid phase extraction: Both acidified extracellular and cellular soluble fractions were solid phase extracted using Bond Elut C18 columns (Varian Inc.). This chromatography enables the isolation of the organic molecules on the basis of their non-specific interaction and retention to the C18 material. This purification removes the high salt charge of the media and extracts, which may interfere during the electrospray procedure by ion-suppression (Li, et al., 2006). The retained fraction was recovered by the use of methanol.

#### 4.2.1 ICR-FT/MS procedure

Broad scan mass spectra were acquired on a Bruker (Bremen, Germany) APEX Qe Fourier transform ion cyclotron resonance mass spectrometer equipped with a 12 Tesla superconducting magnet and an APOLLO I electrospray (ESI) source, whereas high resolution spectra were acquired with an APOLLO II ESI source in positive and negative mode. The samples were infused in methanol with a microelectrospray source at a flow rate of 120  $\mu\text{l}/\text{h}$  with a nebulizer gas pressure of 20 psi and a drying gas pressure of 15 psi (200 °C). Spectra were externally calibrated on clusters of arginine (10 mg/l in methanol), and calibration errors in the relevant mass ranges were always below 100 ppb, which is the prerequisite for an adequate elementary composition assignment. Relative standard deviation in the intensity values of the peaks was routinely lower than 5% in our analysis conditions. The spectra were acquired with a time domain of 1 megaword (1 million bits in size) with a mass range of 150 - 2,000 m/z. The spectra were zero filled to a processing size of 2 megawords. A sine apodization was performed before Fourier transformation of the time-domain transient. The ion accumulation

time in the ion source was set to 0.2 s and 1024 scans were accumulated for samples.

ICR-FT/MS spectra were exported to peak lists at a signal to noise ratio S/N=1. From these lists, possible elemental formulas were calculated for each peak in batch mode by a software tool written in-house. The generated formulas were validated by setting sensible chemical constraints (nitrogen rule, atomic oxygen to carbon ratio  $O/C \leq 1$ , atomic hydrogen to carbon ratio  $H/C \leq (2n+2)$ , element counts: carbon  $C \leq 100$ , oxygen  $O \leq 80$ , nitrogen  $N \leq 5$ , sulphur  $S \leq 1$ ) and only the masses in conjunction with their automated generated theoretical isotope pattern (existence of the  $^{13}\text{C}$  isotope) were taken into consideration (Hertkorn, et al., 2007). The obtained reduced peak lists were compared in m/z at 5 ppm and the corresponding intensity matrices were generated for further statistical analysis.

### 4.3 Targeted approach

The targeted approach allowed a detailed analysis of specific metabolites following a specific chemical structures hypothesis after the previous metabolomic screening of the samples. The analysis, especially in high resolution mode, enables a detailed description of the natural isotopic abundance that in addition allows confirmation of the elementary composition assignments. Figure 4.1 shows the assignment of the elementary compositions (including isotopic peaks) to the m/z as obtained in negative electrospray ICR-FT/MS in two different resolution modes. Mass intensity data related to sulfonolipids were analyzed statistically by one-way analysis of variance (ANOVA) with post-hoc Bonferroni's test for multiple comparisons (Holm, 1979). Probabilities less than 5% ( $P < 0.05$ ) were considered statistically significant.

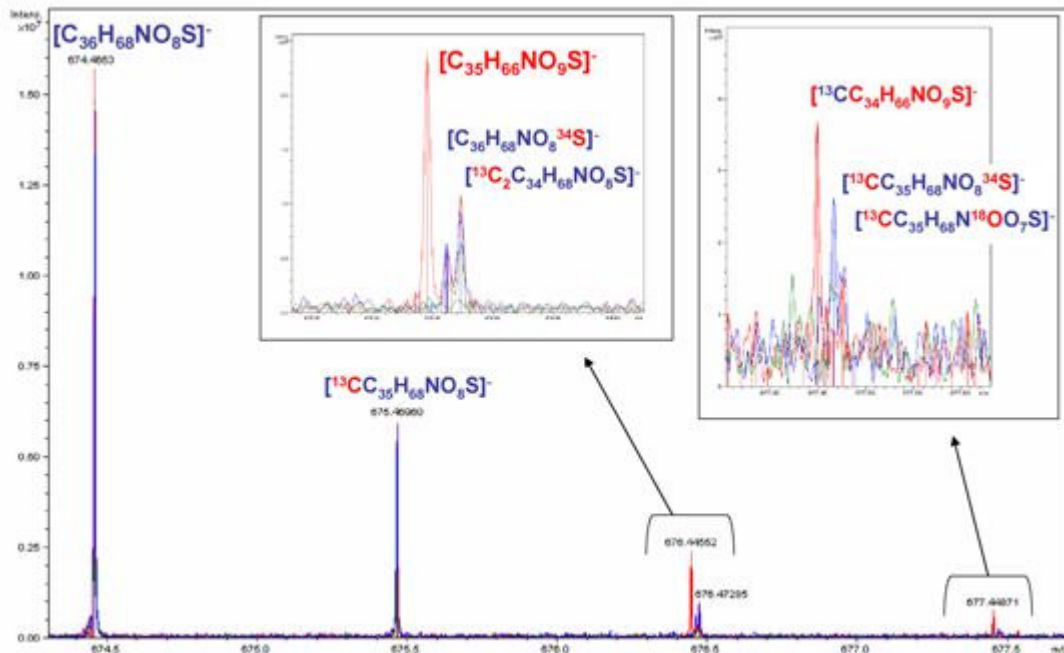


Figure 4.1: Detailed spectra on mass 674.4663 identified as a sulfonolipid in negative mode electrospray ICR-FT/MS within the series described in Table 4.5 and its corresponding natural abundance isotopic pattern. Mass 676.4455 was found only in the Atlantic samples

## 4.4 Statistical analysis

The data were imported and analyzed in SIMCA-P 11.5 (Umetrics, Umea, Sweden). The statistical model used was Partial Least Squares for Discriminant Analysis (PLS-DA), (Sjöstrom, et al., 1986), (Stahle, et al., 1987), (Vong, et al., 1988), (Kemsley, 1996). PLS-DA is a regression extension of the principal component analysis (PCA), (Wold, et al., 1987). It takes advantage of class

information (in this case the geographical origin of the samples) in order to maximize the separation between groups of masses. A list of masses ( $m/z$ ) discriminative for the different geographical area is produced. The PLS-DA uses the X variables (matrix of masses) as predictors, and dummy variables (belonging or not belonging to a given class coded as 1/0; i.e. origin of isolation) as response variables (Y variables). All three modalities (extracellular, cellular soluble and cellular insoluble) were calculated independently and cellular insoluble was chosen as the descriptive power of the model. The descriptive power can be defined by several terms, most directly the fraction of the Sum of Squares (SS) of all the Y explained by the current component ( $R^2Y(\text{cum})$ ) and  $Q^2(\text{cum})$ .  $R^2Y$  provides an estimate of how well the model fits the Y data and  $Q^2$  provides an estimate of how well the model predicts the Y data.

$m/z$	Insoluble cellular fraction	Soluble cellular fraction	Extracellular fraction	Sum of all masses	Cumulative unique masses
Number of unique masses from raw data (S/N=1)	168,444	157,378	161,322	487,144	247,655
Number of unique masses after CHONS calculation	3,456	5,293	5,062	13,811	11,880
Number of masses used for statistical analysis ( $m/z < 550$ )	2,099	3,559	3,450	9,108	8,873
<hr/>					
Number of masses discriminative for Atlantic strains	181	74	80	335	333
Number of masses discriminative for Mediterranean strains	510	655	114	1,279	1,249
Number of masses discriminative for Peruvian strains	287	427	257	971	968
Number of masses from discriminative metabolome	1,121	2,403	2,999	6,523	6,323

Table 4.2: Number of observed masses from the analysis, considered masses for statistics, and masses for geographical discrimination (positive electrospray analysis).



Pareto scaling of the intensity values with a logarithmic transformation of the data was chosen in order to consider all masses equally, including those with medium and low intensity values (Van den Berg, et al., 2006). The cellular insoluble metabolome dataset contained 2,099 variables (see Table 4.2), from 28 observations measured in the three groups (Atlantic - Mediterranean - Peruvian). When analyzing this dataset with PLS-DA using four significant components,  $R^2Y(\text{cum})$  was equal to 0.98 and  $Q^2(\text{cum})$  was equal to 0.45 both, with values indicating high predictive power.

The score scatter plot and loading plots were presented already in figures 3.6a and 3.6b of the third chapter, respectively. The score scatter plot (see figure 3.6a, chapter 3) presents a view of how well the classes (different geographical origin) are separated on the basis of their X variables. In the loading plot (see figure 3.6b chapter 3), the different masses characteristic for each of the three classes are differently colored (red for Atlantic, green for Mediterranean and blue for Peruvian). The variables (single masses) discriminative for each class (origin of isolation) were chosen according to their correlation coefficient value. Those having the highest coefficients were considered to be relevant (i.e. variables (m/z) with a correlation value higher than  $|0.002|$ ). A total number of 180 out of 2,099 masses were considered to be discriminative for the classes (values shown in Table 4.2).

Interpretation of the regression coefficients provides information pertaining to the metabolic explanation of class differences (Holmes, et al., 2002) based on the fact that each coefficient is related to a specific elemental composition. Those masses associated with the highest correlation coefficient were represented in the van Krevelen projection (H/C versus O/C on the basis of their elementary composition values; figure 4.2a and figure 4.2b).

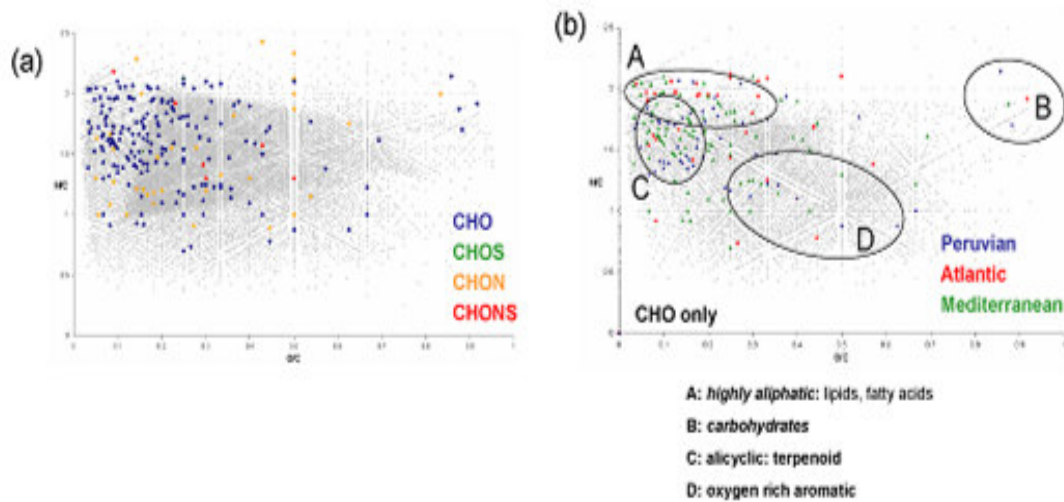


Figure 4.2: A) All discriminating  $m/z$  values independent of their origin but colored only as a function of their attributed elementary composition (CHO, CHON, CHOS or CHONS) and visualized in a van Krevelen diagram (H/C versus O/C). Most of the discriminative metabolites contain only C, H and O (only a few metabolites contain sulfur or nitrogen) and these are compared within a van Krevelen type of diagram to the CHONS containing metabolites of general metabolome databases ([www.metabolome.jp](http://www.metabolome.jp), [www.genome.jp/kegg/](http://www.genome.jp/kegg/)) shown in grey in the figure. Note that the triangular region corresponds to peptides (CHON and CHONS); B) CHO metabolites in a van Krevelen diagram colored as a function of their origin.

Moreover table 4.3 lists the co-ordinate values along the first and second components that numerically represent the similarities and differences among the strains. These values represent the distances resulting from the projection of the points on the first and second components to the origin (0 value). They explain the magnitude (large or small correlation) and the nature (positive or negative correlation) of the samples.

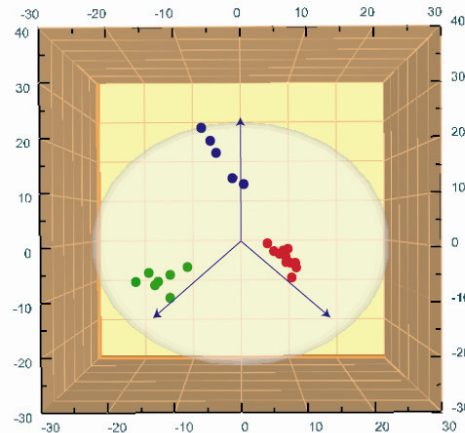


Figure 4.3: Score plot of the partial least square - discriminative analysis (PLS-DA). The interpretation of the figure indicated that each class was "tight" and occupied a small and separate volume in Xspace (X represents the number of the variable). The discrimination derived from this discriminative plane (where the projected observations occur) well separated the tree groups according to the differentiation of the elementary compositions visualized in the Van Krevelen diagram (see figure 4.2b) and their geographic location.

Sites	Strains	Co-ordinate value along the first component (score value)	Sites	Strains	Co-ordinate value along the second component (score value)
Atlantic	C27	9.81995	Peruvian	PR3	28.1714
Atlantic	C3	9.65824	Peruvian	PR8	22.4758
Atlantic	C29	9.63297	Peruvian	PR2	19.0242
Atlantic	C22	9.25211	Peruvian	PR1	13.2149
Atlantic	C12	8.53592	Peruvian	PR6	12.1288
Atlantic	C25A	8.11802	Atlantic	C14	-0.254274
Atlantic	C9	7.54702	Atlantic	C17	-1.17895
Atlantic	C4_Rj	7.52264	Atlantic	C12	-1.28982
Atlantic	C15	7.46169	Atlantic	C26	-1.49031

Atlantic	C6	7.03174	Atlantic	C9	-1.53197
Atlantic	C17	5.94576	Atlantic	C6	-1.87301
Atlantic	C26	5.91143	Atlantic	C4Rj	-1.96649
Atlantic	C14	4.65194	Atlantic	C15	-2.42006
Peruvian	PR6	0.602852	Atlantic	C25A	-2.58963
Peruvian	PR1	-1.11698	Atlantic	C29	-3.75884
Peruvian	PR2	-4.17265	Atlantic	C22	-4.04727
Peruvian	PR8	-5.25783	Atlantic	C27	-4.47822
Peruvian	PR3	-7.62429	Atlantic	C3	-4.78183
Mediterranean	E1	-9.13346	Mediterranean	E1	-4.57788
Mediterranean	E7	-11.6219	Mediterranean	M31	-6.4909
Mediterranean	M31	-12.8236	Mediterranean	IL3	-6.56391
Mediterranean	P18	-14.623	Mediterranean	P18	-8.09294
Mediterranean	E12	-15.1815	Mediterranean	E12	-8.61029
Mediterranean	IL3	-16.8664	Mediterranean	ES4	-8.76525
Mediterranean	ES4	-20.3118	Mediterranean	E7	-9.96454

Table 4.3: Co-ordinate value of the first and second components of the PLS-DA analysis (score). Strains close to each other have similar properties, common metabolites, whereas those far from each other are dissimilar with respect to the origin. From the inspection of the second component values, one factor that might contribute to the differentiation may be related to the geographical location of the origin of the strains. Peruvian is far away from Mediterranean but closer to Atlantic, and Peruvian is in fact negatively correlated with Mediterranean.

The scatter plot score of figure 4.3 summarizes the numerical coordinates present in Table 4.3 to provide a numerical perception of the group distance.

The similarity within stains M8, M31, P13, P18 and IL3 (intensities for each m/z value), was evaluated first by using the Levene's test (Malins, et al., 2002) which evaluates the differences (p-values) in the variances of each group of repetitions. Then, we used a one-way analysis of variance (ANOVA), and a Tukey test for repeated measures. All differences were considered to be significant when  $p < 0.01$ . The analyses were performed in SAS version 9.1 (SAS Institute Inc., Cary, North Carolina). At the  $p < 0.01$  level, the population variations were not significantly different (Table 4.4).

Sample	Sum of squares	Mean squares	F value	P value
M8	6.69x10 <sup>17</sup>	4.60x10 <sup>13</sup>	0.95	0.4331 (NS)
M31	7.67x10 <sup>17</sup>	5.70x10 <sup>13</sup>	0.56	0.6877 (NS)
Pola13	4.44x10 <sup>17</sup>	7.38x10 <sup>13</sup>	2.91	0.0547 (NS)
Pola18	6.96x10 <sup>17</sup>	1.26x10 <sup>13</sup>	2.36	0.0942 (NS)
IL3	5.27x10 <sup>17</sup>	9.67x10 <sup>13</sup>	0.6	0.5469 (NS)

Table 4.4: Analysis of Variance (one-way ANOVA) for the sample M8 (five replicates), M31 (five replicates), Pola13 (three replicates), Pola18 (three replicates) and IL3 (three replicates). For each group of replicates the p-value was greater than the significance level of 0.01, than at the 0.01 level, the population means were not significantly different (NS). Where the Sum of Squares measures variation present in the data, it is calculated by summing squared deviations, the mean square is the sum of squares divided by its associated degrees of freedom, the F value is the F statistic for testing the null hypothesis (the means are the same) and the  $P > F$  is the probability of obtaining a greater F statistic than that observed if the null hypothesis is true.

The discriminative analysis of the Mediterranean strains shown in figure 4.4 was undertaken with Orthogonal PLS-DA (OPLS-DA) based on the cellular soluble fraction. For this kind of sample, OPLS-DA rendered equivalent but clearer results than PLS-DA. In this case, OPLS-DA separates predictive from non-predictive (orthogonal) variations (Bylesjö, et al., 2006). Orthogonal-PLS (OPLS): The objective of OPLS is to accomplish a predictive model  $X \rightarrow Y$  ( $X$  is the matrix of spectral data and  $Y$  the response variables) where the systematic variation in the  $X$ -block is divided into two model parts, one part which models the correlations between  $X$  and  $Y$  and another part which expresses the variation in  $X$  that is not related (orthogonal) to  $Y$  (Eriksson, et al., 2006). The logic is of a regular PLS model, which, after filtering, has been divided in two parts, a predictive part and an orthogonal part. The number of predictive and orthogonal components is decided with cross-validation (Wold, 1978). Parallel raw matrices containing all variable characters among all strains studied, and coded as absence/presence of each peak, were reduced to an informative set by identifying all identical metabolites with different isotopic composition, and by reducing the background

noise by the use of peak thresholds as described in the Material and Methods section.

Improved binary matrices were analyzed by the use of the parsimony tool in the Phylip program package (Felsenstein, 1981) using the default parameters (<http://evolution.genetics.washington.edu/phylip.html>). Clustering analysis of binary matrices: Phenetic analyses were carried out by the use of the TREECON program version 1.3b (Van de Peer, et al., 1994), and by using UPGMA.

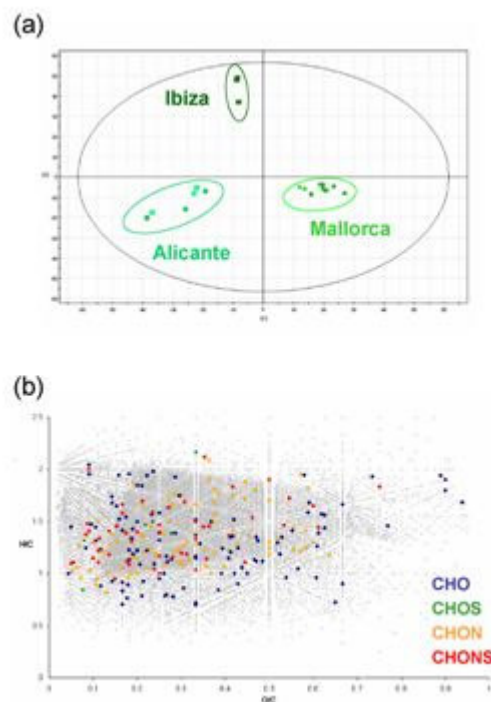


Figure 4.4: (a) Orthogonal partial least square - discriminative analysis (OPLS-DA) score plots of all cellular soluble fractions in the extracts of the Mediterranean strains from the locations of Alicante (P13 and P18), Ibiza (IL3) and Mallorca (M8 and M31). These metabolomes can be differentiated in a two component model with a high level of prediction ( $R^2(Y)=0.99$   $Q^2(\text{cum})=0.47$ ). (b) Representation of the discriminative masses in a van Krevelen diagram showing the high importance of

nitrogen containing metabolites representative of general metabolic pathways (confirmed by searching in public metabolite databases).

## 4.5 Inference on the biogeographic isolation

Biogeographic comparisons at the genetic level: representative members of *Salinibacter* spp. have been reported in several locations in the world either by molecular techniques (Antón, et al., 2002), (Mesbah, et al., 2007) or by culturing approaches (Antón, et al., 2002), (Maturrano, et al., 2006). These members of the Bacteria domain that do not show growth below 15% NaCl concentrations thrive in constrained environments that appear dotted on the earth's surface. An initial study based on our strain collection of about 17 strains isolated from several locations in Spanish coastal salterns indicated that a slight trend for geographic isolation could be discerned at the genetic level (Peña, et al., 2005). Contrary to observations for fluorescent pseudomonads (Cho, et al., 2000), ITS sequences were not suitable for studying biogeographical segregation due to their high sequence similarity. However, both PFGE and RAPD gave weak indications of geographical discrimination of genotypes. In no case were the analyses conclusive in proving allopatric segregation.

In this study, we enlarged the collection with about 28 strains isolated from 5 different locations in the world (Table 4.1). The isolates were obtained from five different locations in the Mediterranean area (Mallorca, Alicante, Tarragona, Ibiza and Israel), the Atlantic Canary archipelago (from a solar saltern on the island of La Palma), and from the 3,500 m high salterns in the Peruvian Andes (Maras). Ten of the isolates were selected to undertake MLSA, which represented the three main geographical areas in the study (west Mediterranean, Atlantic, and Peruvian Andes). The concatenated DNA stretch rendered an alignment of 7,995 homologous sites, 6,513 of them corresponding to seven protein gene sequences, with 129 of them being informative. Phylogenetic analyses were performed by including and excluding indels, as well as by using different datasets (including the 16S rRNA gene in the concatenate, figure 4.5a, or disregarding it, figure 4.5b). In general, the trees agreed with regard to their topology, since only M8 acquired a stable position when including the 16S rRNA gene sequence in the analysis.

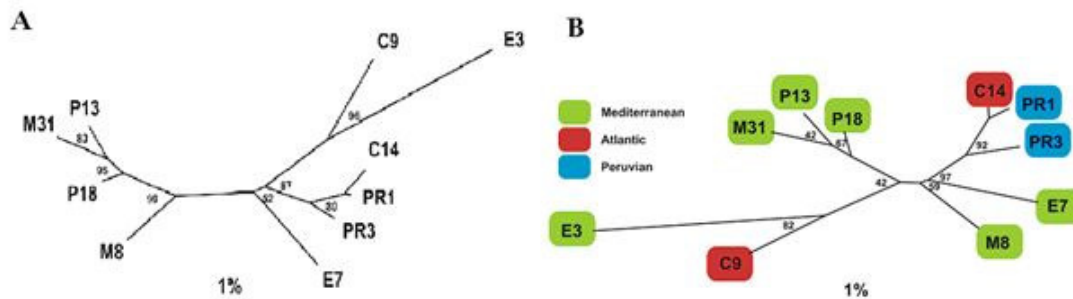


Figure 4.5: Phylogenetic reconstruction based on a PHYML algorithm of a 6,513 nucleotide alignment corresponding to the 7 housekeeping genes. Strains of different geographical areas are marked with their respective colors. The bar indicates 1% sequence divergence. It is remarkable that no geographical trend could be obtained. As indicated in the Material and Methods section, the same dataset was used to calculate reconstructions with several treeing approaches. Most of the trees gave congruent topologies independently of the use of PHYML or ARB (maximum likelihood and neighbor joining algorithms). Only ML showed slightly different topologies when including the indels in the analysis. In any case, none of the obtained tree topologies showed a clear geographic trend. Contrary to the same reconstruction where the SSU rRNA gene was included, bootstrap values were lower. However, despite a lower robustness of the tree topology, there was no doubt about the common affiliations between the Peruvian and Atlantic strains, and between C9 and E3.

Altogether, the reconstructions did not show clear geographical segregation of the selected strains, in contrast to observations made with other extremophiles (Whitaker, et al., 2003). Strains from Alicante (P13 and P18) affiliated together with that of Mallorca (M8 and M31). However, the Mediterranean strains E3 and E7 affiliated together with those from the Atlantic (C9 and C14) and Peru (PR1 and PR3). Neither our previous studies with fingerprinting techniques (Peña, et al., 2005), nor here with a MLSA of gene stretches of nearly 8,000 homologous positions were informative enough to resolve biogeographical segregation. This could be an indication that the process of genetic divergence is still at an early



stage and cannot render clearly resolvable trends. However, given that the size of the *S. ruber* genome is about 3,000 ORFs (Mongodin, et al., 2005), and despite the fact that we selected the genes to be sequenced from a set of putative phylogenetic markers (Sória-Carrasco, et al., 2007), the set of genes may not be adequate for understanding subtle geographical segregation. Intraspecific whole genome comparisons with *S. ruber* might in the future indicate which genes could be useful for understanding allopatric differentiation based on genetic drift.

Biogeographic comparisons at the phenotypic level: as stated above, genomic data is especially useful for solving the main problems in the classification of organisms, as well as understanding speciation processes (Staley, 2006), (Ward, et al., 2007). In most of the fields related to prokaryote diversity (taxonomy, ecology, speciation), phenotype studies are being relegated in favor of those based on genome information, such as MLSA or other genome analyses, due to the ease of the latter. However, standard genotyping techniques may not always help in clearly resolving intraspecific diversity. As has already been requested (Ramette, et al., 2006), there is a need to apply new approaches for understanding allopatric segregation of members of the same species. For this reason, we have evaluated the adequacy of a non targeted metabolite profiling approach, using high field ICR-FT/MS of the chemical extracts of our strain collection. Mass spectrometry has acquired a predominant position in “metabolomics” (Want, et al., 2007) and, especially, high-field ICR-FT/MS (Marshall, 2004). This technique provides ultra-high resolved profiles with thousands of accurate mass values ( $m/z$ ) that can be transformed into real elementary compositions.

For this study, a first experiment with twenty eight isolates of *S. ruber* from seven locations in the world (Table 4.1), divided into three geographical areas (Mediterranean (10 strains), Atlantic (13 strains) and Peruvian (5 strains)), were studied by ICR-FT/MS. All organisms were grown simultaneously under identical environmental conditions to avoid culture-dependent differences. Metabolome comparisons rendered a total of over 247,255 discriminative mass signals at  $S/N=1$  (signal to noise) that could be attributed to distinct elementary compositions containing the elements C, H, O, N and S. Single peak occurrence was reduced from 11,880 (verified by isotopic assignments of elementary composition) to a total of 8,873 metabolites at a  $m/z$  lower than 550 amu (highest probable assignments). The core metabolome (i.e. common peaks for all extracts) consisted of 2,550 single masses, whereas the discriminative metabolome (i.e. peaks not

common to all extracts) consisted of 6,323 single metabolites (Table 4.2). In all cases, the analyses were performed by using the whole metabolome.

With the raw information, the first comparative analyses were based on qualitative data coded as presence or absence of single metabolites. For this, the results were expressed in a binary matrix that was treated either cladistically using parsimony, or phenetically, using UPGMA (see figure 4.6). However, in no case could the profile analysis, based on independent covariant characters, reveal clear geographical trends. Therefore, it seemed that the presence or absence of single metabolite comparisons did not reflect geographical isolation.

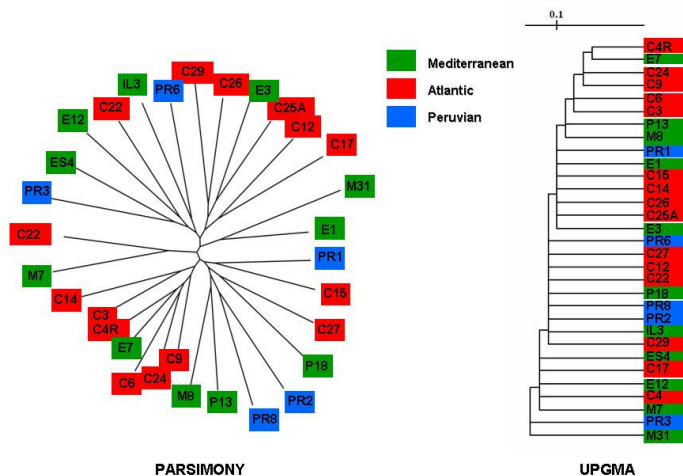


Figure 4.6: Cladistic (left) and phenetic (right) analyses of the binary matrix compiling all variable peaks that proved not to be common to all organisms. The binary matrix contained 9,108 single peaks that were treated as independent covariant characters or as homologous positions for both analyses, respectively. The matrix had been obtained as a reduction of 29,012 peaks from raw spectra by selecting all peaks corresponding to masses smaller than 550, unifying single metabolites with different isotopic compositions, and by removing background noise, as clarified in the Material and Methods section. Colors listed in the

legend indicate the origin of the strains: European (or Mediterranean), Atlantic (or Canary Islands), and Peruvian.

In contrast, weighing the relative intensity of each individual peak and treating the data by using multivariate analysis, revealed statistically significant differences between the different samples. Thus, metabolome comparisons, focusing on geographically discriminative data, yielded clear cut allopatric differences. The non-targeted analysis revealed unique features for each group of isolates (see figures 3.6a, 3.6b chapter 3). In this regard, the most relevant markers were CHO molecules (see figure 4a). Furthermore, van Krevelen plots of atomic O/C versus H/C ratios (Wu, et al., 2004) and comparisons with the total metabolic spaces (see figure 4.2b and table 4.5) showed that the discriminative metabolites may be aliphatic in structure and depleted in oxygen.

Mediterranean	Formula	Atlantic	Formula	Peruvian	Formula
122.00322	C3H5O3S	193.13354	C11H16N2O	121.04953	C4H8O4
122.02703	C3H7NO2S	193.15869	C13H20O	131.14304	C8H18O
126.09134	C7H11NO	194.04478	C9H7NO4	132.04439	C8H5NO
130.12264	C7H15NO	194.08117	C10H11NO3	146.06004	C9H7NO
133.10839	C4H12N4O	195.06518	C10H10O4	146.08117	C6H11NO3
137.03455	C6H4N2O2	209.05971	C14H8O2	165.13862	C10H16N2
139.03897	C7H6O3	217.17982	C12H24O3	167.03388	C8H6O4
157.02438	C5H4N2O4	219.17434	C15H22O	167.04512	C7H6N2O3
157.12231	C9H16O2	221.13835	C10H20O5	167.05501	C5H10O6
157.15869	C10H20O	223.09648	C12H14O4	167.07027	C9H10O3
164.03422	C8H5NO3	233.17474	C10H20O5	177.12337	C7H16N2O3
167.07027	C9H10O3	237.18491	C15H24O2	182.04478	C8H7NO4
167.10666	C10H14O2	279.15909	C16H22O4	182.08117	C9H11NO3
170.99578	C3H6O6S	285.24242	C17H32O3	182.10498	C9H13N2O2
171.02227	C6H6N2O2S	303.21660	C16H30O5	182.97803	C4H6O4S2
176.10297	C6H13N3O3	305.23225	C16H32O5	183.02880	C8H6O5
189.14852	C10H20O3	311.08738	C13H14N2O7	195.06518	C10H10O4
194.15394	C12H19NO	315.25299	C18H34O4	195.08044	C14H10O
201.09101	C13H12O2	343.12349	C12H22O11	195.08631	C7H14O6
201.10224	C12H12N2O	343.28429	C20H38O4	205.09715	C11H12N2O2
203.03388	C11H6O4	371.31559	C22H42O4	209.18999	C14H24O
203.10263	C8H14N2O4	381.37271	C25H48O2	213.03936	C9H8O6
203.10666	C13H14O2	381.40909	C26H52O	218.15731	C11H23NOS
203.12779	C10H18O4	387.34689	C23H46O4	229.14344	C12H20O4

90 | Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter rubber*

203.24817	C12H30N2	463.45096	C31H58O2	230.13868	C11H19NO4
204.06552	C11H9NO3	473.16535	C21H28O12	239.12779	C13H18O4
204.08665	C8H13NO5	493.49791	C33H64O2	243.19547	C14H26O3
204.08799	C9H9N5O	497.45644	C31H60O4	243.23186	C15H30O2
204.10190	C12H13NO2	511.47209	C32H62O4	255.06453	C7H14N2O6S
204.11314	C11H13N3O			263.16417	C16H22O3
205.11828	C8H16N2O4			265.10705	C14H16O5
211.11174	C15H14O			269.21112	C16H28O3
211.16926	C13H22O2			273.18491	C18H24O2
213.07575	C10H12O5			276.07273	C11H9N5O4
213.09101	C14H12O2			279.19547	C17H26O3
215.05501	C9H10O6			279.23186	C18H30O2
217.14344	C11H20O4			291.14383	C13H22O7
221.18999	C15H24O			299.20056	C20H26O2
223.07536	C15H10O2			301.27372	C18H36O3
223.08123	C8H14O7			303.12270	C17H18O5
223.15400	C10H22O5			303.19547	C19H26O3
225.08698	C10H12N2O4			317.21112	C20H28O3
228.19581	C13H25NO2			319.21151	C16H30O6
233.18999	C16H24O			325.16456	C17H24O6
241.14344	C13H20O4			327.21660	C18H30O5
245.12845	C14H16N2O2			327.23186	C22H30O2
257.21112	C15H28O3			327.25299	C19H34O4
262.11006	C17H13N2O			337.18319	C19H28O3S
263.11789	C17H14N2O			337.27372	C21H36O3
263.23694	C18H30O			338.19620	C18H27NO5
285.13326	C14H20O6			343.22677	C22H30O3
285.20604	C16H28O4			354.13697	C17H23NO5S
286.13724	C16H19N3S			357.16965	C21H24O5
293.21112	C18H28O3			359.25807	C23H34O3
297.24242	C18H32O3			359.33084	C25H42O
297.31519	C20H40O			365.24751	C25H32O2
299.29446	C19H38O2			375.25299	C23H34O4
301.06012	C10H12N4O5S			377.30502	C24H40O3
301.10705	C17H16O5			379.24790	C22H34O5
303.23186	C20H30O2			383.18867	C20H30O5S
307.22677	C19H30O3			399.36214	C28H46O
309.24242	C19H32O3			401.18597	C25H24N2O3
309.31519	C21H40O			405.32107	C22H44O6
313.16724	C19H22NO3			409.33124	C25H44O4
317.24751	C21H32O2			411.25299	C26H34O4

91 | Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter rubber*

319.19039	C19H26O4	431.29446	C30H38O2
321.11801	C13H20O9	437.32615	C26H44O5
324.20704	C20H25N3O	439.28429	C28H38O4
331.26316	C22H34O2	453.29994	C29H40O4
334.16240	C22H23NS	460.23432	C26H29N5O3
335.22169	C20H30O4	473.15010	C17H28O15
338.18112	C21H25IN2S	475.45096	C32H58O2
338.18630	C20H23N3O2	481.29485	C30H40O5
339.32576	C22H42O2	501.43022	C33H56O3
343.17513	C17H26O7	515.44587	C34H58O3
347.25807	C22H34O3	524.50372	C33H65NO3
353.26864	C21H36O4		
355.32067	C22H42O3		
359.29446	C24H38O2		
359.31559	C21H42O4		
360.12238	C14H21N3O6S		
361.16456	C20H24O6		
365.34141	C24H44O2		
367.32067	C23H42O3		
371.25807	C24H34O3		
375.20134	C18H30O8		
376.16557	C22H21N3O3		
385.27372	C25H36O3		
385.29485	C22H40O5		
387.15909	C25H22O4		
387.18021	C22H26O6		
403.17513	C22H26O7		
403.23398	C21H30N4O4		
403.35706	C27H46O2		
405.26355	C24H36O5		
405.37271	C27H48O2		
407.31559	C25H42O4		
421.25846	C24H36O6		
425.43531	C28H56O2		
429.29994	C27H40O4		
429.40909	C30H52O		
431.33084	C31H42O		
431.35197	C28H46O3		
433.33124	C27H44O4		
435.31050	C26H42O5		
435.32576	C30H42O2		

441.32107	C25H44O6		
443.35197	C29H46O3		
450.26388	C28H35NO4		
451.48734	C31H62O		
455.31559	C29H42O4		
455.38836	C31H50O2		
457.24321	C23H36O9		
467.44587	C30H58O3		
481.46152	C31H60O3		
483.33163	C27H46O7		
485.34728	C27H48O7		
489.22716	C30H32O6		
489.28468	C28H40O7		
501.32107	C30H44O6		
505.22208	C30H32O7		
511.41457	C34H54O3		
513.43022	C34H56O3		

Table 4.5: List of discriminative m/z values and their corresponding metabolite elementary compositions  $[M+H]^+$  calculated with a tolerance of 1 ppm. These were compared to the available databases ([www.metabolome.jp](http://www.metabolome.jp) and KEGG).

Thus, those components generally associated with cell membranes, such as fatty acids and terpenoids, could be responsible for the geographic differentiation. Among the two principal components that led to geographical discrimination, we could also find, for the second component, a relationship for geographical distance between the isolation sites (table 4.3 and figure 4.3). It seemed that for the second principal component the Atlantic strains might show intermediate differences with those of the Mediterranean and Peruvian strains.

When specifically directing the recognition of discriminative metabolites among geographically distinct metabolomes, we observed that a set of conspicuous compounds could be unambiguously assigned to a sulfonolipid family (table 4.6 and figure 4.2).

Exp. m/z	Proposed composition as [M-H] <sup>-</sup> $\Delta m/z < 0.6$ ppm	Structural variation from C <sub>35</sub> H <sub>66</sub> NO <sub>8</sub> S	Mediterranean (n=6) averaged intensity (x 10 <sup>6</sup> )	Atlantic (n=11) averaged intensity (x 10 <sup>6</sup> )	Peruvian (n=5) averaged intensity (x 10 <sup>6</sup> )
644.4195	C <sub>34</sub> H <sub>62</sub> NO <sub>8</sub> S	- CH <sub>2</sub> & - H <sub>2</sub>	1.27	1.14	1.37
646.4351	C <sub>34</sub> H <sub>64</sub> NO <sub>8</sub> S	- CH <sub>2</sub>	6.40	5.12	5.88
660.4505	C <sub>35</sub> H <sub>66</sub> NO <sub>8</sub> S	(-)	67.20	76.60	<b>92.24</b>
672.4505	C <sub>36</sub> H <sub>66</sub> NO <sub>8</sub> S	+ C	6.42	5.88	7.10
674.4662	C <sub>36</sub> H <sub>68</sub> NO <sub>8</sub> S	+ CH <sub>2</sub>	8.04	6.98	8.88
676.4454	C <sub>35</sub> H <sub>66</sub> NO <sub>9</sub> S	+ O	n.d.	<b>1.50</b>	n.d.
684.4508	C <sub>37</sub> H <sub>66</sub> NO <sub>8</sub> S	+ 2C	0.30	0.24	0.26
686.4663	C <sub>37</sub> H <sub>68</sub> NO <sub>8</sub> S	+ C <sub>2</sub> H <sub>2</sub>	1.44	1.25	1.42
688.4455	C <sub>36</sub> H <sub>66</sub> NO <sub>9</sub> S	+ C & + O	n.d.	<b>0.40</b>	n.d.
688.4819	C <sub>37</sub> H <sub>70</sub> NO <sub>8</sub> S	+ 2CH <sub>2</sub>	0.95	0.87	0.97

Table 4.6: Proposed elemental compositions of various masses assigned to sulfonolipids with their structural variations from C<sub>35</sub>H<sub>67</sub>NO<sub>8</sub>S, originally described by Corcelli, (Corcelli, et al., 2004) as C<sub>35</sub>H<sub>66</sub>NO<sub>8</sub>S, where n indicates the number of strains.

The members of this compound family have been observed to be major components of the cell envelope of *Cytophaga* (Godchaux, et al., 1984), a member of the same phylum as *S. ruber* (Antón, et al., 2002). One of these components (C<sub>35</sub>H<sub>67</sub>NO<sub>8</sub>S, m/z = 660.4505) has been reported to be characteristic of *S. ruber* (Corcelli, et al., 2004). These compounds, which could account for 10% of total cellular lipids, have been proposed as signatures for *S. ruber* identification. The ICR-FT/MS approach, with a mass precision lower than 600 ppb, revealed that *S. ruber* may contain at least nine additional sulfonolipids analogous to C<sub>35</sub>H<sub>67</sub>NO<sub>8</sub>S in the mass range 644 to 688. These components differ from the originally described sulfonolipid in their elementary composition, with variations in their side chain length, insaturation or hydroxylation degree with variations in CH<sub>2</sub>, H<sub>2</sub> and O,

respectively, as described in Table 4.6. All these components were found in all of the analyzed samples with identical intensity ratios between isolates from the same location, except for  $m/z$  676.4454 and  $m/z$  688.4455.

Both of these compounds (C<sub>35</sub>H<sub>68</sub>N<sub>9</sub>O<sub>9</sub>S and C<sub>36</sub>H<sub>68</sub>N<sub>9</sub>O<sub>9</sub>S, respectively) seemed to be exclusive to the Atlantic strains.

The metabolomic approach allowed the targeted search for special metabolic traits considered to be relevant in the organisms' phenotype. Previous biochemical studies on *S. ruber* type strain M31 revealed the presence of an active, hitherto unreported, rhodopsin type of membrane proton translocation system, the xanthorhodopsin, responsible for the putative phototrophy of *S. ruber* (Balashov, et al., 2005). In addition, the genome sequence of the same organism revealed the coding of one halorhodopsin (Peña, et al., 2005) and two sensory rhodopsin homologous genes (Mongodin, et al., 2005). Searching for an indication of the presence of retinal, the chromophore bound to rhodopsins, an experimental positive mass 285.22125 (theoretical 285.22129) was present in all samples. However, the  $m/z$  value was only discriminative for the Mediterranean strains.

An independent “fine tuning” experiment was undertaken by growing four replicates of 5 Mediterranean strains (P13 and P18 from Alicante, M8 and M31 from Mallorca and IL3 from Ibiza). Metabolome comparisons validated the replicates by first applying a Levene's test (Malins, et al., 2002) to evaluate differences in the variance, and after applying analysis of variance (one-way ANOVA) and the Tukey test to evaluate the differences in the means of each replicate group. Nevertheless, the results between both latter tests were equivalent. At the  $p < 0.01$  level, population variations were not significantly different (see table 4.4 for the ANOVA results). Therefore, the differences observed between different strains could be attributed to strain-specific metabolisms rather than sample to sample variations. In contrast to previous results (Peña, et al., 2005), when searching for discriminative phenotypes at a more reduced geographical scale, we observed a phenotypic segregation in individual locations (see figure 4.4a), using the ICR-FT/MS approach. The main discriminative metabolomic profile features were different from those giving resolution at a larger geographical scale. In such cases, geographical differences were associated to strain specific compositions of N-containing molecules (see figure 4.4b). The confrontation of their exact masses with the Kyoto Encyclopaedia of Genes and Genomes (KEGG) and the Japanese Metabolome Database (metabolome.jp), indicated that the discriminative molecules were involved primarily in the core metabolism (i.e. carbohydrate, amino acid and fatty acid biosynthesis and metabolism).



## 4.6 Conclusions

Our findings reveal that intraspecific metabolic diversity of *S. ruber* can be readily detected by the ICR-FT/MS approach and that such diversity can be associated with different geographical patterns at different metabolic levels. In principle, the standard genetic methods used to assess biogeography (Ramette, et al., 2006), (Whitaker, et al., 2003) do not have the resolving power needed for a fine geographic discrimination of our model organism. The MLSA approach, based on different gene datasets, does not resolve putative genetic-geographic patterns, as the genetic divergence may be too subtle for the given selection of genes. However, one must take into account that, despite the fact that large sets of concatenated genes tend to reflect the organismal phylogeny (Sória-Carrasco, et al., 2007), perhaps only full genome sequences may reflect geographical isolation in the strain collection of *S. ruber*, in accordance with taxa segregation that correlates with the average nucleotide or amino acid identity of shared genes (Konstantinidis, et al., 2005). However, the backlogs in the current state of full genome sequencing makes the metabolomic approach a fast and less expensive alternative for revealing prokaryotic biogeography, with the added value of being discriminative at different levels at the geographical scale.

It seems clear that different locations led to the isolation of strains sharing common metabolic traits, such as, for instance, the distinct production of sulfonolipid derivatives. However, differences were generally related to quantitative composition yields, rather than qualitative production of distinct compounds. In addition, the metabolic differences correlated with the geographical locations, influenced perhaps by environmental conditions such as climate and distance, since in the second component Peruvian and Mediterranean strains were shown to be the most different. The discriminative metabolites were mainly aliphatic structures related to terpenoids or fatty acids, which might be membrane components and these differences, could be related to different environmental conditions (Sajbidor, 1997). Altogether, the results seem to indicate that the differences found could be attributed to transcriptional or posttranscriptional regulations rather than composition changes in genes at the genomic level. The major forces for these differences between strains should be related to their distinct response to the environmental conditions of the sites where they had been isolated, since, for example, the Peruvian salterns are not only over 10,000 km away from the rest of our sampling sites, but they are also at

an altitude of 3,500 m. At this site the temperature changes and solar radiation are clearly different from those at sea level. ICR-FT/MS was shown to have a higher resolution in revealing fine scale diversity. This method has a great potential for revealing biogeographical patterns in many other non extremophilic microorganisms.

# Chapter 5

## 5 EXPRESSING FOREST ORIGINS IN THE CHEMICAL COMPOSITION OF COOPERAGE OAK WOODS AND CORRESPONDING WINES BY ICR-FT/MS

### 5.1 Introduction

Here, we report the first non targeted chemical characterization approach using organic structural spectroscopy/spectrometry, based primarily upon ultra high resolution ICR-FT/MS analysis (Hertkorn, et al., 2007) of cooperage oak wood extracts and related wines, with the aim of drawing comprehensive "chemical pictures", which would allow to establish significant correlations between these samples. It must be noted that recently, a similar non-targeted approach, based on the "electronic tongue" analysis, has been able to nicely discriminate wines

with respect to the origin of oak barrels they were aged in (Parra, et al., 2006). However, these discriminations were only based on the high cross-selectivity of voltammeter sensors, and provided no structural information on any active molecules involved on a molecular level.

The main goal of this study is to identify families of metabolites that could discriminate both the species and the geographical origin of woods. Based on 12 Tesla ICR-FT/MS of wood extracts, hundreds of mass peaks were identified as possible significant biomarkers of the two species, with phenolic and carbohydrate moieties leading the differentiation between two wood species (*Quercus robur* L. and *Quercus petraea*) as corroborated by both FTMS and NMR data. For the first time, it is shown that oak woods can also be discriminated on the basis of hundreds of forest-related compounds, with a particular emphasis on sessile oaks from the Tronçais forest, for which hexoses are significantly discriminant. Despite the higher complexity and diversity of wine metabolites, forest-related compounds can also be detected in a wine aged in related barrels.

Initially aimed at serving as suitable wine containers, oak barrels have today become practical means of modulating fine sensory characteristics of wine (Garde-Cerdan, et al., 2006). Several studies have revealed the influence of oak wood on the organoleptic properties of wines matured in oak barrels (Waterhouse, et al., 1994), (Jarauta, et al., 2005). This influence is considered to be due to the variation of physical and chemical properties of oak, which mainly depend on both the geographical origin and the species (Doussot, et al., 2000), (Doussot, et al., 2002). So far, attempts to establish correlations between oak wood chemical properties and origin or species have relied on targeted analyses of selected compounds. These studies particularly revealed significant species effects: for instance, it is recognized that among the two predominant west European oak species, *Quercus robur* L. (pedunculata oak) exhibits larger ring widths and contains more ellagitannins than *Quercus petraea* Liebl. (sessile oak), which in contrast generally contains more volatile compounds, such as cis- and trans- $\beta$ -methyl- $\gamma$ -octalactones (whisky-lactones), eugenol, vanillin or furfural, although discrepancies can be found in the literature (Doussot, et al., 2000), (Chatonnet, et al., 1998). A similar trend, but restricted to ellagitanins and whisky lactones, has been generalized to east European pedunculate and sessile oaks (Prida, et al., 2006). When considering more specifically aromatic whisky-lactones, French sessile oaks are generally poorer than American white oaks (*Quercus Alba*) and east European sessile oaks (Prida, et al., 2006), (Towey, et al., 1996). Besides the species effect, effects that forests could impose upon the chemical composition of

oak wood and ultimately on wines, have also been investigated (Waterhouse, et al., 1994), (Mosedale, et al., 1999), (Dousot, et al., 2000), (Dousot, et al., 2002), (Mosedale, et al., 1996), (Cadahia, et al., 2001). These studies showed that forest effects on the chemical composition of wood are less pronounced than species effects and significant discriminations, regardless of the species, could only be made between American, west and east European forests, on the basis of their eugenol, 2-phenylethanol, vanillin and syringaldehyde contents (Prida, et al., 2006). A huge inter-individual variability of the chemical composition of oak trees, even within a given species in a given forest, is actually the major acknowledged reason for the current absence of established significant correlations relating a forest and its oak woods composition, regardless of the species and location (Dousot, et al., 2000), (Dousot, et al., 2002), (Guchu, et al., 2006), (Feuillat, 2003).

When considering the further chemical composition correlations that can be made between the geographical origin or the species of oaks and wines matured in related barrels, the only acknowledged generalization is that the American white oak species provides higher amounts of cis- whiskylactone to wines than the European sessile oak species (Garde-Cerdan, et al., 2006), (Waterhouse, et al., 1994). The cis- whiskylactone is often mentioned as the major discriminant compound, because its content in wood correlates well with its content in wines aged in respective oak barrels and also with the coconut, toasty or vanilla sensory descriptors of these wines (Sauvageot, et al., 1999). In contrast, despite the abundance of heartwood ellagitanins and their solubility in wines, the concentration in oak-aged wines is generally lower than expected (Puech, et al., 1999). Therefore, in terms of chemical composition, no unambiguous forest effects of general validity on wines have been reported yet, and effects on the chemical composition of cooperage oak woods have heavily relied upon the species-based identification of natural forests (Dousot, et al., 2000).

Consequently correlations between a forest classification and the wine aged in a barrel made of oaks from this forest are at best feeble. In addition, a multi-stage process operates between the cutting of oaks and the end of the barrel ageing period of wines. First, wood staves undergo natural seasoning and then toasting, designed to shape barrels. Both of these processes contribute to modulate the chemical composition of wood (Dousot, et al., 2002), (Cadahia, et al., 2001), (Chatonnet, et al., 1989) and subsequently of wine (Hale, et al., 1999), (Spillman, et al., 2004). However, although heating does form new compounds as a result of lignin and cellulose degradation, many heartwood constituents are

barely or not affected by the heating intensity normally used, and instead of eliminating the intrinsic variation between wood samples, heating would rather appear to complement it (Mosedale, et al., 1999). Second, several concurrent processes do take place during the ageing period of wine (Garde-Cerdan, et al., 2006), (Jarauta, et al., 2005), (Chassagne, et al., 2005), (Barrera-Garcia, et al., 2007). This has been recently illustrated by Jarauta et al. (Jarauta, et al., 2005), who have identified at least seven processes responsible for the evolution of the 79 aroma compounds analyzed in wines aged in oak barrels. These authors have confirmed that, in addition to the most studied extraction processes from the barrel, microbiological transformation, weak oxidation reactions enabled by the porosity of this container, condensation reactions and sorption to wood, also modulate wine compounds during barrel aging. Another example of the complex mechanisms involved in wine chemistry related to barrel ageing has been provided by Quideau et al., who highlighted the fact that many ellagitannin derivatives would probably exist as a result of nucleophilic substitution reactions with wine relevant nucleophiles (Quideau, et al., 2003), (Quideau, et al., 2005).

All these studies have fundamentally contributed to the knowledge of the chemical composition of oak wood related to its species and to a lesser extent to its origin, and also to its impact on the composition and flavor of barrel aged wine. However, as shown by Jarauta et al. (Jarauta, et al., 2005), most studies have failed to consider oak casks as a physical, chemical and biochemical active system. Oak wood itself is already a complex living system for which environmental conditions, such as the forest ecosystem where it has grown, may modulate its chemical composition as extensively as genetic diversity between species; genetic analyses have actually shown rather minor differentiation between the two species (*Quercus robur* L. versus *Quercus petraea* L.) (Mosedale, et al., 1999), (Curtu, et al., 2007).

In 1998, a full-scale integrated study ("Tonnellerie 2000") initially involving nine French forests providing twelve lots of 24 trees (5 lots of pedunculate and 7 lots of sessile oak), was designed to evaluate the influence of both the geographic origin and the species of oak on the quality of wines matured in oak barrels (Feuillat, et al., 1999). We hypothesized that such sets of wood and wine samples would become unique panels of chemical compositions with little variations, and as such, ideal candidates for a non-targeted analysis of the correlations that could possibly exist among the species and/or the forest origin of oak wood and the wine aged in barrels made of this wood. This chapter is based on the article:

“Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines in ICR-FT/MS” (in review, Chemistry European Journal, 2008).

## 5.2 Wood samples collection

The "Tonnellerie 2000" experiment (Feuillat, et al., 1999) has been designed to particularly take into account the high interindividual variability which had already been observed even between trees from a same forest. Therefore, the selected procedure was based on the combination of lots of trees considered as representative of one species from one forest.

The detailed procedure followed to select trees has already been described elsewhere (Feuillat, et al., 1999). In brief, twelve lots (5 pedunculate and 7 sessile) of 24 trees were selected from nine French forests. During the cutting of trees, a disk was cut at a one-meter height up the bole of each tree, for further analyses. From each disk, a radial strip (oriented along the diameter), centered around the outer part of heartwood was kept. For this study, we only considered the three forests where both the pedunculate (P) and the sessile (S) species were represented, i.e. Citeaux (C), Darney (D) and Tronçais (T). Therefore, we had six lots of 24 strips at our disposal, which had been stored in plastic boxes in the basement of our university building without any further care. After a careful examination of the 144 strips, 4 strips per lot, which showed visual traces of mould, were excluded. For our study, we therefore had 6 lots of 20 wood samples (120 samples), each corresponding to one species from one geographical origin. It must be noted that laboratory morphological analyses realized later after the original identification on standing trees, revealed that errors had been made on the assignment of species from the Darney forest: 3 out of the 20 pedunculate oaks were actually sessile oaks, and conversely, 4 out of the 20 sessile oaks were actually pedunculate oaks (Feuillat, et al., 1999). Our sets of sawdust samples were prepared regardless of these errors, meaning for instance that 15 % of the Darney pedunculate set actually corresponds to sessile species.

### 5.2.1 Barrels and wine elaboration

To one lot of 24 trees corresponded one barrel. Each barrel has thus been assembled from 24 trees which stood each for  $1/24^{\text{th}}$  of the toasted surface (body) and  $1/24^{\text{th}}$  of the untoasted surface (head and bottom). After one year of natural seasoning of staves, 48 barrels (12 lots x 4 repeats) were assembled and subsequently medium toasted for 45 minutes.

A first set of two experiments was designed during the 1998 harvest; one with the appellation "Mercurey rouge 1er cru" with "Pinot noir" variety from Domaine Michel Juillot (12 lots x 2 repeats + 1 reference stainless steel tank), and the other with the appellation "Beaune blanc 1er cru" with "Chardonnay" variety from Maison Bouchard Père et Fils (12 lots x 2 repeats + 1 reference stainless steel tank). At the end of the wine ageing period (12 months for the red, and 14 months for the white), bottling has been realised after blending of the two repeats for each lot, thus providing us with 13 bottles of Mercurey and 13 bottles of Beaune.

### 5.2.2 Wood and wine samples preparation

On each of the 120 wood samples, the outer duramen zone has been planed at different locations to obtain few millimeter-thick coarse shavings. Hence, each lot was made of 20 sets of wood shavings equally represented and mixed together. The 6 lots of wood shavings thus obtained were then ground to powders of less than  $250\ \mu\text{m}$  granulometry.

20 mg of each sawdust sample were then extracted with 1 ml ethanol/water solution (8:2 v/v) at room temperature for 30 minutes in an ultrasonic bath. Each of the six mixtures was then centrifuged (10 mn, 18000 rpm) and further filtrated on  $0.2\ \mu\text{m}$  filters. Three repetitions were realized for each of the six lots, which provided us with 18 hydroalcoholic extracts. Although the hydroalcoholic solution does not necessarily exhibit the best extracting efficiency for non-volatile compounds, we chose it to minimise the preparation steps prior to the injection to the mass spectrometer. For NMR analysis, deuterated solvent was used for extraction and the ethanol extract was analysed after centrifugation.

Wine was sampled directly from the bottles through the cork using a Hamilton needle. Only  $20\ \mu\text{L}$  of wine was diluted into 1 mL methanol from which only  $50\ \mu\text{L}$  was used for one experiment (i.e. only a total aliquot of  $2\ \mu\text{L}$  wine was necessary to reach the spectral quality presented herein).



### 5.3 ICR-FT/MS analysis

High-resolution mass spectra for molecular formula assignment were acquired on a Bruker (Bremen, Germany) APEX Qe Fourier transform ion cyclotron resonance mass spectrometer (ICR-FT/MS) equipped with a 12 Tesla superconducting magnet and a APOLO II ESI source in the negative ionisation mode. Samples were introduced into the microelectrospray source at a flow rate of 120  $\mu\text{l}/\text{h}$  with a nebuliser gas pressure of 20 psi and a drying gas pressure of 15 psi (200 °C). Spectra were externally calibrated on clusters of arginine (10mg/l in methanol) and accuracy reached values lower than 0.1 ppm in day to day measurements. Further internal calibration was done for each sample using fatty acids and accuracy reached values lower than 0.05 ppm. The spectra were acquired with a time domain of 1 MW with a mass range of 100-2000 m/z. The spectra were zero filled to a processing size of 2 MW and an average resolution of 250.000 was reached at m/z 200 (100.000 at respectively m/z 600) in full scan. Before Fourier transformation of the time-domain transient, a sine apodization was performed. The ion accumulation time in the ion source was set to 0.2 s for each scan. 1024 scans were accumulated for samples.

### 5.4 NMR Spectroscopy

All experiments in this study were performed with a Bruker DMX 500 spectrometer and a  $^{13}\text{C}/^1\text{H}$  dual 5 mm cryogenic probe at 283 K on forest-consolidated wood samples from both species dissolved in 184 mg 99.95%  $^2\text{H}$   $\text{CD}_3\text{OD}$  (reference for  $^1\text{H}/^{13}\text{C}$  -NMR was 3.30/49 ppm;  $(90^\circ(^1\text{H})) = 10.1 \mu\text{s}$ ;  $90^\circ(^{13}\text{C}) = 10.0 \mu\text{s}$ ). 1D  $^1\text{H}$ -NMR spectra were also recorded for each of the six lots of hydroalcoholic solutions. 1D  $^1\text{H}$ -NMR spectra were recorded using the first increment of the presat-NOESY sequence (solvent suppression with presaturation and spin-lock, 5.0 s acquisition time, 10.0 s relaxation delay, 320 scans, 1 ms mixing time, 1 Hz exponential line broadening).  $^{13}\text{C}$ -NMR spectra were acquired, using inverse gated WALTZ-16 decoupling (13.75 s relaxation delay; 42153 scans for  $^{13}\text{C}$  NMR, 75821 scans for DEPT-135 and 32768 for DEPT-90) with an acquisition time of 1.25 s and an exponential line broadening of 1.5 Hz.

The one bond coupling constant  $^1J(\text{CH})$  used in 1-D  $^{13}\text{C}$  DEPT and proton-detected 2D NMR spectra was set to 150 Hz. Sensitivity-enhanced, carbon decoupled  $^1\text{H}$ ,  $^{13}\text{C}$ -HSQC (heteronuclear single quantum coherence) NMR spectra were acquired under the following conditions:  $^{13}\text{C}$ -90-deg decoupling pulse, GARP (70 $\mu\text{s}$ ); F2 ( $^1\text{H}$ ): acquisition time: 291 ms at spectral width of 6009 Hz,  $^1J(\text{CH}) = 150$  Hz, 1.21 s relaxation delay; F1 ( $^{13}\text{C}$ ): SW = 22009 Hz (175 ppm); number of scans (F2)/F1-increments ( $^{13}\text{C}$  frequency) for  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC experiments: 144/800; for absolute value;  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC (heteronuclear multiple bond correlation): 320/270;  $^1\text{H}$ ,  $^1\text{H}$  COSY (correlated spectroscopy): 64/1056;  $^1\text{H}$ ,  $^1\text{H}$  TOCSY (total correlated spectroscopy) (70 msec mixing time): 64/938;  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC-TOCSY: 160/513 (70 msec mixing time), respectively. HSQC and DEPT-HSQC spectra were calculated to a 2048 x 512 matrix with exponential line broadening of 2 Hz in F2 and a shifted sine bell ( $\pi/3$ ) in F1. Gradient sequences (1 ms length, 450  $\mu\text{s}$  recovery) were used for all proton detected spectra.

#### 5.4.1 Analysis of NMR spectra

NMR integrals were measured manually from printed spectra. Bucket analysis (Brindle, et al., 2002) was performed on the experimental  $^{13}\text{C}$  NMR spectra of six wood extracts; these were decomposed into 87 equidistant integral segments with 0.1 ppm bandwidth, ranging from 0.4 - 8.1 ppm.

### 5.5 Statistical analysis

Raw data (mass spectra) were normalized, and then transformed to  $\log(X + 0.00001)$ . The constant 0.00001 was added to provide non-detectable components with a small non zero value (Sjödin, et al., 1989). Transformed variables were then mean centered and Pareto scaled and represented as an X matrix. Pareto scaling gives each variable a variance equal to its standard deviation by dividing by the square root of the standard deviation of each column (Eriksson, et al., 2001). The sample classification and the prior information about the sample are done using the Hierarchical clustering analysis (HCA) unsupervised method. On the other hand, partial least square - discriminative analysis (PLS-DA), performed with SIMCA

11.5, is used to discover characteristic biomarkers (Quideau, et al., 1996). This multivariate procedure provides bioinformatics clues for the selection of a limited number of masses most effective in discriminating different species and forests.

The primary advantage of using targeted profiling as an input to PLS-DA is that the resulting variables are combinations of measured metabolites concentrations. The positive regression coefficient indicates that there is a relatively greater concentration of the considered metabolites with respect to the others, whereas the negative value indicates a relatively lower concentration with respect to the other samples-classes (Herve du Penhoat, et al., 1991). As such, these variables are easier to interpret as factors in the underlying classification model. Thus, targeted profiling provides meaningful and interpretable factors describing the input data. PLS-DA is a regression extension of PCA that takes advantage of class information to attempt to maximize the separation between groups of observations.

The feature selection procedure comprises two steps: i) identification of those masses that best describe each classes (a list based on the modeling power of the original variables), ii) scoring and ranking of the variables in every class-related list according to their abilities to discriminate the class they model from all other categories. The ranking and score take place after computation of the minimum number of masses through the formula generator (in-house code written in FORTRAN). The generated formulas were validated by setting sensible chemical constraints (N rule, O/C ratio  $\leq 1$ , H/C ratio  $\leq 2n + 2$ , element counts: C  $\leq 100$ , O  $\leq 80$ , N  $\leq 5$ , S  $\leq 1$ ) and only the masses in conjunction with their automated generated theoretical isotope patterns were taken into consideration.

## 5.6 Result and discussion

### 5.6.1 Wood differentiation

Figure 2.4 (chapter 2) shows an exemplary full mass spectrum of forest-averaged oak wood extracts for the two species. Within the 150-1000 m/z range explored, these spectra exhibit several thousands of peaks, which represent all ionisable metabolites under the selected experimental conditions (electrospray negative mode). Although hydroalcoholic solutions do not necessarily provide the

best extracting efficiencies for non-volatile compounds, 5727 distinct peaks were observed at  $S/N = 1$  for the P species, of which 1045 that could be assigned elementary formula containing CHONS. Similarly, 7677 resolved peaks are observed for the S species with 1562 assignments of elementary formula. A cumulative total of 8354 different peaks and 1797 assignments of detected non-identical molecular formula indicate the occurrence of both common and divergent molecules for P and S species.

Hierarchical cluster analysis (HCA) readily identifies two major groups (see figure 5.1), and shows more uniformity among P oak samples than among S samples. Clearly, a correct classification of each of the six sets of three repetitions is available upon their negative ion mode ICR-FT/MS in order to assess similarity/dissimilarity. The modi employed and the choice of linkage methods used for clustering greatly affects the numerical outcome of the HCA results. Following careful examinations of available similarity/dissimilarity assessments, Pearson correlation coefficient distance (straight line distance between two points in  $c$ -dimensional space defined by  $c$  variables) as similarity descriptor in conjunction with the complete linkage method, were found to produce the most distinctive grouping, in which each member within the group is more similar to its fellow members than to any member from outside the group. This is a confirmation that the complete linkage method performs quite well in cases where object form naturally distinct “clumps” (Taylor, et al., 2000).

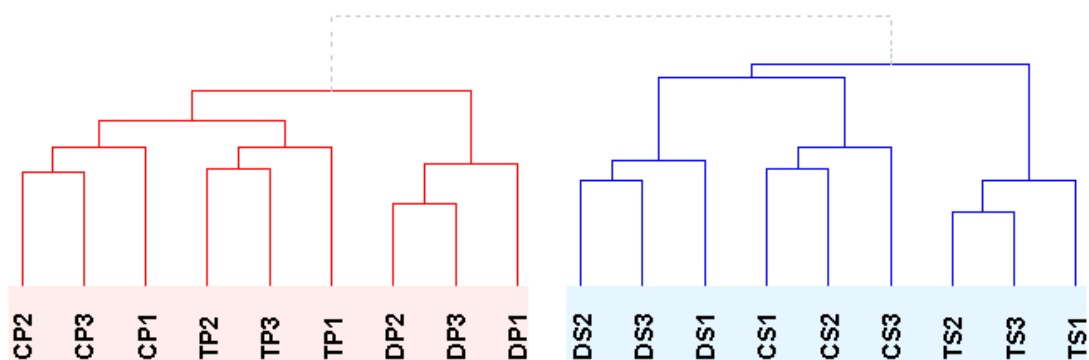


Figure 5.1: Dendrogramm for HCA for classifying 18 sets (3 times 6) of samples (Minimum similarity=0.44).

HCA does not provide a statistical test of group dissimilarity; however external tests like the Kolmogorov-Smirnov test can be applied for this purpose. This elaboration was done with SAS version 9.1 (SAS Institute Inc., Cary, NC, USA) with the Kolmogorov-Smirnov hypothesis that two groups of observations have identical distributions. With this test the difference among S subgroups defined by the HCA (see figure 5.1) were determined to be statistically significant for DS and CS subgroups ( $p < 0.0001$ ) and for TS and CS ( $p < 0.0001$ ), whereas for subgroups DS and TS, the asymptotic p-value (0.0052) indicates rejection of the null hypothesis that the distributions were identical also for these two subgroups. The PLS-DA score plot of wood species (see figure 5.2) provides a representation of how forests from a given species are grouped together. The two predictive components of the PLS-DA model,  $R^2(Y)$ : 99% and the prediction accuracy  $Q^2(\text{cum})$ : 0.96 were obtained through a typical seven-fold cross-validation and guaranteed that this model is satisfactory. In agreement with the cluster analysis (see figure 5.1), P oak samples exhibit a narrower distribution between the three forests than S oaks. These findings corroborate the previously observed higher inter-individual variability in whiskey-lactone contents among S oaks in comparison with P oaks in which only traces have been found (Feuillat, 2003).

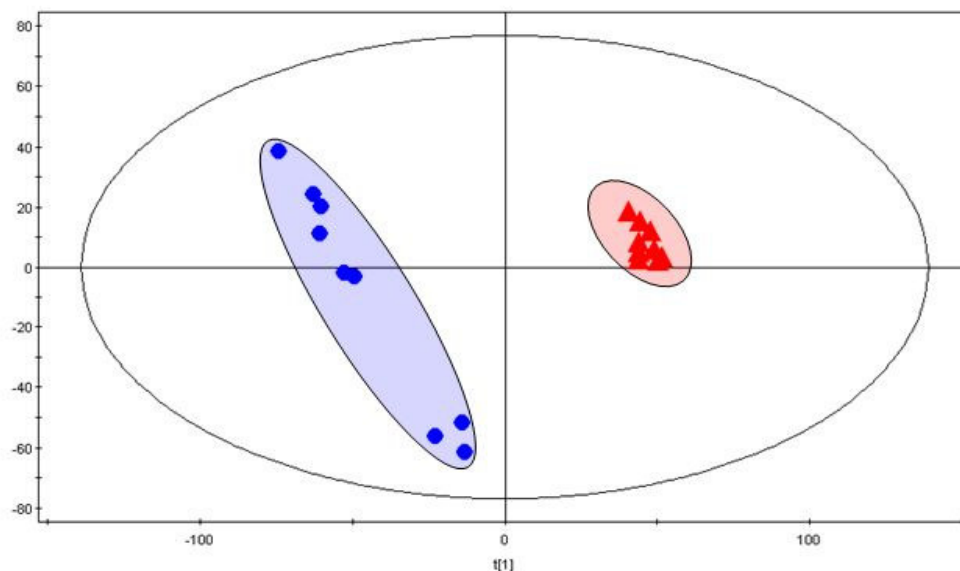


Figure 5.2: Partial least square - discriminant analysis (PLS-DA) score plot for the first two components indicating the separation between the two species (P: red filled triangles; S: blue filled circles).

## 5.6.2 The species effect

The mass spectral peaks ( $m/z$  values) that drive the differentiation between species are extracted according to exceeding given positive values of the regression coefficient. For the S group, 159 mass peaks with a regression coefficient greater than 0.001, are considered possible significant biomarkers; in case of P, 207 mass peaks with a regression coefficient in excess of 0.0004 were selected.

The selected mass range of 334.95-335.30  $m/z$  (see figure 5.3) illustrates the remarkable resolution of the 12 Tesla ICR-FT/MS. Within this nominal mass, more than a dozen resolved peaks are identified in the S and P samples from the Citeaux forest. Only at this extent of resolution, clear and unambiguous differentiation of species is feasible.

Within the frame of a full scale metabolomic approach, both the mass resolving power and the mass accuracy should be precise enough to enable an unambiguous identification of the elemental compositions at a same nominal mass. Even if these two conditions would appear to be fulfilled for most of the mass peaks in this study, the lack of experimental databases and the chemical complexity of wood would make the task of identifying all of the corresponding molecules extremely tedious. Yet, in particular cases, tentative assignment of peaks to known wood-related compounds is feasible without the need of other analytical tools. As illustrated by the peak at  $m/z$  335.17114 (averaged value), found only in the mass spectra of S oaks (see figures 5.3a and b), the corresponding  $[M-H]^-$  ion  $C_{15}H_{28}O_3$  can most likely be assigned from literature data to 3-methyl-4-hydroxyoctanoic acid 3-O- $\beta$ -D-glucopyranoside, a common precursor of whisky-lactone (Masson, et al., 2000), (Hayasaka, et al., 2007). This attribution is acknowledged by the higher contents of both the trans and the cis isomers of whiskylactone in S oaks (Masson, et al., 1995). Therefore, along with whiskylactone, its precursor can logically be considered as a bio-marker of S oaks.

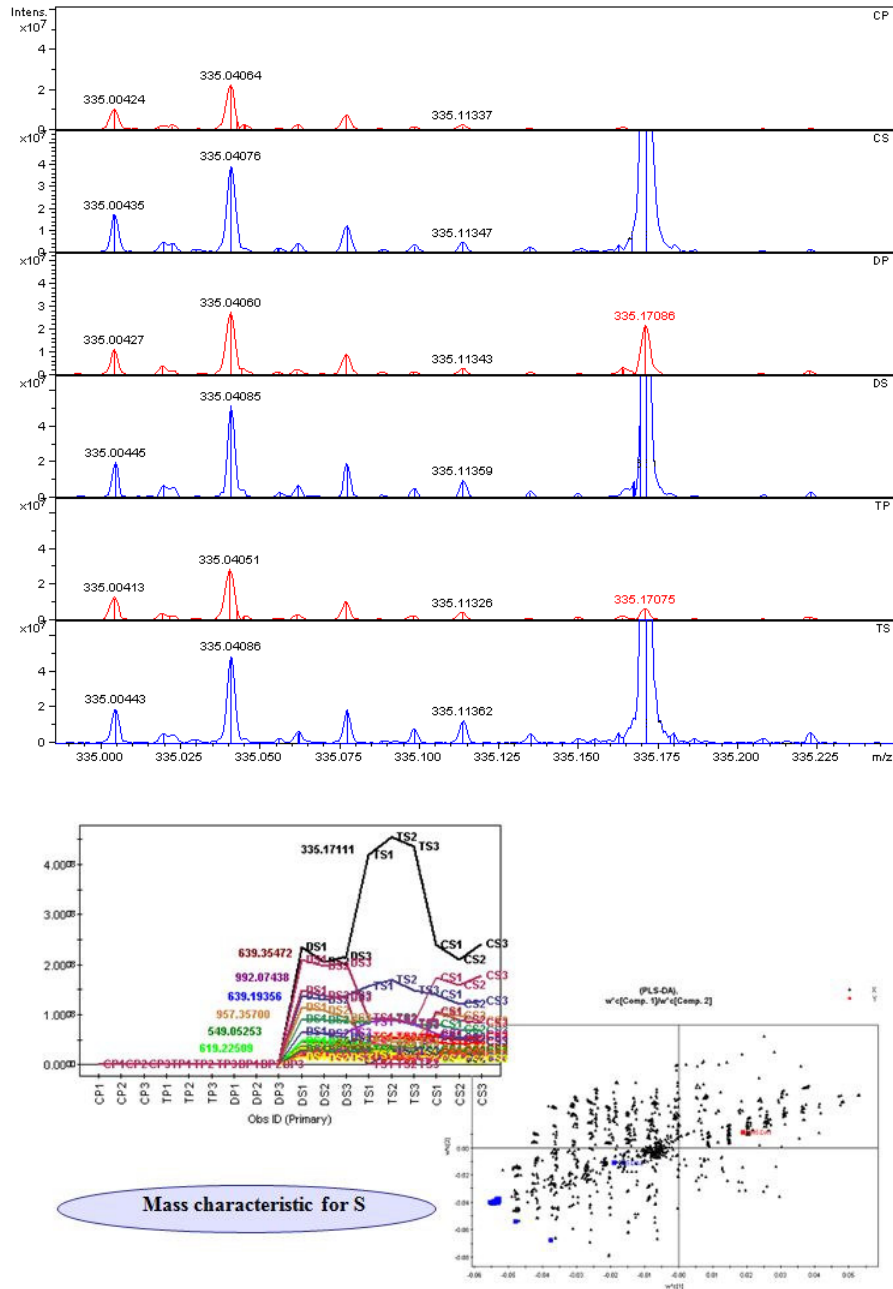


Figure 5.3: A): Details of the mass spectra on the 335.000-335.225 m/z range for the 2 species in the 3 forests, spotting the "contamination" of S

the pedunculate sample from Darney by sessile wood. B) Loading Plot with the trend plot of the masses particular for sessile wood, is delineated the characterization of the  $m/z$  335.17114 for this type of wood as defined in figure 5.3a.

As already mentioned, errors were made on the original identification of standing trees from the Darney forest : three out of the 20 P oaks were actually S oaks, and conversely, four out of the 20 S oaks were actually P oaks (Feuillat, et al., 1999). Since our sets of sawdust samples were prepared in ignorance of these erroneous attributions, we should observe peaks specific of the S species in the mass spectra of P samples from Darney, and vice versa. This is illustrated in figure 5.3a for the  $m/z$  335.17114 peak attributed to the whiskylactone precursor, which is only detected for the P sample from Darney, and not for the other P samples.

It has to be noted that each of the six samples studied actually represent the average polled sample of 20 wood pieces (as used for one barrel), each piece coming from one distinct tree. In contrast, all of the previously reported studies were based upon the detailed analysis of any single trees, which obviously favours the detection of singularities, but at the expense of excess instrument time. The major consequence of working with "averaged" mass spectra is that singularities at the species or forest level (inter-individual variability) might be attenuated beyond recognition. On the other hand, any differentiation based upon "averaged" spectra will represent more solid evidence of actual tree distinction. To the best of our knowledge, this is the first study of non-targeted analysis of "averaged" oak metabolites leading to a clear species differentiation between *Quercus robur* L. and *Quercus petraea* Liebl., and as shown below, to a forest differentiation.

The 1D  $^1\text{H}$  NMR analysis confirms this differentiation as shown in figure 5.4. Principal component analysis has been found to be a suitable method for the comparison of forest-consolidated wood extracts (for both P and S species) on NMR data. In the present study, this analysis led to one statistically significant principal component accounting for 62% of the data variability, which confirmed the chemical characteristic of P and S wood. This conclusion was supported by the score plot (see figure 5.5 upper inset) showing positive values only for P and negative values for all the S woods.



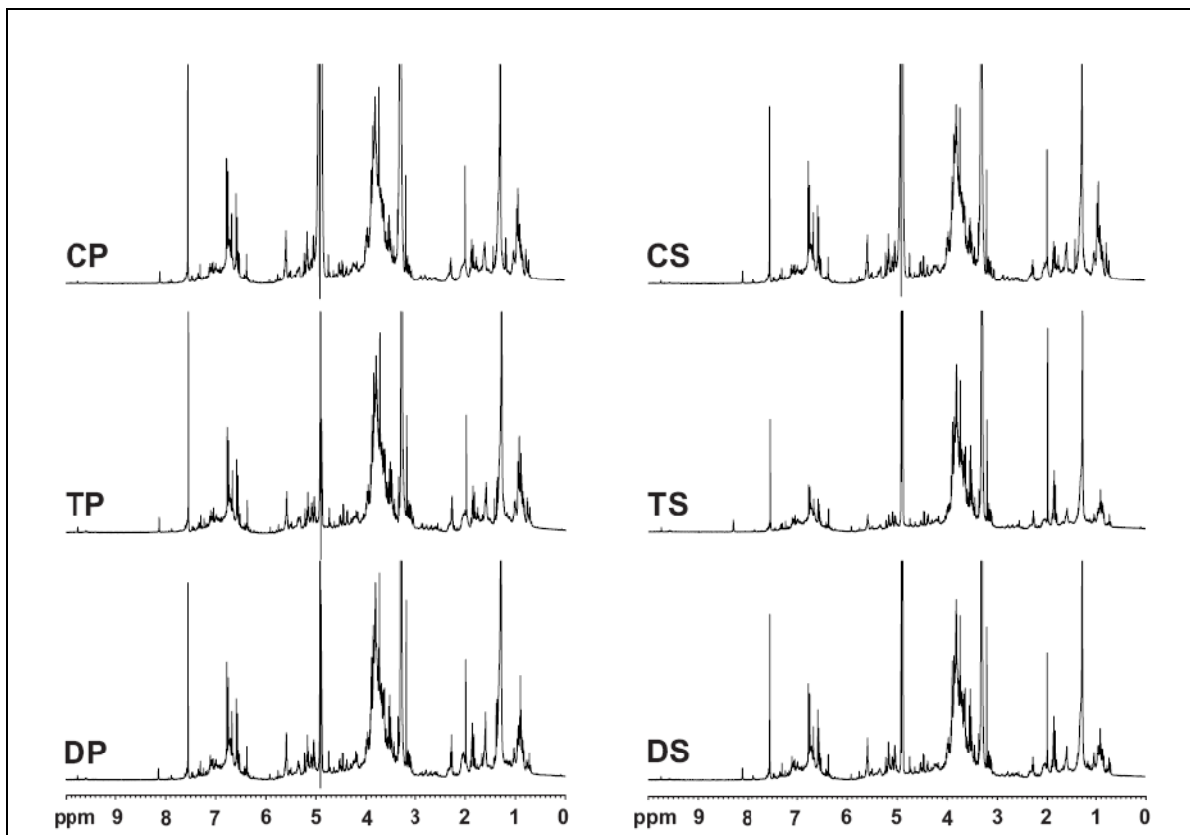


Figure 5.4:  $^1\text{H}$  NMR spectra of the different consolidated wood extracts.

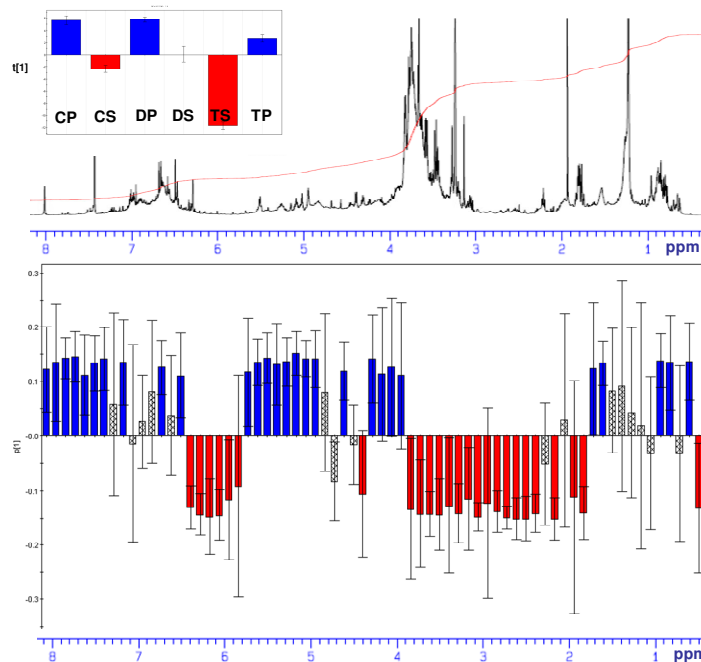


Figure 5.5: Typical  $^1\text{H}$  NMR spectrum for the CS wood extract, along with the corresponding Principal component analysis over the six lots of methanolic extracts. The score plot (upper left inset) and the loading plot (bottom) of the first principal component from the analysis of the 6 spectra are also shown. Only significant variables are colored in the loading plot, which highlight discriminant chemical shifts weightings for the two species.

As indicated by the congruence in the line shapes of the one-dimensional  $^1\text{H}$  NMR spectra which indicates molecular environments, the six wood extracts investigated showed considerable similarity at the level of coarse molecular fragments (see figure 5.4). However, one-dimensional  $^1\text{H}$  NMR spectra showed variation in both the NMR integrals of these coarse substructures and in the fine detail of line shapes (see figure 5.5 bottom). Clearly, the proportion of  $^1\text{H}$  chemical shifts that weight for the discrimination of P extracts is higher

(comprising the 0.2 - 1.8 ppm, the 3.8 - 5.6 ppm regions and the section downfield from 6.4 ppm) than the proportion of chemical shifts which discriminate S extracts (1.8 - 3.8 ppm and 5.6 - 6.4 ppm regions). The variance of NMR integrals allows quantifying the occurrence of substructures, and pattern analysis in multiple 2D NMR spectra aids in structural assignment of classes of molecular environments down to individual molecules, nicely complementary to mass spectral findings. Accordingly, we considered the selection of a single sample sufficient for in-depth NMR characterization by a suite of two-dimensional NMR experiments, which in combination, provide single bond ( $^1J$ ), geminal ( $^2J$ ) and vicinal ( $^3J$ ) connectivity, allowing the assignment of molecular fragments across three bonds (see figure 5.6).

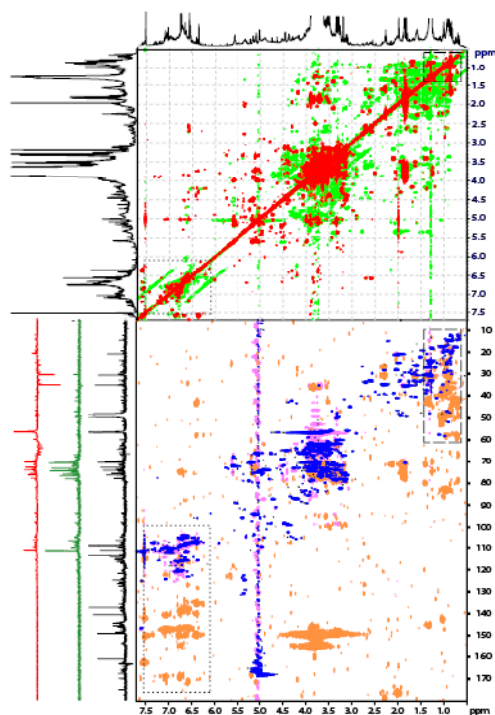


Figure 5.6: top:  $^1\text{H}, ^1\text{H}$  COSY (red) and  $^1\text{H}, ^1\text{H}$  TOCSY NMR spectra (light green) of the CS wood extract with  $^1\text{H}$  NMR projection spectra and

(bottom) overlay of  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC (blue),  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC (orange), and  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC-TOCSY (bright purple) NMR spectra, together with edited  $^{13}\text{C}$  NMR projection NMR spectrum: DEPT-135 (methylene CH<sub>2</sub> down; red), DEPT-90 (methine only; green) and standard  $^{13}\text{C}$  NMR spectrum (black); no appreciable  $^{13}\text{C}$  NMR signal intensity was found below  $\delta(^{13}\text{C}) = 180$  ppm at this S/N ratio. The dotted box indicates cross peaks derived from oxygenated aromatics, while the dashed box denotes cross peaks from branched aliphatics.

In the aromatic region of chemical shifts [ $\delta(^1\text{H}) > 6.2$  ppm], cross peak positions in COSY and HSQC NMR spectra were shielded in both  $^1\text{H}$  and  $^{13}\text{C}$  frequencies, indicating predominance of multiply oxygenated aromatic rings (Perdue, et al., 2007) in agreement with the  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC NMR cross peaks (figure 5.6, dotted box), which showed multiple quaternary carbon atoms in the  $^{13}\text{C}$  NMR shift range from  $\delta(^{13}\text{C})$ : 106-170 ppm (with maximum cross peak amplitude between  $\delta(^{13}\text{C})$ : 130-150 ppm). Furthermore, the minor contribution of aromatic environments to the COSY cross peak integral (see figure 5.6 top) as compared with the sizable integrated intensity of aromatic hydrogen obtained from one dimensional  $^1\text{H}$  NMR spectra (see figure 5.5) indicated the occurrence of many isolated aromatic protons - suggesting convincingly rather extensive degrees of aromatic substitution. All these spectral features are typical of ellagitannins, which are well established wood constituents (Quideau, et al., 1996), (Herve du Penhoat, et al., 1991), therefore confirming that the  $^1\text{H}$  NMR spectral region downfield from 6.2 ppm, which discriminates P extracts, mostly corresponds to ellagitanins. Similarly, the proton NMR resonances in the 5-6 ppm chemical shift range were not generated from olefinic protons but were attributed to the phenolic ester type because of  $^1\text{H}$ ,  $^1\text{H}$  COSY,  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC and  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC cross peak positions, which occupied typical shift ranges of phenolic esters rather than those of double bonds. The binding partners as identified from  $^1\text{H}$ ,  $^1\text{H}$  COSY cross peaks most likely are various carbohydrates, themselves providing strong signatures in all NMR spectra (e.g. > 35 methylene carbon signals (OCH<sub>2</sub>) with  $\delta(^{13}\text{C})$ : 60-66 ppm as well as methylene-derived HSQC cross peaks). Again, these features agree with the 5-5.6 ppm region of the  $^1\text{H}$  1D spectra, which contributes to the P species discrimination, being correlated to oak tannins, possibly of the galloyl ester type (Mämmela, et al., 2000). Conversely, the presence of several cross peaks between  $^1\text{H}$  signals in the 3 - 3.8 ppm region and  $^{13}\text{C}$  signals in the 60 - 80 ppm region of the  $^1\text{H}$ ,  $^{13}\text{C}$  HSQC and  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC spectra (see figure 5.6 bottom) indicate that carbohydrates probably participate to the discrimination of

the S species by the 2.8 - 3.8 ppm region of the  $^1\text{H}$  NMR spectra (see figure 5.6). Recently, dehydro- and deoxyellagitannins have been identified in toasted oak wood (Glabasnia, et al., 2007). In general, extended proton spin systems were rather found in the aliphatic section; sizable degrees of branching in purely aliphatic structures ( $\delta(^1\text{H}) < 1.2$  ppm) are also indicated by positions of  $^1\text{H}$ ,  $^{13}\text{C}$  HMBC cross peaks, with carbon chemical shifts up to 60 ppm (cf. dashed box in figure 5.6).

### 5.6.3 The forest effect

Partial least square discriminant analysis (PLS-DA) of three times six sets of samples resulted in clear differentiation according to species and to the geographic localization of the forests. This is the first time that a molecule-based differentiation according to geographical origin is demonstrated between oak trees from distinct forests in a given country. Figure 5.7 shows the 3D score plot of the three times six sets of samples, which indicates good discrimination of the six forests. Most interestingly, the closer correlation between P groups compared to S groups, defines a much higher homogeneity among the former group. From this analysis, it was possible to draft a list of masses characteristic of each of the six forests, based on correlation coefficient values. Although selected for this study, P oaks are actually scarce in the Tronçais forest, which is much more renowned for the quality of its sessile oaks for wine ageing (Mosedale, et al., 1996). Therefore, these findings not only agree with this fame, largely attributed to the higher whiskylactone contents and finer grains of S oaks from Tronçais, but they also provide molecular evidence for this distinction. Indeed, S oaks from Tronçais can be discriminated on the basis of more than 194 mass peaks (Table 5.1), all of them being unambiguously associated with absolute formulas. A particular emphasis should be put on the differentiation of TS samples on the basis of hexoses as discriminating molecules (see figure 5.7). Indeed, the PCA and PLS-DA analyses of  $^1\text{H}$  NMR spectra (see figure 5.5) already showed that the 2.8 - 3.8 ppm region was weighting for the discrimination of the S species, and that this weight was maximum for the Tronçais forest. Bearing in mind that 2D NMR experiments supported the correlation of this  $^1\text{H}$  chemical shift range to carbohydrates, together these results bring insights into a possible sweetness that would particularly characterise oaks from the Tronçais forest.

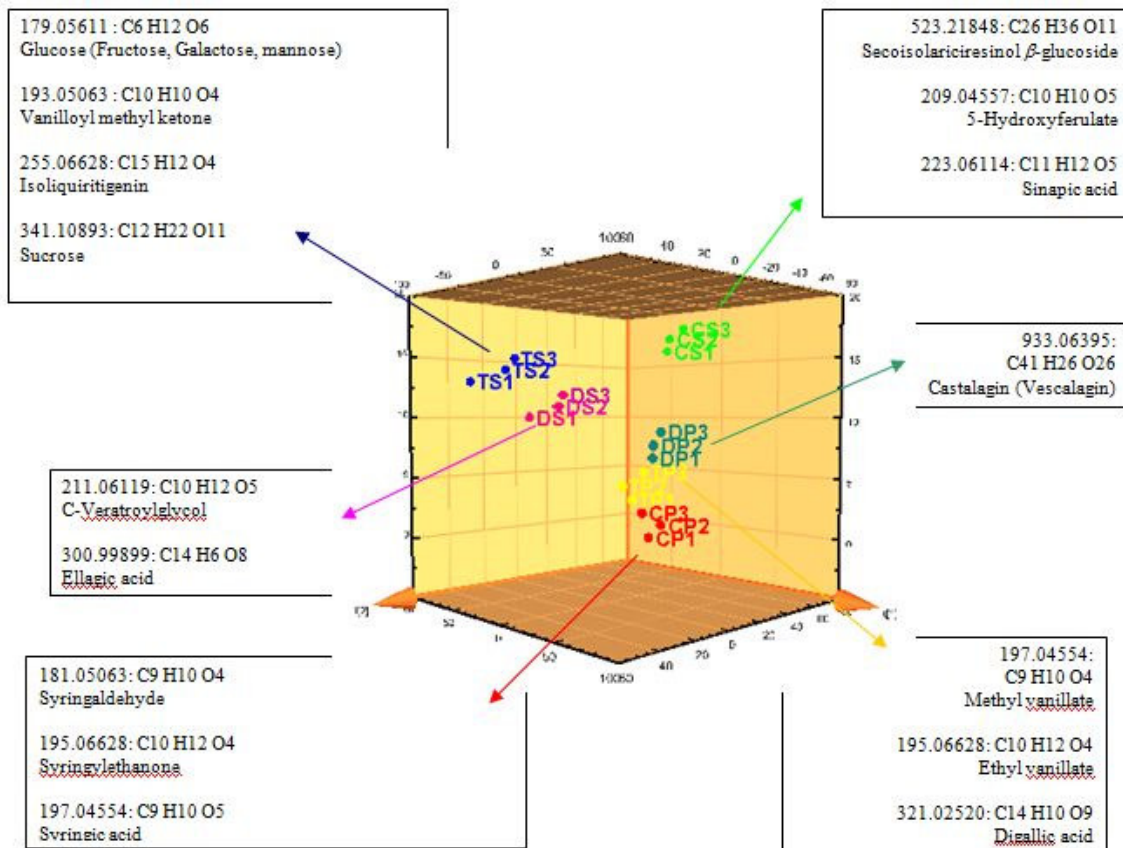


Figure 5.7: PLS-DA score plot of the 18 sets of samples ( $Q^2(\text{cum}) = 0.80$ ,  $R^2(Y) = 0.98$ ), grouped by forests, with the indication of some of the discriminating absolute masses (from negative ESI), the corresponding unique formulas (from neutral mass) and possible associated molecules, known to be related to wood.

117 | Expressing forest origins in the chemical composition of coopeage oak woods and corresponding wines by ICR-FT/MS

SESSILE					PEDONCULATE				
Forest	Mass	Coeff	Mean		Forest	Mass	Coeff	Mean	
			$\frac{I - (I_{min})}{I_{(max)} - I_{(min)}}$					$\frac{I - (I_{min})}{I_{(max)} - I_{(min)}}$	
				stdev					stdev
CS	209.04557	0.0012269	0.0532841	0.1860487	CP	261.13431	0.0016867	0.0159536	0.0216156
CS	223.02477	0.0012233	0.0592463	0.2120610	CP	310.20947	0.0012681	0.0008400	0.0016207
CS	223.06114	0.0012206	0.0516499	0.1829567	CP	337.31110	0.0022599	0.0132086	0.0178031
CS	265.14787	0.0022034	0.0077040	0.0122218	CP	359.07732	0.0016086	0.0178241	0.0208826
CS	283.04599	0.0010271	0.0106260	0.0122387	CP	382.03010	0.0011505	0.0001859	0.0005414
CS	295.04604	0.0014965	0.0463954	0.1695424	CP	387.05696	0.0016664	0.0166002	0.0227721
CS	300.97761	0.0017240	0.0457148	0.1732080	CP	401.01519	0.0016015	0.0173745	0.0223815
CS	301.00358	0.0020980	0.0983185	0.1497242	CP	411.34803	0.0016311	0.0138563	0.0191180
CS	301.20211	0.0014966	0.0625165	0.2341163	CP	412.04076	0.0013906	0.0012469	0.0023735
CS	315.25383	0.0012180	0.0377002	0.1268625	CP	420.05199	0.0030483	0.0004320	0.0012652
CS	335.17111	0.0015807	0.0367204	0.0486027	CP	483.27578	0.0012014	0.0023743	0.0039558
CS	339.01461	0.0012459	0.0059080	0.0075186	CP	517.33893	0.0018963	0.0159182	0.0242691
CS	359.04091	0.0017332	0.0489994	0.1854479	CP	517.36032	0.0026088	0.0043515	0.0067424
CS	369.04641	0.0012737	0.0643462	0.2337736	CP	533.28266	0.0019880	0.0037731	0.0049320
CS	389.07261	0.0010613	0.0607169	0.2186562	CP	533.30812	0.0020881	0.0686038	0.1589409
CS	449.05469	0.0013793	0.0075487	0.0170316	CP	534.04803	0.0013946	0.0005551	0.0012927
CS	467.21377	0.0011869	0.0608641	0.0846695	CP	550.03223	0.0022100	0.0060335	0.0089751
CS	468.01847	0.0019012	0.0006294	0.0009693	CP	557.02160	0.0024764	0.0075955	0.0112075
CS	486.15524	0.0027694	0.0022789	0.0051445	CP	559.29161	0.0023601	0.0007765	0.0018004
CS	487.14106	0.0022284	0.0043383	0.0076528	CP	577.04779	0.0016513	0.0181509	0.0228962
CS	487.16418	0.0015638	0.0053006	0.0076645	CP	659.05498	0.0011374	0.0028838	0.0056119
CS	487.20018	0.0017429	0.0048537	0.0075541	CP	665.35648	0.0022460	0.0013923	0.0023503
CS	498.13394	0.0022098	0.0064487	0.0107712	CP	667.37178	0.0021849	0.0025413	0.0042765
CS	506.05318	0.0018477	0.0572406	0.2352967	CP	677.35527	0.0014676	0.0094945	0.0088840
CS	511.16118	0.0015064	0.0038604	0.0046780	CP	679.28516	0.0010941	0.0051613	0.0060651
CS	515.41131	0.0010590	0.0119153	0.0131847	CP	681.38822	0.0019168	0.0056089	0.0091718
CS	523.21848	0.0015176	0.0030385	0.0037300	CP	695.27530	0.0013524	0.0024042	0.0031434
CS	525.08892	0.0012376	0.0044192	0.0054589	CP	709.37463	0.0022507	0.0010773	0.0025726
CS	525.12517	0.0015182	0.0035196	0.0043386	CP	722.09564	0.0011215	0.0004584	0.0013788
CS	529.17180	0.0015162	0.0026698	0.0033376	CP	750.09028	0.0012621	0.0003338	0.0009819
CS	532.02102	0.0019295	0.0012302	0.0019616	CP	782.95466	0.0033281	0.0011086	0.0021979

118 | Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines by ICR-FT/MS

CS	549.05254	0.0015568	0.0100280	0.0130868	CP	821.04574	0.0012363	0.0143886	0.0236617
CS	555.13613	0.0017404	0.0034108	0.0047479	CP	831.25355	0.0014553	0.0017152	0.0022323
CS	558.15527	0.0022063	0.0055236	0.0090203	CP	832.38681	0.0018282	0.0082471	0.0160175
CS	559.37666	0.0017157	0.0502292	0.1373532	CP	843.08298	0.0011849	0.0018338	0.0037427
CS	567.04230	0.0018969	0.0220603	0.0829797	CP	843.43449	0.0021892	0.0066168	0.0082570
CS	595.21887	0.0015007	0.0022934	0.0028000	CP	845.35367	0.0019125	0.0007128	0.0016451
CS	609.18285	0.0015083	0.0024203	0.0029845	CP	849.38598	0.0012885	0.0024270	0.0046792
CS	609.48930	0.0015066	0.0021442	0.0026567	CP	860.43822	0.0017274	0.0067240	0.0080929
CS	613.22938	0.0014851	0.0034026	0.0040580	CP	867.04545	0.0014312	0.0008217	0.0020139
CS	617.11535	0.0015264	0.0261327	0.0930840	CP	877.02891	0.0022221	0.0006521	0.0015591
CS	619.20471	0.0012134	0.0045383	0.0053406	CP	891.04528	0.0012160	0.0030237	0.0088571
CS	619.22509	0.0015478	0.0084680	0.0109094	CP	895.03768	0.0040816	0.0007921	0.0018700
CS	621.43782	0.0015102	0.0022191	0.0030240	CP	905.02198	0.0018173	0.0008362	0.0019827
CS	625.17788	0.0015530	0.0028386	0.0042063	CP	940.04226	0.0023076	0.0013750	0.0024320
CS	629.07922	0.0016984	0.0223969	0.0826541	CP	943.09564	0.0023807	0.0093598	0.0165457
CS	629.17280	0.0013158	0.0333663	0.1136252	CP	947.03097	0.0013384	0.0088455	0.0173105
CS	631.16700	0.0014983	0.0031127	0.0037102	CP	961.04853	0.0030347	0.0084774	0.0169710
CS	633.10966	0.0015217	0.0034604	0.0044579	CP	969.02966	0.0017101	0.0017774	0.0037019
CS	635.12483	0.0017476	0.0047086	0.0068085	CP	981.03734	0.0011608	0.0008395	0.0024454
CS	639.19357	0.0015755	0.0174131	0.0229083	CP	982.06049	0.0040674	0.0006312	0.0014704
CS	639.35472	0.0016031	0.0186053	0.0273154	DP	299.98437	0.0016302	0.0004214	0.0009725
CS	643.09439	0.0015150	0.0027616	0.0034570	DP	303.00217	0.0032120	0.0332270	0.0641488
CS	653.39184	0.0022051	0.0022046	0.0039329	DP	332.98860	0.0022482	0.0058883	0.0080909
CS	655.27481	0.0015770	0.0028619	0.0049050	DP	401.01519	0.0017405	0.0173745	0.0223815
CS	655.31598	0.0017451	0.0021107	0.0034168	DP	438.02714	0.0033289	0.0028481	0.0077915
CS	655.38274	0.0028057	0.0012379	0.0031033	DP	439.10709	0.0013281	0.0071691	0.0084135
CS	679.05837	0.0015164	0.0018896	0.0024640	DP	465.01206	0.0013479	0.0020085	0.0038798
CS	683.40200	0.0015646	0.0064708	0.0089184	DP	466.00485	0.0015255	0.0043058	0.0094614
CS	699.08568	0.0012627	0.0245120	0.0729247	DP	497.29124	0.0011513	0.0119058	0.0099282
CS	707.06317	0.0015317	0.0037530	0.0048609	DP	517.28272	0.0014314	0.0073838	0.0078549
CS	709.07086	0.0010997	0.0013518	0.0024138	DP	517.36032	0.0019440	0.0043515	0.0067424
CS	723.04868	0.0017608	0.0019007	0.0031275	DP	518.32115	0.0019567	0.3518602	0.3226987
CS	725.06431	0.0012806	0.0064346	0.0160024	DP	533.28266	0.0014073	0.0037731	0.0049320
CS	737.06457	0.0015250	0.0025999	0.0034435	DP	547.32848	0.0011319	0.0075840	0.0065633
CS	755.07592	0.0015241	0.0027111	0.0035662	DP	608.02708	0.0016982	0.0155053	0.0214343
CS	757.09157	0.0019186	0.0019890	0.0032013	DP	679.28516	0.0011749	0.0051613	0.0060651
CS	757.12750	0.0015055	0.0023040	0.0028393	DP	788.10699	0.0018052	0.0059196	0.0086943
CS	761.08691	0.0014596	0.0031429	0.0043489	DP	806.22154	0.0021139	0.0042061	0.0072405
CS	767.07656	0.0013831	0.0285138	0.0749425	DP	806.72049	0.0022614	0.0055344	0.0073579
CS	773.08610	0.0013530	0.0023386	0.0052432	DP	819.32354	0.0016476	0.0101990	0.0140316
CS	777.73580	0.0014921	0.0022197	0.0027569	DP	829.37055	0.0016621	0.0006556	0.0015406



119 | Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines by ICR-FT/MS

CS	782.96185	0.0031400	0.0008991	0.0025254	DP	833.39047	0.0017967	0.0042773	0.0062779
CS	789.08234	0.0012917	0.0075358	0.0187259	DP	835.32255	0.0017921	0.0022082	0.0034084
CS	797.08740	0.0015329	0.0022524	0.0031522	DP	843.44355	0.0018032	0.0040121	0.0060067
CS	801.41027	0.0012996	0.0793173	0.2150946	DP	845.36327	0.0018819	0.0024317	0.0031736
CS	805.05454	0.0011405	0.0156528	0.0389879	DP	848.38274	0.0017806	0.0035877	0.0084340
CS	807.36073	0.0013148	0.0009546	0.0019495	DP	859.43544	0.0012562	0.0207063	0.0151553
CS	811.06605	0.0012560	0.0029931	0.0036318	DP	872.24284	0.0017930	0.0009661	0.0022736
CS	817.28631	0.0015686	0.0026266	0.0043445	DP	875.43380	0.0016854	0.0026248	0.0038965
CS	817.35121	0.0021706	0.0013795	0.0026217	DP	971.34062	0.0016774	0.0015617	0.0036554
CS	819.40932	0.0017564	0.0112452	0.0166203	DP	990.07304	0.0017115	0.0035878	0.0086051
CS	831.09283	0.0017438	0.0018011	0.0027341	DP	993.08067	0.0023385	0.0032521	0.0054838
CS	833.40056	0.0024286	0.0014746	0.0029169	DP	997.37831	0.0020440	0.0012285	0.0020494
CS	841.14932	0.0015090	0.0022849	0.0028753	TP	261.13431	0.0017859	0.0159536	0.0216156
CS	847.41361	0.0021502	0.0013264	0.0023619	TP	310.20947	0.0014294	0.0008400	0.0016207
CS	855.09340	0.0017317	0.0016944	0.0026072	TP	325.23811	0.0046763	0.0065457	0.0151199
CS	855.23705	0.0015090	0.0022126	0.0027443	TP	332.98860	0.0012463	0.0058883	0.0080909
CS	861.04503	0.0015279	0.0019348	0.0026325	TP	350.99944	0.0038688	0.0023708	0.0046696
CS	861.08257	0.0019435	0.0015643	0.0025499	TP	355.08256	0.0029958	0.0027428	0.0086042
CS	877.04121	0.0017366	0.0019963	0.0030182	TP	359.07732	0.0017743	0.0178241	0.0208826
CS	885.04482	0.0015391	0.0035880	0.0049579	TP	368.36492	0.0017842	0.0045779	0.0088826
CS	887.06141	0.0017320	0.0017241	0.0033451	TP	384.04575	0.0032633	0.0001444	0.0004204
CS	901.04035	0.0017352	0.0017283	0.0025414	TP	385.22329	0.0045791	0.0019436	0.0045513
CS	905.03750	0.0010685	0.0023226	0.0044913	TP	387.05696	0.0017378	0.0166002	0.0227721
CS	907.39932	0.0017604	0.0014933	0.0024341	TP	409.40522	0.0014103	0.0301243	0.0331032
CS	909.10800	0.0019211	0.0140801	0.0534089	TP	411.34803	0.0016971	0.0138563	0.0191180
CS	909.21787	0.0015545	0.0021769	0.0033953	TP	470.05412	0.0021883	0.0003522	0.0008125
CS	917.03415	0.0017597	0.0037535	0.0058513	TP	497.29124	0.0013347	0.0119058	0.0099282
CS	919.09351	0.0019572	0.0011989	0.0021110	TP	499.27103	0.0029786	0.0019666	0.0032821
CS	921.06754	0.0011862	0.0910058	0.2284956	TP	499.30683	0.0016591	0.0247883	0.0171574
CS	925.75386	0.0014801	0.0017727	0.0022554	TP	503.30197	0.0017803	0.0147759	0.0137394
CS	927.73257	0.0022083	0.0010970	0.0019585	TP	517.06264	0.0015559	0.0143094	0.0199886
CS	929.09462	0.0016149	0.0008125	0.0020592	TP	517.28272	0.0019820	0.0073838	0.0078549
CS	931.05111	0.0013527	0.0803147	0.2325855	TP	518.27100	0.0011233	0.0067655	0.0083374
CS	941.09682	0.0011014	0.0012593	0.0024996	TP	518.32115	0.0025396	0.3518602	0.3226987
CS	945.11721	0.0031785	0.0018026	0.0050390	TP	529.28154	0.0027361	0.0017984	0.0030041
CS	951.42523	0.0017111	0.0015141	0.0022010	TP	531.31850	0.0031175	0.0054125	0.0087572
CS	957.09027	0.0019300	0.0265129	0.0733212	TP	533.26785	0.0025536	0.0012252	0.0028229
CS	957.35701	0.0015930	0.0103947	0.0165065	TP	533.33416	0.0029267	0.0063766	0.0083190
CS	958.36075	0.0017548	0.0054035	0.0090674	TP	533.35785	0.0029294	0.0018764	0.0031933
CS	992.07439	0.0022311	0.0089543	0.0162573	TP	534.31629	0.0011675	0.1724440	0.1424947
DS	449.05469	0.0014349	0.0075487	0.0170316	TP	535.47404	0.0038180	0.0057637	0.0111280

120 | Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines by ICR-FT/MS

DS	773.08610	0.0014060	0.0023386	0.0052432	TP	547.32848	0.0013239	0.0075840	0.0065633
DS	807.36073	0.0013992	0.0009546	0.0019495	TP	550.04478	0.0033279	0.0004197	0.0012220
DS	817.35121	0.0011724	0.0013795	0.0026217	TP	565.30301	0.0036399	0.0010653	0.0020603
DS	847.41361	0.0011866	0.0013264	0.0023619	TP	565.44846	0.0033236	0.0003935	0.0011480
DS	992.07439	0.0012191	0.0089543	0.0162573	TP	577.04779	0.0017028	0.0181509	0.0228962
TS	265.14787	0.0009535	0.0077040	0.0122218	TP	579.11500	0.0045147	0.0008590	0.0019950
TS	283.04599	0.0001351	0.0106260	0.0122387	TP	585.02546	0.0029112	0.0062628	0.0129483
TS	300.97761	0.0006664	0.0457148	0.1732080	TP	587.36771	0.0045062	0.0007591	0.0017472
TS	301.00358	0.0006683	0.0983185	0.1497242	TP	595.03773	0.0027678	0.0042104	0.0100045
TS	301.05646	0.0003238	0.0054314	0.0066780	TP	607.05904	0.0011659	0.0137452	0.0162142
TS	305.03039	0.0002656	0.0076083	0.0100992	TP	630.98452	0.0011519	0.0029559	0.0034226
TS	311.16858	0.0003278	0.0071890	0.0096830	TP	659.05498	0.0015478	0.0028838	0.0056119
TS	311.29553	0.0003310	0.0111191	0.0140187	TP	667.31480	0.0033226	0.0003873	0.0011301
TS	325.18417	0.0001379	0.0284232	0.0346318	TP	673.06911	0.0012956	0.0329905	0.0757852
TS	325.31109	0.0001491	0.0161045	0.0171110	TP	686.32868	0.0026091	0.0073597	0.0170485
TS	335.04084	0.0002602	0.0050183	0.0069534	TP	693.35101	0.0011225	0.0121031	0.0101619
TS	373.04290	0.0003107	0.0007407	0.0009458	TP	695.27530	0.0016133	0.0024042	0.0031434
TS	420.98376	0.0002536	0.0030358	0.0045103	TP	698.38607	0.0026802	0.0029494	0.0052892
TS	438.02002	0.0002364	0.0008288	0.0012317	TP	699.30436	0.0033227	0.0003876	0.0011305
TS	466.02870	0.0003233	0.4444444	0.5113100	TP	703.33669	0.0025889	0.0006889	0.0016425
TS	467.21377	0.0001974	0.0608641	0.0846695	TP	709.34636	0.0015846	0.0032479	0.0034992
TS	468.01847	0.0006116	0.0006294	0.0009693	TP	709.38329	0.0028044	0.0015257	0.0025683
TS	473.09393	0.0002402	0.0028119	0.0042274	TP	713.37695	0.0020040	0.0026480	0.0031889
TS	475.07319	0.0002438	0.0029594	0.0044787	TP	725.37945	0.0033461	0.0005693	0.0016578
TS	477.12521	0.0002515	0.0038276	0.0056731	TP	734.09621	0.0032868	0.0002125	0.0006188
TS	481.22921	0.0002748	0.0067837	0.0105754	TP	742.06417	0.0014202	0.0004780	0.0009286
TS	483.07867	0.0001271	0.3616735	0.3793750	TP	762.99120	0.0033379	0.0004971	0.0014486
TS	487.14106	0.0009153	0.0043383	0.0076528	TP	763.09363	0.0025055	0.0115138	0.0265930
TS	487.20018	0.0003700	0.0048537	0.0075541	TP	788.10699	0.0027364	0.0059196	0.0086943
TS	493.11998	0.0002466	0.0038095	0.0056091	TP	799.21438	0.0012718	0.0007870	0.0023577
TS	495.15106	0.0003304	0.0028300	0.0036798	TP	800.39749	0.0031634	0.0010302	0.0029997
TS	498.13394	0.0001334	0.0064487	0.0107712	TP	806.20999	0.0022616	0.0041908	0.0064247
TS	525.08892	0.0001139	0.0044192	0.0054589	TP	806.22154	0.0012544	0.0042061	0.0072405
TS	532.02102	0.0006090	0.0012302	0.0019616	TP	809.07827	0.0027545	0.0021850	0.0037509
TS	558.15527	0.0009393	0.0055236	0.0090203	TP	825.25634	0.0033299	0.0004364	0.0012729
TS	569.18783	0.0003137	0.0023925	0.0028742	TP	829.42420	0.0011866	0.0079873	0.0068787
TS	575.43217	0.0002599	0.0021925	0.0029662	TP	833.39047	0.0027205	0.0042773	0.0062779
TS	585.48966	0.0003162	0.0014224	0.0017522	TP	835.32255	0.0026614	0.0022082	0.0034084
TS	589.19341	0.0003408	0.0025357	0.0035482	TP	843.08298	0.0014848	0.0018338	0.0037427
TS	619.13064	0.0002514	0.0032279	0.0042826	TP	843.44355	0.0027055	0.0040121	0.0060067
TS	647.12569	0.0003059	0.0016805	0.0019849	TP	847.05585	0.0011326	0.0086641	0.0106134

TS	649.21414	0.0002833	0.0029212	0.0040020	TP	847.37602	0.0012147	0.0297288	0.0248544
TS	653.39184	0.0009026	0.0022046	0.0039329	TP	849.38598	0.0014610	0.0024270	0.0046792
TS	657.07404	0.0002552	0.0027250	0.0038326	TP	851.05036	0.0027113	0.0025933	0.0043748
TS	663.12093	0.0003078	0.0019321	0.0022862	TP	853.06721	0.0027138	0.0023774	0.0039928
TS	663.19275	0.0003187	0.0028924	0.0035202	TP	859.43544	0.0016283	0.0207063	0.0151553
TS	667.40707	0.0002623	0.0050046	0.0086129	TP	863.05110	0.0011292	0.0095904	0.0116597
TS	677.13662	0.0002394	0.0023390	0.0033107	TP	865.08161	0.0015994	0.0293315	0.1059131
TS	691.15199	0.0002532	0.0033436	0.0044809	TP	867.08282	0.0027494	0.0027558	0.0049568
TS	725.06431	0.0001170	0.0064346	0.0160024	TP	879.08006	0.0017649	0.0016251	0.0031403
TS	749.70398	0.0003167	0.0016978	0.0020886	TP	897.06655	0.0012031	0.0093306	0.0174223
TS	779.07485	0.0002304	0.0020079	0.0030826	TP	905.06045	0.0035171	0.0038715	0.0074937
TS	787.06894	0.0002397	0.0027684	0.0043919	TP	909.05635	0.0034220	0.0020917	0.0061315
TS	789.08234	0.0001189	0.0075358	0.0187259	TP	917.55956	0.0032937	0.0002369	0.0006896
TS	799.39234	0.0002913	0.0101956	0.0119571	TP	926.06322	0.0020077	0.0014569	0.0033910
TS	805.05454	0.0001200	0.0156528	0.0389879	TP	935.07572	0.0017651	0.0228998	0.0450481
TS	819.40932	0.0003334	0.0112452	0.0166203	TP	947.03097	0.0014717	0.0088455	0.0173105
TS	833.40056	0.0004203	0.0014746	0.0029169	TP	951.06356	0.0011587	0.0147818	0.0175322
TS	837.72026	0.0002400	0.0013325	0.0017898	TP	953.77103	0.0033037	0.0002822	0.0008223
TS	861.08257	0.0006106	0.0015643	0.0025499	TP	953.78963	0.0029999	0.0051140	0.0181424
TS	865.75181	0.0003092	0.0012661	0.0015155	TP	963.02703	0.0011162	0.0224662	0.0755017
TS	877.04121	0.0003395	0.0019963	0.0030182	TP	965.04392	0.0024223	0.0015816	0.0036917
TS	881.07291	0.0001430	0.0338800	0.0775514	TP	973.06654	0.0021945	0.0009244	0.0022008
TS	915.05694	0.0001817	0.1518868	0.3303776	TP	977.04098	0.0026864	0.0011804	0.0019730
TS	917.03415	0.0003275	0.0037535	0.0058513	TP	987.33855	0.0026036	0.0011793	0.0027869
TS	921.06754	0.0001352	0.0910058	0.2284956	TP	997.36213	0.0017402	0.0007841	0.0015187
TS	983.79391	0.0002850	0.0012685	0.0014721	TP	997.37831	0.0012562	0.0012285	0.0020494

Table 5.1: List of the specific peaks (absolute mass and coefficient) specific to the sessile and pedonculate species; additional information on the three forests origins (Citeaux, C; Darney, D and Troncey, T).

#### 5.6.4 Wood-wine correlations

Based upon the metabolomic differentiation of wood extracts, the discrimination of selected wines that were aged in barrels made of the particular averaged woods seems very promising. This is illustrated in figure 5.8 by the analysis of a Mercurey wine grown in P and S barrels. The different m/z distribution in the lower mass range as compared to the wood extract in figure 2.4

(chapter 2) is clearly visible; the expansion at nominal mass 227 shows the likely presence of resveratrol (not present in figure 2.4 (chapter 2)) that traces its origin from the grapevine.

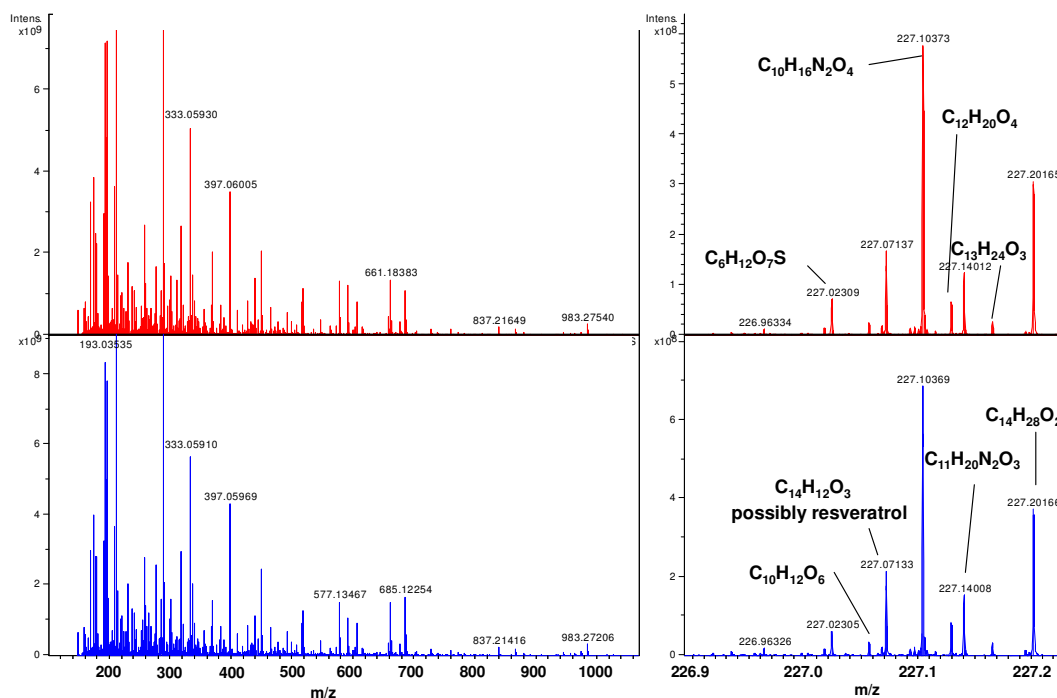


Figure 5.8: Typical negative ion mass spectrum of the mercury wine aged in sessile and pedunculate oak wood barrels from the Tronçais forest for 12 months; detail on masses 226.90-227.20 with elementary composition assignment of the major intensities (similar intensity of all peaks in mass 227 show no influence of wood species for that particular m/z).

Most interestingly, the expansion at nominal mass 335 (see figure 5.9) demonstrates the higher molecular diversity of the wine compared with the wood extract (see figure 5.3a) but nevertheless allows to verify the presence of oak wood biomarkers in the wine. Indeed, only wines aged in barrels made of S oak

woods exhibit the peak attributed to the whisky-lactone precursor, considered to be a bio-marker of S oaks. Analogous relationships apply throughout the entire mass range.

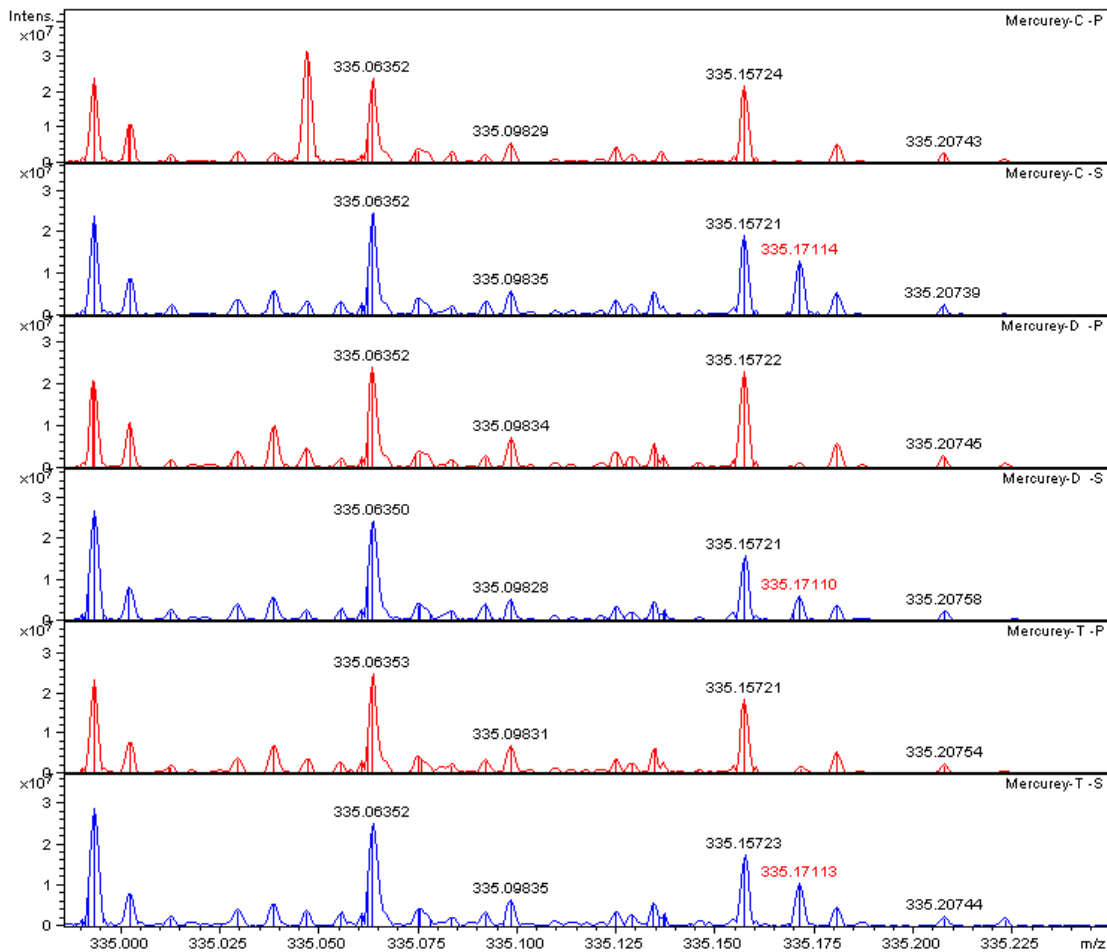


Figure 5.9: Details of the mass spectra on the 335.000-335.225 m/z range for the wine aged in barrels made of the 2 species in the 3 forests, spotting the presence of the m/z 335.17114 peak only in the wine aged in barrels made of sessile oaks.

## 5.7 Conclusion

Many studies of the variability of wood properties have concluded that the largest variations are observed between trees within a same forest (Mosedale, et al., 1996), (Feuillat, et al., 1997). However, all of these studies have relied upon targeted analyses of initially preselected compounds, which in the case of cooperage oaks, had been identified as responsible for organoleptic properties. In this study, we have applied an advanced ICR-FT/MS technique at the highest commercially accessible field strength, to assess the opportunity to molecularly discriminate a series of oak wood extracts and corresponding wines on a non-targeted basis. The major information provided by such an approach is the relative quantities of all the molecules that can ionise under the selected experimental conditions for any sample. In this context, wood is considered as a complex biological system that can evolve because of many environmental conditions related to the local ecosystem where it has grown, with the consideration that this multiparametric variation will express itself through a particular chemical space. Our results demonstrate that ultrahigh resolution ICR-FT/MS allows defining such chemical space down to single species in single forests. Furthermore, these lists of molecules, allow defining and identifying chemical sub-spaces, which could be associated to a forest regardless of the species, and alternatively, selectively associated to a species regardless of the forest.

These results provide insights of considerable novelty, referring to the identification of the chemical composition of oak woods as feasible by ultrahigh resolution ICR-FT/MS, capable of identifying thousands of distinct molecular compositions directly out of mixtures. Even if such full-scale metabolomics approach including identification of molecular structures remains at present extremely tedious due to the lack of experimental databases, a promising alternative approach is metabonomics. Here identification of any single peak (molecular structure) is not necessarily required. Instead, whole sub-spaces are considered and their variations from one sample to the other are monitored by the use of advanced processing tools, able to handle very large data sets. We are currently investigating the possibility of applying such metabonomics approach to our set of oak wood samples, in order to assess the feasible correlations with the sensory attributes that these woods can transfer to wine.

Finally, we envision general value and applicability in this non-targeted molecular level traceability, not only for cooperage, but more generally for vine

(beverages) forensics assessments on European or larger scale levels or for botanical science and silviculture to record environmental changes (such as climate modifications over decades), and to improve nutritional value and sensory properties of agricultural products based upon knowledge of molecular composition.





# Chapter 6

## 6 THE CHEMODIVERSITY OF WINES: FROM OENOLOGY TO “SYSTEMS OENOLOGY”

### 6.1 Introduction

As far as history recalls, wine has always been an unique beverage for humans, acting as dietary, religious, sensory or therapeutic commodity. Its chemical composition, the result of a complex interplay history between environmental factors (bio-, geo-, pedoclimatic), genetic factors (grape varieties) and viticultural practices, is considered to constitute the origin of this fame. Here, we show that an unprecedented chemical diversity of wine composition can be unravelled through a non-targeted *oenolomics* approach by ultrahigh resolution mass spectrometry which provides an instantaneous image of the thousands of metabolites present in exceedingly small quantities in wine, thereby integrating

the consequences of gene and enzyme regulation within metabolic pathways in the grapes, the yeast fermentation, the barrel-wood ageing, along with influences by the “terroir” and viticultural practices. In particular, the statistical analysis of series of barrel-aged wines revealed that nine-year old wines still express a metabo-geographic signature of the forest location where oaks of the barrel they were aged in have grown. Beyond *oenomics*, these data demonstrate that including dynamic changes of a wine chemical composition within the frame of a data driven “*systems oenology*” approach, allows to envision new directions for characterising the intricacy of wines, which results from complex interacting systems and processes, not easily or possibly resolvable into their unambiguous individual contributions.

Metabolic changes occur throughout the growth and maturation of grape berries, and at harvest time the berries contain the major grapevine compounds contributing to the body and flavour of the wine (Lund, et al., 2006). During winemaking processes and in particular during fermentation, these compounds act as carbon, nitrogen and element source for yeasts, and are either further metabolised, chemically transformed or directly transferred to the wine. Yeasts metabolism will further contribute to the wine enrichment through, for instance, the enzymatic liberation of particular volatile organic molecules responsible for the aroma of wine. Even if the biochemical and functional-genomics approach of enzyme signaling definitely help to clarify how the accumulation of active compounds is regulated at the molecular level in the grapevine or the grape (Goes da Silva, 2005), (Burns, et al., 2001), it would not be sufficient for providing an integrated picture of the actual organoleptic properties or therapeutic activities associated with these compounds in wine, because process-related synergistic effects certainly modulate these properties, to finally result in a unique beverage (Burns, et al., 2001). In traditional winemaking practices in particular, several processes can indeed subtly modulate the characteristics of wine, and in most cases, these modulations involve ‘trace’ amounts and interplay of metabolites within a complex matrix. As a consequence, it is likely that deeper understanding of organoleptic or therapeutic activities of wine will rely on its consideration as a complex blend of wine active compounds (WAC, in Wine Active Compounds-OenoPluri Media, Beaune - France, 2008). Recent findings indicate that similarly to red wines, certain white wine extracts could also exert cardioprotective effects on rats, with a pronounced antioxidant activity (Cui, 2002), although white wines are known to exhibit much lower amounts of antioxidant polyphenolic compounds than red wines.

Considerable progress has been made in recent years, in the characterisation of grape and wine metabolites (Jeandet, et al., 2007). Beyond a basic chemotaxonomic approach, today's ambition to understand the subtle aspects of wine composition is undoubtedly fostered by the various reports on the acknowledged therapeutic effects attributed to its moderate consumption (Marmot, et al., 1981), (Soleas, et al., 1997). Since the triggering works of Renaud (Renaud, et al., 1992), and the so-called '*French paradox*', numerous researchers have indeed attempted to identify the metabolites or the family of metabolites of wine, which could subsequently be considered as biomarkers of therapeutic activity (Jang, 1997), (Corder, 2006).

If therapeutic issues definitely contribute to the current progress in the identification of wine metabolites, organoleptic issues have fuelled, by far, the largest number of analytical studies over the past decades, because of the crucial role played by grapevine and wine macromolecules and secondary metabolites on the flavour and stability, and consequently on the wine industry (Bisson, et al., 2002). Numerous studies have therefore reported the identification of anthocyanins, tannins and their combinations that coexist in model wine solutions or that have been actually observed in wines, since these compounds are chiefly responsible for the colour and taste (Bakker, et al., 1997), (Cheynier, 2006). Similarly, and based on the assumption that any WAC may more or less contribute to sensory properties, other families of compounds such as organic molecules responsible for the aroma, organic acids, polysaccharides, amino acids, peptides and proteins have been the object of various studies, in particular tracking their way of transfer to wine from the grape or from yeast metabolism or from yeast lees (Mongay, et al., 1996), (Doco, et al., 1999).

To that respect, all of these previous and current analytical results contribute to an "*oenomics*" approach of wine, which we define, in accordance to the "*metabolomics*" definition of J. Nicholson and co-workers (Nicholson, et al., 1999) (Lindon, et al., 2007) as *the quantitative description of all low molecular weight metabolites in a specified biological sample or compartment* (here the local system = vine grapes, yeast or wood). The vast majority of wine analyses up to date rely on a molecular targeted basis and have assumed and often confirmed the presence of molecules in wine, in correlation to the particular property under investigation (organoleptic, therapeutic...). Here, we report the non-targeted metabolite analysis of a set of wine samples, which reveals the extremely high, yet unknown diversity of wine metabolites. In particular, we concentrated our analysis on a set of wines which were initially part of a full-scale study involving

nine French forests, designed to evaluate the influence of the geographic origin and the species of oak wood on the quality of wines matured in oak barrels (Feuillat, 2003).

This chapter is based on the article: “The chemodiversity of wines: from Oenology to “System Oenology”” (submitted to Science, 2008)

### 6.1.1 *Oenolomics*: describing the chemical spaces of wine

The *Oenolomic* approach, which enables an instant molecular picture of wines, requires both the mass resolving power and the mass accuracy of high-field Ion Cyclotron Resonance-Fourier Transformed Mass Spectrometry (ICR-FT/MS), up to considerable mass ranges. We recorded electrospray ionisation ICR-FT/MS mass spectra of samples representative of distinct steps of the elaboration of wine, proceeding from vine grape extracts to fully aged wines (see figure 6.1). Within the mass range explored (150-2000 m/z), the spectra exhibit several thousands of peaks, which correspond to the metabolites that can be ionised under the selected experimental electrospray conditions (see figure 6.2).

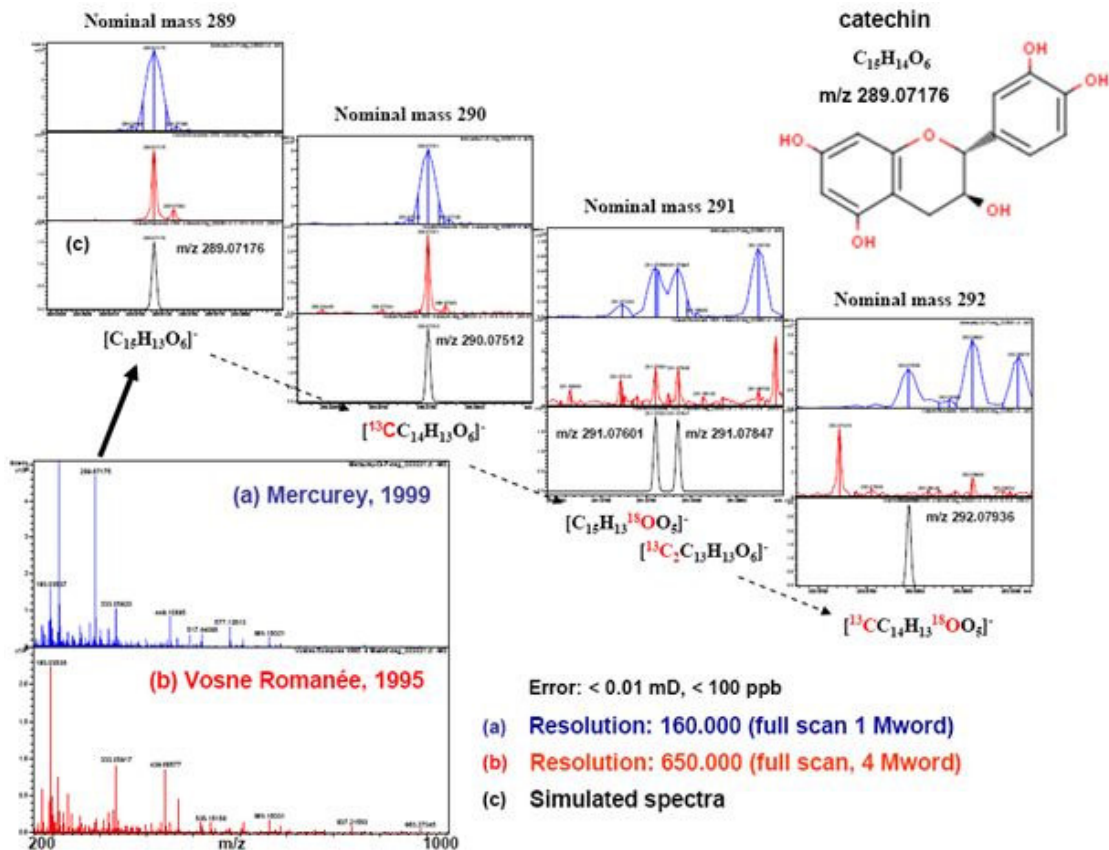


Figure 6.1: ICR-FT/MS spectra of a Mercrey 1999 and Vosne Romanée 1995 in electrospray negative mode from  $m/z$  200 to 1000. The mass peak 289.07178 typically dominant in the Mercrey is represented in detail with corresponding resolutions of around 160.000 and 650.000 for the Mercrey and Vosne Romanée wines, respectively, and compared to the simulated spectrum, showing the presence of corresponding isotopologues at nominal mass  $m/z$  290, 291 and 292. The presence of the isotopologue in the mass spectra is used to confirm the assigned elemental compositions of the signals in the spectra and ultrahigh resolution combined to high mass accuracy is needed to avoid false positive assignments.

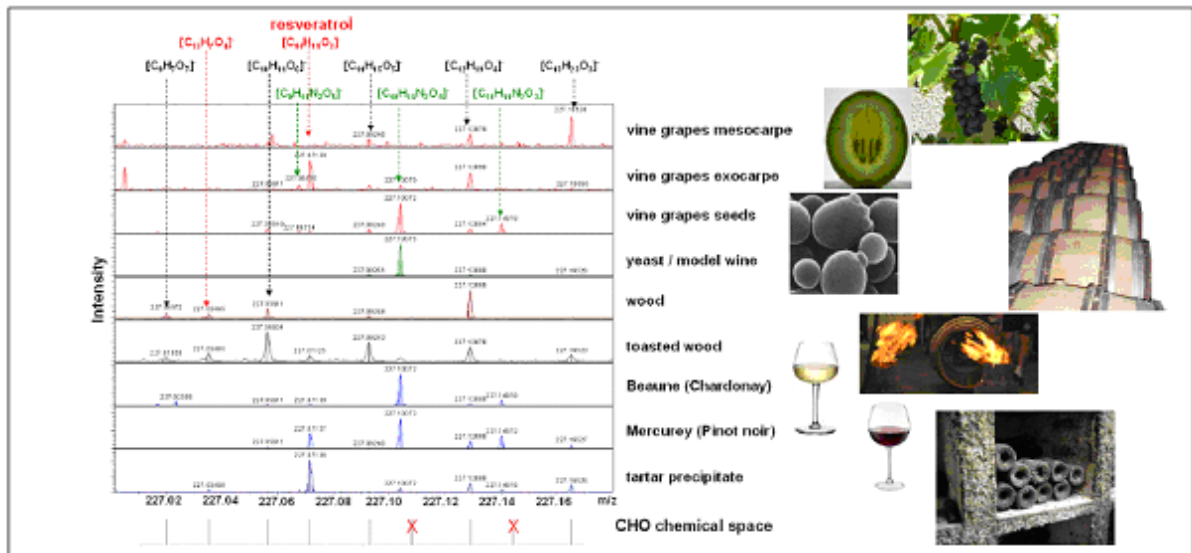


Figure 6.2: Detail on nominal mass 227 with all major signals and their attributions to CHONS elemental compositions from the grape berry to the wine. The tartar precipitate shows molecules that co-precipitated with the tartaric acid in the Vosnes Romanee 1995 bottle and thus are withdrawn from the wine during aging processes. Nitrogen containing compounds at mass 227.1037 are a signature of grape seeds and yeast; compound at mass 227.0713 attributed to stilbene resveratrol finds its origin in the exocarpe of grapes and its sink in the tartar precipitate. The chemical space of a wine can already be partially observed in the mass distributions within a single nominal mass (made possible by the ultrahigh resolution obtained with the 12 Tesla ICR-FT/MS); the 10 different elemental formulae shown in line (red, green, black) are related by a formal exchange of O by  $\text{CH}_4$  and represent 59% of the 17 feasible C,H,O-molecular compositions at this nominal mass.

Data reduction was followed according to elementary composition assignments using isotopic abundance patterns (Kind, et al., 2006) prior to any further data treatments (Rossello´-Mora, 2008). For example, the spectrum of a red wine from Burgundy (i.e. Vosnes Romanée, 1995) can lead up to 17400 peaks at a signal-to-noise = 2, (115000 at a signal-to-noise = 1), which can be unambiguously attributed

to 1180 unique elemental CHONS compositions (see figures 6.2 and 3.16 chapter 3) with 200 ppb tolerance and confirmation with  $^{13}\text{C}$ -signal (3890 compositions at 500 ppb tolerance), from which only a few hundred may correspond to masses of metabolites such as those gathered in our database (see figures 6.1c and 6.3), that have already been observed in model solutions or in wines with targeted analyses.

The diversity of chemical spaces of wine can already be observed in the mass distributions within the 200 millimass range of a single nominal mass (see figure 6.1a); the 10 different CHO elemental formulae shown (red, green, black traces) vary by a formal exchange of O with  $\text{CH}_4$  and occupy all possibilities within the feasible CHO compositional space in the mass range from 227.02 to 227.16 Dalton. For instance, when considering only the compositions based on C, H and O (CHO chemical space in figure 6.1a), 7 out of a total of 9 theoretically possible combinations appear in the different spectra within this 140 mDa mass range (Hertkorn, et al., 2007). The peak at  $m/z$  227.01714, which is present in the spectrum of the grape skin extract, but absent from the spectrum of the grape flesh extract corresponds to the  $[\text{M}-\text{H}]^-$  ion with absolute mass formula  $[\text{C}_{14}\text{H}_{11}\text{O}_3]^-$  and can most likely be assigned to resveratrol isomers. This attribution is further supported by the presence of an analogous mass peak in the spectrum of the Mercurey red wine, whereas it is absent from the spectrum of the Beaune white wine (see figure 6.1a). Interestingly, figure 6.1a shows reveratrol along with many other metabolites (see the corresponding full spectra in figure 6.2) in tartar precipitates that may appear in bottles upon ageing. Finally, figure 6.1a shows that, in this mass range, and in particular at the nominal mass 227.1037, nitrogen containing molecules are a signature of grape seeds and yeast metabolites.

We endeavoured to identify or structurally relate as many of these peaks as possible to known compounds, by questioning topic related available databases (KEGG, MassTRIX, KNApSAcK) (Suhre, et al., 2008) and/or implementing separation/purification techniques for subsequent structural elucidation. However, due to the deficiency of current experimental databases and the chemical complexity of wine, such task remains still out of range and only limited to the elementary composition analysis of a few known wine components (Cooper, et al., 2001). An initial interpretation of such compilations is made following assignment of elemental compositions with two-dimensional van Krevelen diagrams (Rossello´-Mora, 2008), (Wu, et al., 2004), which sort each elemental composition onto two axes according to its H/C and O/C atomic ratios (see figure 2.16a).

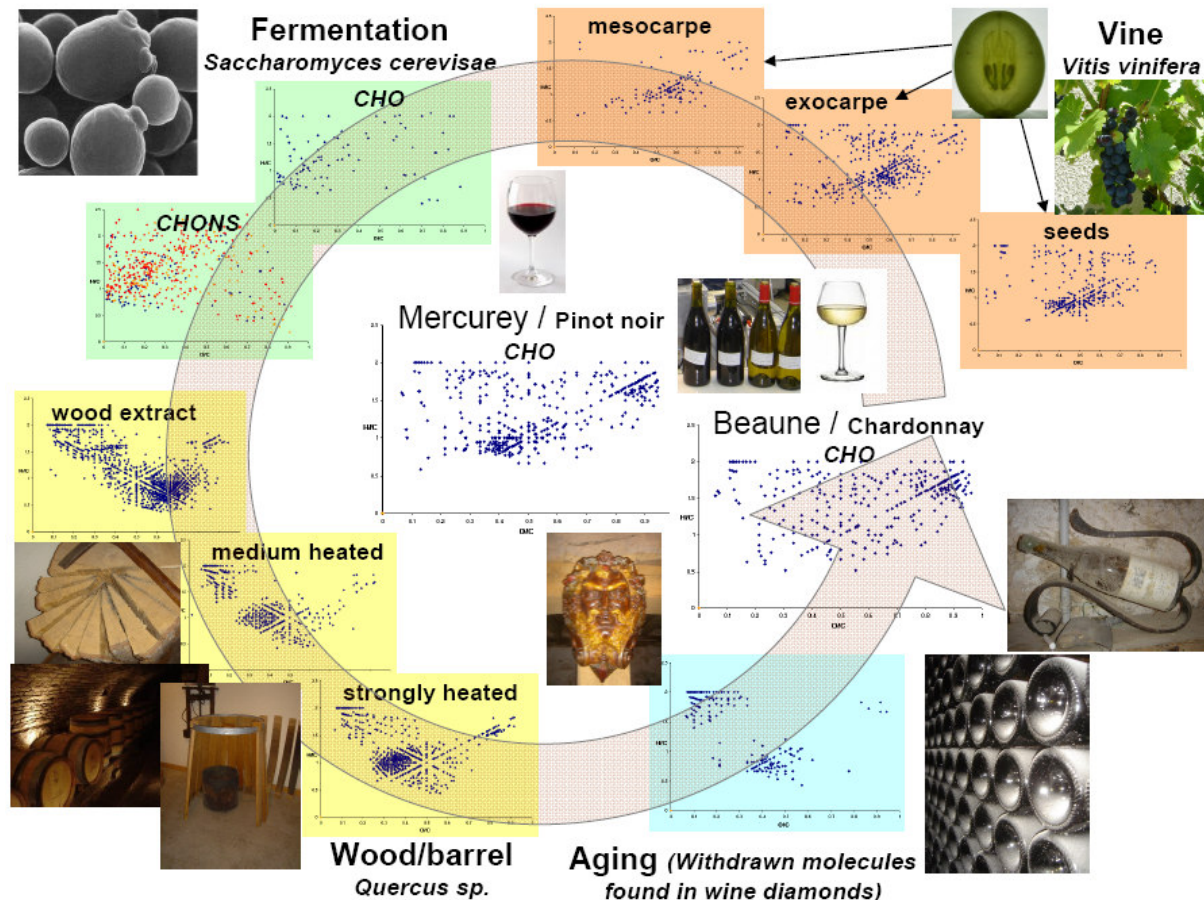


Figure 6.3: CHO-chemical space in the elaboration history of a wine; “from vine to wine” as visualised with the (O/C, H/C) van-Krevelen diagrams. Complementarities of the factors are visualised in the superimposition of the molecular footprints; interestingly the fermentation step has much more importance in the CHON and CHONS space.

Unprecedented graphical representations of the various chemical spaces (CHO, CHOS, CHON, CHONS) of wines are then obtained, which visually highlight specific



cluster series of elementary compositions observed within nominal masses (see figure 6.1a). Using a home-compiled database of compounds that can exist in model wine solutions or that have been actually observed in wines, allows to similarly represent the specific contributions of phenolics, peptides, polysaccharides, nucleotides and any other classes of compounds present in wines, and which can be positively or negatively ionised (see figures 6.1c 6.3). It must be noted however, that many of the compounds responsible for the aroma of wines, which exhibit  $m/z$  values below 150, are not detected under our experimental conditions.

### 6.1.2 From *oenomics* to oenonomics and systems oenology

When a wine's spectrum is transposed into van Krevelen diagrams, the result not simply reflects the superposition of all separate diagrams that can be assembled from each separate steps of its elaboration. Instead, it provides an instantaneous metabolite picture of a complex biological system (super organism approach), which encompasses all the initial contributions of genetic factors modulated by constantly evolving environmental factors (see figure 6.3). When analysed separately, each of these steps can be characterised by the potential release into the wine of thousands of compounds of extensive molecular diversity (see figure 6.3). In the CHO compositional space, seed and skin extracts would be dominated by tannins and anthocyanins polyphenolic structures, but many other compounds with higher H/C ratios could also be extracted. In contrast, in a chemical space restricted to CHO molecules, a flesh extract or a *Saccharomyces cerevisiae* culture medium would appear poorer, though still containing hundreds of distinct molecules. However, expanding the chemical space to CHONS elemental compositions, as illustrated for the *Saccharomyces cerevisiae* culture medium (see figure 6.3), reveals a much larger chemical diversity of nitrogen and/or sulfur containing metabolites typical of yeast core metabolome (Suhre, et al., 2008). It must be borne in mind that, ICR-FT/MS alone does not allow to distinguish isomers, therefore, it is likely that in any of the observed chemical spaces, the actual chemical diversity is considerably higher than that derived from mass peaks alone (Hertkorn, 2007). During the elaboration of wines, barrel aging is an important environmental factor. Indeed, initially aimed at being suitable containers, oak barrels became practical means of modulation of fine sensory characteristics of wine, and several studies have been devoted to the modifications undergone by the wine during oak barrel aging, with particular

emphasis on the aromatic complexity provided by the contact with more or less toasted wood staves, in conjunction with low oxidation conditions enabled by this porous container (Garde-Cerdan, et al., 2006). Barrel aging is a striking example of the extremely complex modifications that a wine can undergo (Jarauta, et al., 2005). In addition to natural clarification, colour stabilisation favoured by ellagitannins extracted from oak wood and other acid catalysed reactions between hydrolysable tannins and wine nucleophiles (Quideau, et al., 2005), oak wood can act as a sorbent, with an appreciable selectivity towards hydrophobic metabolites of wine (Barrera-Garcia, et al., 2006). Thousands of molecules can actually be extracted from oak wood barrels, with a clear distinction on their nature according to the level of toasting (see figure 6.3).

In particular, at increased temperatures of toasting, more extensively oxygen depleted derivatives are formed, with molecules of O/C and H/C elemental ratios around 0.35 and 1 dominating. As mentioned before, one particular not necessarily desired step is the formation of tartar solid precipitate upon aging. In that case, although numerous molecules are involved, most of them belong to small acids including notably tartaric acid, and polyphenolic molecules such as anthocyanins, as can be observed from the red colour of these precipitates. As a whole, ICR-FT/MS clearly provides an instantaneous chemical picture of a wine, where the overall molecular composition is more than the sum of individual molecular contributions.

In that context, the metabolomic approaches in oenology would require the analysis of countless samples in order to gather a comprehensive description of wine metabolites. Even advanced protocols such as the “*Architecture for Metabolomics*” (Jenkins, 2005), would have to integrate the decisive, yet so versatile ‘*human*’ factor, since in essence, wine producers are providing a sensory experience to the consumer (Bisson, et al., 2002). Alternatively, the non-targeted metabolomics approach (Nicholson, et al., 1999), which combines multivariate statistics with high-dimensional unannotated variables, offers the possibility to integrate all the history of time-related metabolic changes of wine throughout its elaboration process. In this context, we define for a given grape genotype, and following Nicholson and co-workers (Nicholson, et al., 1999), (Lindon, et al., 2007), (Nicholson, 2006), (Lindon, et al., 2007), *Oenomics as the sums, products & interactions of the individual compartments/metabolomes in a complex organism (here the ‘Global’ System = wine)*. *Oenol(n)omics* thus becomes a non-targeted top down approach, non hypothesis-driven in the molecular level analysis of wine, and means ‘*understanding biochemical mechanisms, identifying*

*biomarkers, quantitatively analyzing concentration and fluxes, probing molecular dynamics and interactions*'. Accordingly, the systems oenology goal for a given grape genotype is a description of the qualitative and quantitative dynamic and multiparametric metabolic response of wine to environmental modifications.

## 6.2 When systems oenology witnesses to the story that a wine tells

In 1998, a full-scale integrated study involving 9 French forests and 4 sets of French wines was designed to evaluate the influence of the geographic origin and the species of oak wood on the quality of wines matured in oak barrels (Feuillat, 2003). Each of these 4 sets corresponded to 12 repetitions of the same wine, which only differed by the oak wood species and origin of the trees used for the elaboration of barrels they were aged in. We hypothesised that such sets of samples would represent unique panels of wine compositions with subtle variations, and as such, ideal candidates for the assessment of a systems approach in oenology.

We recorded the negative and positive-ion electrospray ionisation mode ICR-FT/MS mass spectra of each of 60 wines and these data were further statistically processed in order to identify possible discriminations among wines (only the negative ion electrospray data are shown here / positive ion data showed the same differentiations). PLS-DA score plots of wines according to their colour or geographical origin, and therefore variety provided an illustration of the diversity of metabolites that could basically lead to significant discriminations (see figure 6.4 a, b). The two predictive components of the PLS-DA model,  $R^2(Y)=0.99$  (6.4a and 6.4b) and the prediction accuracy  $Q^2(\text{cum})=0.96$  (6.4a) and  $Q^2(\text{cum})=0.95$  (6.4b) were obtained through a typical seven-fold cross-validation and guaranteed that this model is satisfactory.

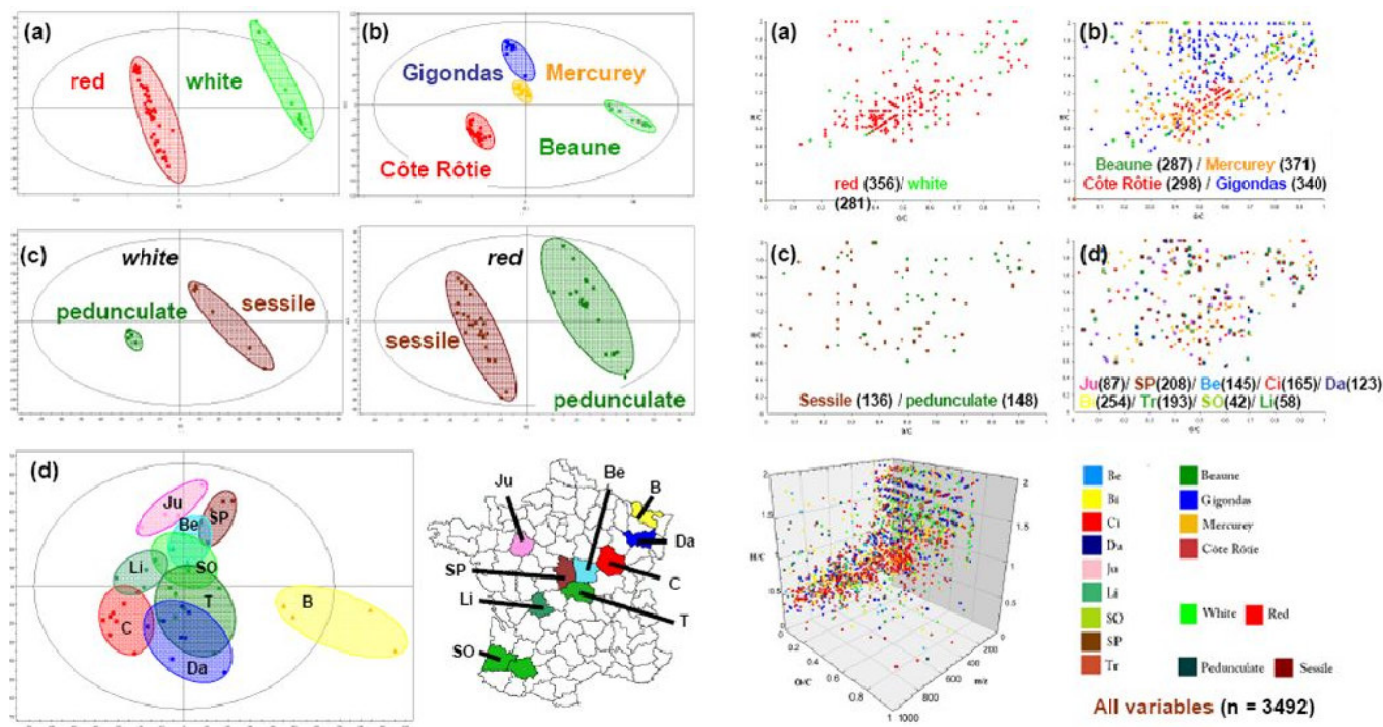


Figure 6.4: a to d left: PLS-DA score plots; Classes are (a) white (■) and red (■) wines; (b) Gigondas (■), Mercurey (■), Beaune (■), and Côte Rôtie (■) wines; (c) White and red wines aged in Sessile (■) and Pedunculate (■) barrels; (d) wines sorted according to forests of origin of oaks of barrels they were aged in, regardless of the species: (Ju) Jupilles, (SP) Saint Palais, (Be) Bertrange, (Li) Limousin, (SO) Sud Ouest, (Tr) Tronçais, (Ci) Citeaux, (Da) Darney, (Bi) Bitsch, along with their location on the map of France; a to d right: van Krevelen representations of discriminating masses (highest correlation coefficients) for the different classes shown on the left side; e: 3-D van Krevelen diagram (H/C vs O/C vs m/z) representing together all of the discriminating mass.

For instance, within the 150-2000 m/z range explored, 356 signals corresponding to unique CHO formulae were found representative for red wines, whereas 281 signals were found representative for white wine. From the two-dimensional van Krevelen representation of the corresponding signals, it can be seen that anthocyanins (O/C region between 0.4 and 0.6, and H/C region around 1.0) obviously discriminate red wines (see figure 6.4a). However, several other signals in the regions of fatty acids, amino acids or carbohydrates are also discriminant. Similarly, several hundreds of peaks were selectively observed in each of the four wines of this study (see figure 6.4b), which led to a clear discrimination of the wines according to their geographical origin or to the variety of the grape they were elaborated from (see figure 6.4b). Interestingly, Burgundy red wines from Mercurey appeared to differ more from their neighbours from Côte Rotie (Côte du Rhône North) than from the southern wines of Gigondas. Analytical discriminations of wines, based on their colours or on the grapes varieties have already been largely reported (Vogels, et al., 1993), and our results clearly appear to complement them in terms of the chemical diversity that is responsible for these discriminations. Figure 6.4b indeed shows for instance, that signals specific to Gigondas wines (made of Grenache grapes) span from the lower left corner of the van Krevelen diagram (O/C about 0.2 and H/C about 0.6) up to the upper right corner, indicating that molecules ranging from rather condensed weakly oxygenated to saturated fully oxygenated are specific to these wines. Similarly, the possibility to discriminate wines according to the oak species of the barrels they were aged in (see figure 6.4c) has already been demonstrated, with the particular identification of significantly higher amounts of aromatic whiskylactones in wines aged in European sessile or American white oak barrels (Wu, et al., 2004), (Waterhouse, et al., 1994). However, the PLS-DA score plots for both red and white wines again provide an enhanced representation of how wines aged in barrels from a given wood species are grouped together (see figure 6.4c). Most interestingly, these results show a significantly narrower distribution among white wines aged in pedonculate barrels than among those aged in sessile barrels, whereas no such difference in distribution is observed for red wines. For white wines, these findings corroborate the previously observed narrower distribution among pedonculate oak wood extracts than among sessile oak extracts (Gougeon, submitted). In contrast, the broader distribution among red wines aged in pedonculate barrels witnesses to the multiple - yet to discover - products of the possible reactions between ellagitannins and wine nucleophiles such as polyphenolic compounds characteristic of red wines (Quideau, et al., 2005). The major outcome of this non targeted approach is the previously unavailable

opportunity to discriminate wines according to the forest origin of the oaks used for barrel aging of these wines, and to provide a significance in terms of related chemical spaces (see figure 6d), regardless of the colour, the origin of production (and grape variety), and the barrel oak species. The three-dimensional van Krevelen representation (see figure 6.4d) of the cumulated 3492 discriminant signals, to which unique CHO-based chemical formulae could be assigned, illustrates the chemical diversity which is responsible for this discrimination. Within, the 150-1000 m/z mass range, a major part of the discriminating signals correspond to masses lower than 400 Dalton. However, discrimination of all the forests is specifically associated with molecule masses up to 1000, with both decreased O/C and H/C elemental ratios of the corresponding molecular formulae at higher masses (see figure 6.4d). Low H/C values at high m/z values could be associated with condensed - possibly aromatic - structures such as derived from native lignols (see figure 2.16b chapter 2), subject to further condensation reactions during the toasting process of staves. Hence, our approach not only allows to integrate the intrinsic cooperage variability arising from the fact that all of the staves and barrels of the "Tonnellerie 2000" experiment did not undergo the same drying procedure and were not made by the same cooper (Feuillat, 2003), but also illustrates that the different steps of elaboration of barrels can complement the chemical signature of a given forest without necessarily erasing it (Mosedale, et al., 1996).

## 6.3 Methods

**Tonnellerie 2000 samples:** The detailed procedure followed to select trees has already been described elsewhere (Feuillat, et al., 1999), and in Supplementary Methods. In brief, twelve lots (5 pedunculate and 7 sessile) of 24 trees were selected from nine French forests. To one lot of 24 trees corresponded one barrel. Each barrel has thus been assembled from 24 trees which stood each for 1/24<sup>th</sup> of the toasted surface (body) and 1/24<sup>th</sup> of the untoasted surface (head and bottom). These twelve barrels (representing twelve identified forest/species couples) were used for wine ageing experiments with a red Pinot noir wine from Mercurey, a white Chardonnay wine from Beaune, a red Syrah wine from Côte Rotie and a red Grenache wine from Gigondas. For a given wine, we had therefore

twelve bottles (or 24 if duplicates) which only differed by the species and the forest origin of oaks used for barrel ageing.

**Grape extracts, yeast fermentation medium, tartar precipitate, oak wood extracts:** All of the samples were obtained with extraction or dilution in methanol.

**ICR-FT/MS analysis:** High-resolution mass spectra for molecular formula assignment were acquired on a Bruker (Bremen, Germany) APEX Qe Ion Cyclotron Resonance-Fourier Transform Mass Spectrometer (ICR-FT/MS) equipped with a 12 Tesla superconducting magnet and a APOLO II ESI source in the negative ionisation mode. Samples (typically 20 to 50  $\mu\text{l}$  diluted into 1 ml methanol) were introduced into the microelectrospray source at a flow rate of 120  $\mu\text{l}/\text{h}$  with a nebuliser gas pressure of 20 psi and a drying gas pressure of 15 psi (200 °C). Other details are given in Supplementary Methods.

High-resolution mass spectra for molecular formula assignment were acquired on a Bruker (Bremen, Germany) APEX Qe Ion Cyclotron Resonance-Fourier Transform Mass Spectrometer (ICR-FT/MS) equipped with a 12 Tesla superconducting magnet and a APOLO II ESI source in the negative ionisation mode. Samples were introduced into the microelectrospray source at a flow rate of 120  $\mu\text{l}/\text{h}$  with a nebuliser gas pressure of 20 psi and a drying gas pressure of 15 psi (200 °C). Spectra were externally calibrated on clusters of arginine (10mg/l in methanol) and accuracy reached values lower than 0.1 ppm in day to day measurements. Further internal calibration was done for each sample using fatty acids and accuracy reached values lower than 0.05 ppm. The spectra were acquired with a time domain of 1 megaword (4 megaword for selected samples) with a mass range of 100-2000 m/z. The spectra were zero filled to a processing size of 2 megawords and an average resolution of

250.000 was reached at m/z 200 (100.000 at respectively m/z 600) in full scan. Before Fourier transformation of the time-domain transient, a sine apodization was performed. The ion accumulation time in the ion source was set to 0.2 s for each scan. 1024 scans were accumulated per samples.

### 6.3.1 Statistical analyses

Raw data (mass spectra) were normalised, and then transformed into variables which were further mean centered, Pareto scaled and represented as an X matrix<sup>50</sup> for further processing.

Raw data (mass spectra) were normalised, and then transformed to  $\log(X + 0.00001)$ . The constant 0.00001 was added to provide non-detectable components with a small non zero value (Sjödin, et al., 1989). Transformed variables were then mean centered and Pareto scaled and represented as an X matrix. Pareto scaling gives each variable a variance equal to its standard deviation by dividing by the square root of the standard deviation of each column (see chapter 3.3). The sample classification and the prior information about the sample were done using the Hierarchical clustering analysis (HCA) unsupervised method. On the other hand, partial least square - discriminative analysis (PLS-DA), performed with SIMCA 11.5, was used to discover characteristic biomarkers (Wold, et al., 2006). This multivariate procedure provided bioinformatics clues for the selection of a limited number of masses most effective in discriminating different species and forests.

The primary advantage of using targeted profiling as an input to PLS-DA is generating variables that represent combinations of measured metabolites concentrations. Positive regression coefficients indicate a relatively greater concentration of the considered metabolites with respect to the others, whereas negative values indicate a relatively lower concentration with respect to the other samples-classes (Rossello'-Mora, 2008). As such, these variables are easier to interpret as factors in the underlying classification model. Thus, targeted profiling provides meaningful and interpretable factors describing the input data. PLS-DA is a regression extension of PCA that takes advantage of class information to attempt to maximize the separation between groups of observations.

The feature selection procedure comprises two steps: i) identification of those masses that best describe each classes (a list based on the modelling power of the original variables), ii) scoring and ranking of the variables in every class-related list according to their abilities to discriminate the class they model from all other categories. The ranking and score take place after computation of the minimum number of masses through the formula generator (in-house code written in FORTRAN). The generated formulas were validated by setting sensible chemical constraints (N rule, O/C ratio  $\leq 1$ , H/C ratio  $\leq 2n + 2$ , element counts: C  $\leq 100$ , O  $\leq 80$ , N  $\leq 5$ , S  $\leq 1$ ) and only the masses in conjunction with their generated theoretical  $^{13}\text{C}$ -isotope patterns were taken into consideration.



## 6.4 Discussion and conclusions

For the analytical chemist, wine is the complex mixture of water, ethanol, and countless compounds which represent less than 5% of the composition, but which actually govern its identity. For the wine maker and the chemist gourmet, altogether these compounds are gathered to form a delicate equilibrium, which confers its flavour, aroma, colour, stability, and aptitude for ageing to wine. Recent studies also indicate that this equilibrium is likely to define particular therapeutic activities of wine (Corder, 2006). We believe that our results represent a great step towards a more holistic overview (Dixon, 2006) of this unique beverage. Our findings show that approaches aiming at the most comprehensive representations of wine through its particular chemical spaces, considerably enhance the opportunities of discriminating metabolites related to distinct environmental modifications and their impacts on organoleptic or therapeutic activities. In the particular case of barrel ageing, this study reveals that even after several years in a bottle, a wine can still express a chemical imprinting of the forest where the oaks of the barrel have grown. As such, our systems oenology approach provides an unprecedented example of metaboledgeography (Green, et al., 2008) translated into the chemical representation of the way such noble nectar can shape on the papillas of the wine taster some of the outlines of the scene of its birth.

Through the means made available by this study, we envision general value and applicability of this non-targeted molecular level traceability for purposes as diverse as wine or other beverages, forensics assessments on European or larger scale levels, improvement of the nutritional value and sensory properties, understanding of mechanisms responsible for undesired evolutions (untimely oxydation) or even recording of environmental changes such as climatic modifications over decades.



# Chapter 7

## 7 METABOLOMICS APPROCH IN HEALTH

### 7.1 Introduction

The metabolomics study, applied to the health evaluation, has the main goal to diagnose diseases and identify factors that cause them. These studies can enhance the understanding of disease mechanisms of drug or xenobiotic effect and can lead to new diagnostic markers. Thus, this approach allows increasing the ability to predict individual variation in drug response phenotypes (Kaddurah-Daouk, et al., 2008).

The final result will be the definition of a list of biomarkers, which is a list of relevant masses that measure or indicate the effects or progress of pathology. Their biological trend is influenced by many environmental factors. Initial metabolomic signatures have already been reported for several disease states,

including motor neuron disease (Rozen, et al., 2005), depression (Paige, et al., 2006), schizophrenia (Holmes, et al., 2006), (Van Der Greef, et al., 2007), Alzheimer disease (Han, et al., 2002), cardiovascular and coronary artery disease (Sabatine, et al., 2005), (Brindle, et al., 2002), hypertension (Brindle, et al., 2003), subarachnoid hemorrhage (Dunne, et al., 2005), preeclampsia (Kenny, et al., 2005), type 2 diabetes (Van Der Greef, et al., 2007), (Wang, et al., 2005), (Yang, et al., 2004), liver cancer (Yang, et al., 2004), ovarian cancer (Odunsi, et al., 2005), breast cancer (Fan, et al., 2005), and Huntington's disease (Underwood, et al., 2006).

In this respect the metabolomics field has enormous potential to improve human health in a number of ways listed here below:

- prognostics of risk of disease or diagnose disease
- determination whether a treatment is working or not
- monitor healthy people to reveal early signs of disease
- information about mechanisms of disease

Two different surveys are presented here: the first one dealing with exhaled breath condensates (EBC), as non invasive tool to study the pulmonary diseases, and the second one dealing with plasma to study the pre-diabetic state in the frame of the TULIP study, a project in collaboration with the University of Tuebingen.

These studies will facilitate a range of integrated profiling analysis, improving the understanding of disorder mechanisms and develop new diagnostic, prognostic and monitoring strategies in the areas of obesity, diabetes and pulmonary disease.

## 7.2 Metabolomic analysis of exhaled breath condensate for smokers, no-smokers COPD patients with ICR-FT/MS

Exhaled breath condensate (EBC) is a noninvasive method to collect samples in relationship with the airways and the lungs. In EBC samples, a large number of

mediators including adenosine, ammonia, hydrogen peroxide, isoprostanes, leukotrienes, nitrogen oxides, peptides and cytokines were identified and analyzed (Horváth, et al., 2005). The concentration levels of these mediators varied by lung diseases and can be modulated by therapeutic interventions. Similarly, the pH - value of EBC can also change in respiratory diseases. Other publications inform about important molecules dealing with lung/respiratory disorders (Bloemen, et al., 2007), (Kharitonov, et al., 2001) and (targeted) analysis of these mediators, but all methods were limited to few components, using component selective methods including for example immuno- and bioassay methods.

The ESI ICR-FT/MS technique used in our laboratory opens the possibility to measure thousands of individual molecules, present in the samples simultaneously, with exact molecular weight, with semi-quantitative intensity, allowing the use of metabolomic analysis.

EBC-samples were collected at Helmholtz Zentrum Muenchen, Institute of Inhalation Biology, with a commercial instrument Ecosacreen-2 (Filt GmbH, Germany). This system gained two separated samples, from the patients: the first part (150 ml) of the exhaled air was collected separately as “bronchoalveolar” (AW) sample representing the upper airways, and the second part as the “alveolar” (AV) sample. The samples taken were stored by minus 20°C in Eppendorf vials. The defrosted samples are centrifuged by 30.000 rpm for 15 min. 50 µl transferred in an another Eppendorf vial, and 50 µl MeOH with 0.2% formic acid was added. After the solution was homogenized and the samples were transferred into 96 well probe holder of the -chip- Nano ESI instrument (Advion TriVersa NanoMate, Advion BioSciences, Inc, 19 Brown Road, Ithaca, NY 14850 USA).

Broad scan mass spectra were acquired on a Bruker (Bremen, Germany) APEX Qe ICR-FT/MS with 12 T superconducting magnet and an Apollo I electrospray (ESI) source, whereas high-resolution spectra were acquired with an Apollo II ESI source in positive mode. Spectra were externally calibrated on clusters of arginine (10 mg/l<sub>1</sub> in methanol), and calibration errors in the relevant mass ranges were always below 100 ppb, which is the prerequisite for an adequate elementary composition assignment. The spectra were acquired with a time domain of 1 MW and with a mass range of 150-2000 m/z. A sine apodization was performed before Fourier transformation of the time-domain transient. The ion accumulation time in the ion source was set to 0.2 s and 1024 scans were accumulated for the samples.

Overall 60 samples were investigated, which are divided in smokers (28 samples, 6 of them with chronic obstructive pulmonary disease (COPD)), former

smokers (6 samples, all of them with COPD) and no-smokers (26 samples, 2 with COPD). From the literature it is known that cigarette smoking reduces life span by an average of 7 years, and tobacco consumption accounts for a shortening of disease free life by 14 years (Bernhard, et al., 2006). The exact mechanisms by which smoking causes disease and death are generally not well understood, but evidence continues to mount that cigarette smoking exhausts cellular defense and repair functions, leading to an accumulation of damage e.g. mutations and malfunctioning proteins. Here we investigate the phenomenon in the view of the metabolites-changing.

### 7.2.1 Statistical elaboration

Spectra were exported to a peak list at different levels of S/N. According to these different ratios, a possible approach for standard analysis method was set up (see chapter 3.2). The first step is to fix the level of S/N equal to 1, subsequently the data are processed and submitted to the formula calculation program. Through this software tool we obtain only the realistic masses (see chapter 3.2.1). Mass signals at S/N=1 could be attributed to distinct elementary compositions, containing the elements C, H, O, N and S.

Only the  $^{13}\text{C}$  validated peaks after the submission of formula calculator are used for further analysis. The generated formula were validated by setting sensible chemical constraints (nitrogen rule, atomic oxygen to carbon ratio  $\text{O}/\text{C}<1$ , atomic hydrogen to carbon ratio  $\text{H}/\text{C}<(2n+2)$ , element counts: carbon  $\text{C}<100$ , oxygen  $\text{O}<80$ , nitrogen  $\text{N}<5$  and sulphur  $\text{S}<1$ ) and only the masses in conjunction with their automated generated theoretical isotope pattern (existence of the  $^{13}\text{C}$  isotope) were taken into consideration (Hertkorn, et al., 2007).

The results obtained with S/N=1 were compared to those obtained with S/N=3, in which the formula calculator is used only at the end of the process, when a list of molecule of interest (see chapter 3.2) was isolated. Specifically the first method, with S/N=1, was considered better because it eliminates redundant information at first (with the formula calculation). This gives discriminative masses also with lower abundances and speeds up all informative process, at least it does not exclude low mass molecules of potential interest.

All spectra are aligned using the software written in-house (see chapter 3.2). The measurements are arranged into a data table where each row constitutes an

observation (in our case mass spectra) and the columns represent the variables (m/z).

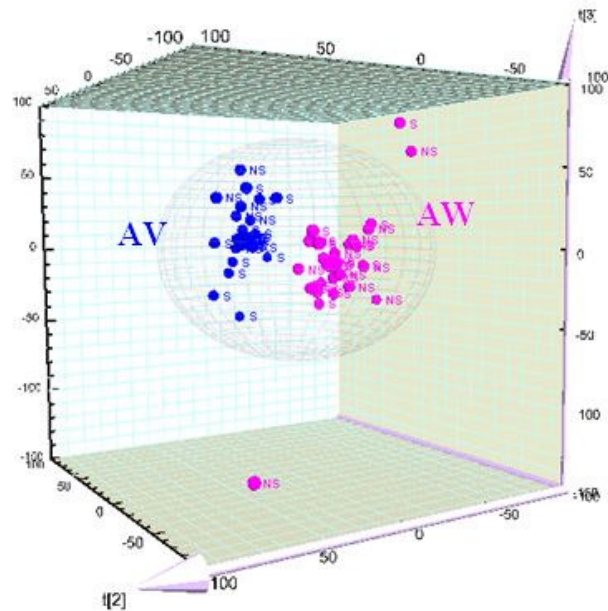
Once we have obtained the matrix, we provide to give a corresponding representation of the variable, which are hard to summarize and visualize without appropriate tools. Using the Chemometrics tools (see chapter 3), it is important to include efficient, validated, and robust methods for modeling chemical and biological data.

The data that are submitted to the statistical evaluation are transformed according to the methods reported in chapter 3.3. Logarithmic transformation ( $\log_{10}$ ) was found to be beneficial prior to multivariate analysis. Data transformation is suitable to reduce the effect of peaks with high abundance and the no-constant variance of the noise (heteroscedasticity of noise structure). For the presence of null values a constant (0.00005) is normally included before the log transformation.

For this particular data we choose Pareto scaling, which provides more flexibility in data analysis. In particular it reduces the relative importance of large values, but keeps data structure partially intact.

To detect differences among the different groups of samples we use multivariate technique as well as partial least-squares discriminant analysis (PLS-DA). The first result of the analysis gives a list of significance masses which are possible biomarker candidates, characteristic of the groups. From the analysis are excluded the samples which were proved to be contaminated. Before analyzing the difference between smokers and no-smokers, we divided the samples into AV and AW samples (see figure 7.1).

A)



B)

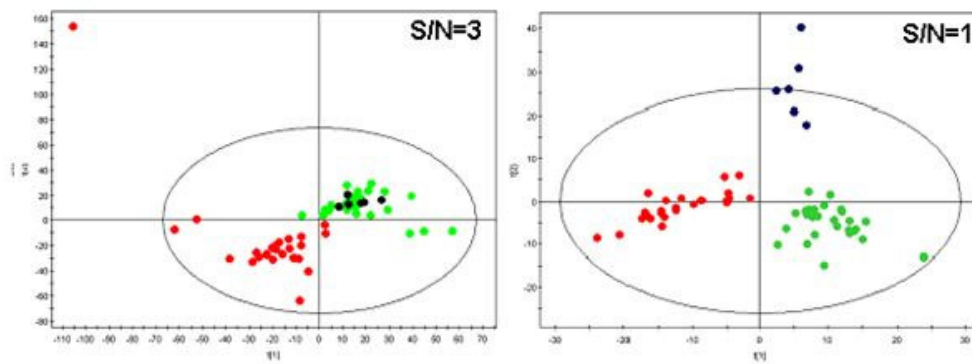


Figure 7.1: (A) Score Plot representing the differentiation between AW (●) vs AV (●) ( $Q^2=0.74$ ,  $R^2(Y)=0.98$ ), internal of both groups is the presence of smoker and no-smoker samples with COPD disease. (B) Differentiation of the score plot of the sample extracted at  $S/N=1$  and  $S/N=3$ .  $S/N=3$  has a valid model only with two classes (● = no smokers; ● = smokers).



= smokers), former smoker (●) are not identified; instead with S/N=1 we could also differentiate the former smokers. This must be attributed to the fact that the masses valid with low intensity are not discarded like it is happened in S/N=3.

Based on the discriminative masses we obtain pathways that confirm the presence of differentiation at the level of AV and AW (see figure 7.2).

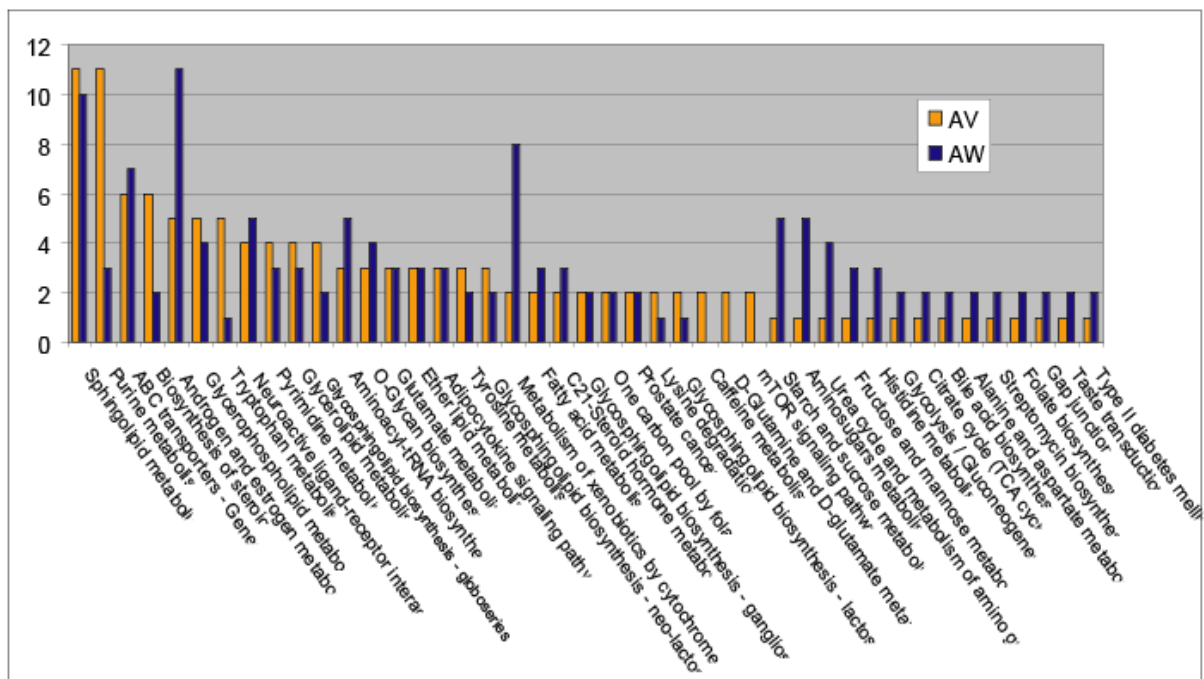


Figure 7.2: Pathways characteristic for AW samples and AV samples.

At S/N=1 (see figure 7.1 b) the former smokers are classified as independent group with its own characteristics. Instead using the S/N=3 ratio, this group exhibits the same characteristics of smokers. According to this similarity and since former smokers are not significant in number, a valid model is developed considering only two classes merging former smokers and smokers against no-

smokers. The PLS-DA model was validated with regard to fitted  $R^2$  and cross-validated  $Q^2$  values (Eriksson, et al., 1997), ( $Q^2=0.48$  and  $R^2(Y)=0.98$ ), where these two indices reassume the validity and the predicative capability of the model. A model is considered acceptable for biological data if:  $R^2>0.7$  and  $Q^2>0.4$  (Lundstedt, 1998).

In addition to cross-validation, the model is also validated using permutation validation (using 100 validation rounds, (Eriksson, et al., 1997)). The permutation validation gives  $R^2$  and  $Q^2$  intercepts (see chapter 3.6). The model is valid being  $Q^2$  value below zero and the estimated  $R^2$  value considerably smaller than the  $R^2$  value of the model (Eriksson, et al., 2004), (Eriksson, et al., 1997). Moreover,  $Q^2$  is used to determine how many PLS latent variables should be included in the PLS models (Wold, 1978).

Once we found the list of biomarkers, we submitted it directly to the MasSTRIX to find pathways and investigate directly the compounds with the KEGG database, adding any additional genomic or transcriptomics information by highlighting the corresponding enzyme boxes (see figure 7.3), (Suhre, et al., 2008).

From this list we selected a list of biomarker candidates, responsible for the differentiation of smokers and no-smokers (see table 7.1). The molecule: C19H28O2 is pointed out. One possible assignment is to the testosterone. The natural level of testosterone, in blood, ranges 9-30  $\mu\text{m}/\text{l}$ . This could be detected by the instrument but other investigation will be necessary to confirm this finding. The inspection of the biomarkers can also serve as indicators of disease progression.

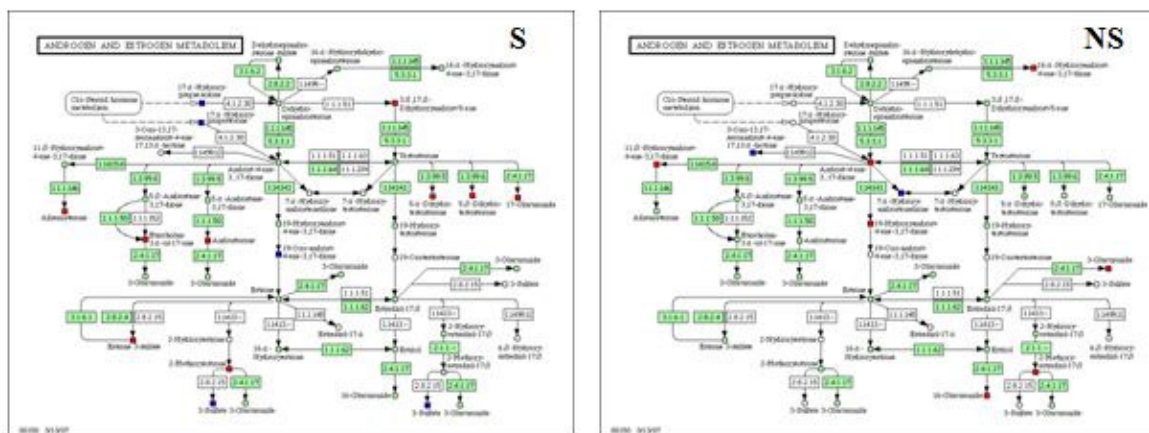


Figure 7.3: An example of pathways: “Androgen and estrogen metabolism”, it reveals different probable compounds in smokers samples compared to no smokers group.

METABOLITE	Significance probable compounds in the Smoker
ABC transporters - General	Glutamine, D-Allose Fructose
Aminoacyl-tRNA biosynthesis	L-glutamine, L-Tyrosine
Aminosugars metabolism	
Androgen and estrogen metabolism	5beta and 5alfa Dihydrotestosterone, 17 Glucuronide, testosterone
Arachidonic acid metabolism	Arachidonate
Biosynthesis of steroids (characterize also AL)	
C21-Steroid hormone metabolism	Cortisone, Urocortisol, Progesterone
Caffeine (characterize also AL)	
Folate biosynthesis	
Galactose metabolism	
Glycerophospholipid metabolism (characterize also AL)	
Glycosphingolipid biosynthesis - ganglioseries (characterize also AL)	
Metabolism of xenobiotics by cytochrome P450	
Neuroactive ligand-receptor interaction	
Purine metabolism (characterize also AL)	
Pyrimidine metabolism (characterize also AL)	
Sphingolipid metabolism (characterize also AL)	
Starch and sucrose metabolism	D-Glucose, D-Fructose
Tryptophan metabolism (characterize also AL)	
Tyrosine metabolism (characterize also AL)	Tyrosine
Urea cycle and metabolism of amino groups	

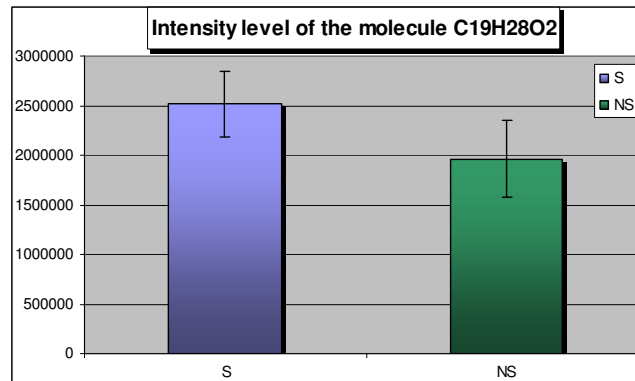


Table 7.1: List of metabolites responsible for the differentiation of the smokers in confront to the no-smokers with the correspondent list of possible compounds present in smokers people, (to the list of masses, selected with the high regression coefficient, is applied the t-test to select the most representative, with a level of significance of  $p < .05$ ). In the bottom there is the graphic of the level of annotate and probable testosterone in smoker vs no-smokers. In same instance it should be interesting to monitor changes in this metabolite.

### 7.3 Metabolomic analysis of plasma of pre-diabetic patients with various insulin resistant index values

Diabetes is a condition with a long “silent or asymptomatic period where the patient is not aware of the disease but complications gradually develop. At the time of clinical diagnosis as many as 20-30% have microvascular complications and diabetes is often diagnosed after retinal or renal problems are clinically overt (Yoon, et al., 2006). Furthermore, diabetes, particularly type 2 diabetes, is frequently diagnosed in intensive care units in patients with acute myocardial infarction (Maki, et al., 1995) indicating that also macrovascular complications develop clinically silent during the undiagnosed period. The affected individuals experience greatly elevated morbidity and mortality from nearly all of the common macrovascular diseases (e.g. myocardial infarction and stroke), and in overt diabetes from diabetic late complications (e.g. nephropathy, blindness, and neuropathy). Important to note, the pre-diabetic state precedes the manifestation of overt type 2 diabetes for decades (DeFronzo, 2004), (Eckel, et al., 2005). However, the impairment of insulin sensitivity and the development of type 2 diabetes can be retarded and even prevented by therapeutic and/or lifestyle interventions, which was demonstrated in several recent studies (Knowler, et al., 2002), (Tuomilehto, et al., 2001), (Schafer, et al., 2007), but currently only little is known about the multiple metabolic alterations reflecting these subtle abnormalities in asymptomatic individuals.

In order to gain new insights in complex metabolic processes, non-selective but specific information-rich analytical approaches are required. Metabolomics is the

non-targeted analysis of metabolites typically carried out to generate a specific fingerprint of a current metabolic state at a given time point of the metabolic pattern of an organism (Lindon, et al., 2004), (Lu, et al., 2008), (Lenz, et al., 2007). It is a rapidly advancing field that complements genomics and proteomics, promised to add significant information to the understanding of physiological and pathophysiological processes (Lindon, et al., 2004), (Gross, et al., 2007). The very complex data are evaluated by pattern recognition techniques, i.e. multivariate statistic methods (Eriksson, et al., 2004), (Jonsson, et al., 2005). Furthermore, metabolomic investigations have the potential to identify molecular species differentiating physiological states (Chen, et al., 2008). Thus, mass spectra of biofluids serve in two distinct but closely related modes: as a metabolic fingerprinting tool and as means of metabolite biomarker elucidation.

The aim of our study was to investigate for the first time the metabolic pattern in plasma of individuals at high risk to develop type 2 diabetes by a metabolomics approach to detect the conversion from the physiological to the pathological metabolic state by an individual metabolic fingerprint. Furthermore, this let us to elucidate pathways and to discover metabolite biomarkers altered in the pathogenesis of insulin resistant. Thereby opening new perspectives in the study on the pathogenesis of this epidemic metabolic disease.

### 7.3.1 Pre-Diabetic state definition

The pre-diabetic state, which precedes overt type 2 diabetes for decades, results in multiple metabolic alterations in insulin sensitive target tissues like liver, fat and skeletal muscle. Nevertheless, the identities of biomarkers for such changes are largely unknown. Applying metabolomics, a non-targeted top down approach, we aimed to investigate these specific metabolic traits indicated by the identification of metabolites of altered pathways in plasma of individuals at high risk to develop type 2 diabetes.

All individuals underwent a 75 g oGTT according to the recommendations of the WHO/IDF (WHO, definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of WHO/IDF Consultation. Ed. World Health Organization. Geneva: WHO Press, 2006, 1-46) to determine the insulin sensitivity index (ISI). Venous blood samples were obtained at 0, 30, 60, 90 and 120 minutes for determination of plasma glucose and insulin. Insulin sensitivity was calculated from glucose and insulin values during the oGTT as proposed by Matsuda and DeFronzo (Matsuda, et al., 1999) using the formula:

$$ISI = \frac{10,000}{\sqrt{(FGP \cdot FPI) \cdot (\bar{x} \text{ oGTT gluc. concentration} \cdot \bar{x} \text{ oGTT insulin concentration})}}$$

where FGP = fasting plasma glucose; FPI = fasting plasma insulin concentration;  $\bar{x}$  oGTT gluc. concentration=average of glucose concentration during the oGTT;  $\bar{x}$  oGTT insulin concentration=average insulin concentration during the oGTT. Low levels of the ISI indicate that the body is more resistant to insulin action. Where insulin resistant has a value <8.5 and insulin sensitive has a value >8.5.

## 7.4 Data analysis

The non-target analytical approach was applied also to a set of 47 non-diabetic individuals but with a high risk to develop type 2 diabetes.

Bioinformatics data evaluation, analyzing differences in individual pattern by multivariate analysis, revealed three clusters representing distinct metabolic plasma pattern:

- insulin sensitive individuals
- insulin resistant subjects
- a “transition” group between insulin sensitive individuals and insulin resistant subjects

Following the above result, the major goals of our investigations were, firstly to elucidate alterations in metabolic pathways to further understand the pathophysiological changes in the transition-group as well as the insulin resistant state (pathobiochemicak aspect), and secondly identifying within these pathways distinct metabolite biomarkers showing significant differences between the different groups (diagnostic aspect).

The statistical elaborations were done on the dataset calculated at 1 ppm.

For the typology of data and the goals to achieve, a PLS model was developed using as dependent variable Y=ISI (Insulin Sensitivity Indices). The strong outlier (it

is the sample 1051\_HL with a value of ISI=28.65) from the initial model was not included. The model is represented in figure 7.4 with its validation (100 row permutations, it is represented in figure 7.4 a). The internal validation was acquired by randomizing the positions of the Y data in relation to their corresponding rows in the dataset and observing the effect of that randomization on the  $R^2$  and  $Q^2$  values. If the original model was confirmed, randomization of the Y data would be expected to considerably reduce  $Q^2$  (see chapter 3.6). The model has three valid components resulting in:  $Q^2(\text{cum})=0.90$  and  $R^2(Y)=0.94$  (see figure 7.4a).

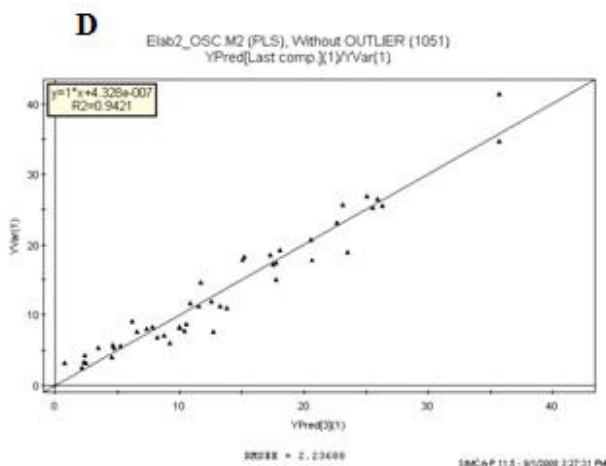
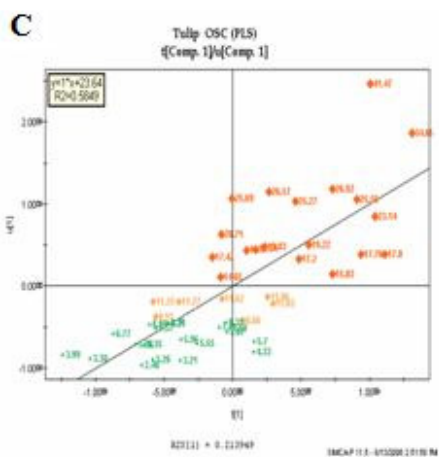
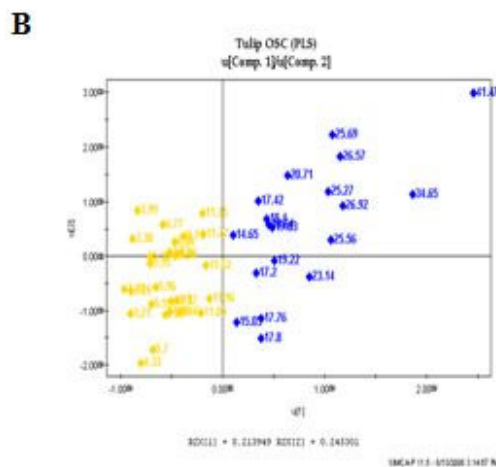
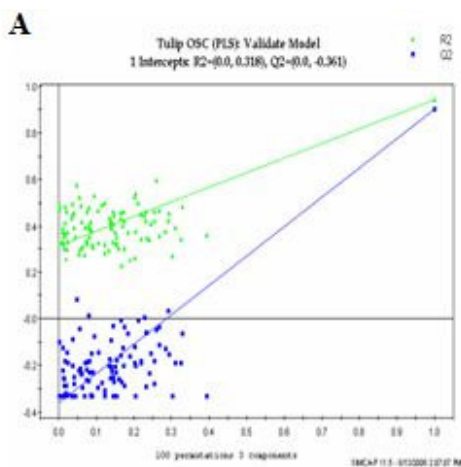


Figure 7.4: OSC PLS Validation. The validation presents good values for  $Q^2$  and  $R^2$ , within the permitted limits. B) and C) are the score scatters plot  $u[1]u[2]$  and  $t[1]u[1]$ , where the relation between the first summary of X variables and the Y underlines a spread of the data more enhanced in the samples with a very high value of ISI. It will be improved with the use of the OPLS-DA model and with the re-assignment of the classes (see figure 7.4b). The score scatter plot  $u[1]u[2]$  reveals the separation of two groups (the first one with a low value of ISI the other one with a high ISI level). The PLA model uses 3 valid components. D) The last plot describes the relation between the observed vs the predicted values. The prediction is particularly good.

The interpretation of the model is easier if examined through the score scatter plot  $u[1]u[2]$ , (see figure 7.4b). Once investigated, we built up several new models in order to investigate deeper the limit of ISI index, taking care of the transit region of ISI, ranging from 8.66 and 14.65.

At first an OPLS-DA model (see figure 7.5a) with two groups was developed (Wold, et al., 1998), which reflect the standard classification of the ISI value:  $>8.5$  (first group) and  $<8.5$  (second group). This reveals that the transition group seems to have the characteristics similar to the group with a low ISI level. The transition samples are validated as the samples with an ISI value ranging from 8.66 to 14.65. We re-classify the samples of the transition region and the new OPLS-DA model which is more consistent (see figure 7.5b). The built model was an OPLS with two classes showing the value of  $Q^2$  and  $R^2(Y)$  being respectively 0.94 and 0.89

With the new re-classification of seven samples, only one has a low probability to belong to the ISI low level group (Probability=0.2), the other samples are ranging between  $Pr=0.6$  and 0.9.

However to prove this, the samples belonging to the transition region are investigated separately in the pathway (revealing and confirming characteristics more similar to the ISI low level group).

The sensitivity and specificity values, obtained using soft independent modeling of class analogy (SIMCA), are 100%, for both classes. This approach, used in (Van Der Greef, et al., 2007), allows one to build a class model by a distribution of probability (namely the PLS response). In our sensory modeling problem,  $\alpha$  is the probability of false non-low ISI level and  $\beta$  the probability of false low ISI level (see chapter 3.5).



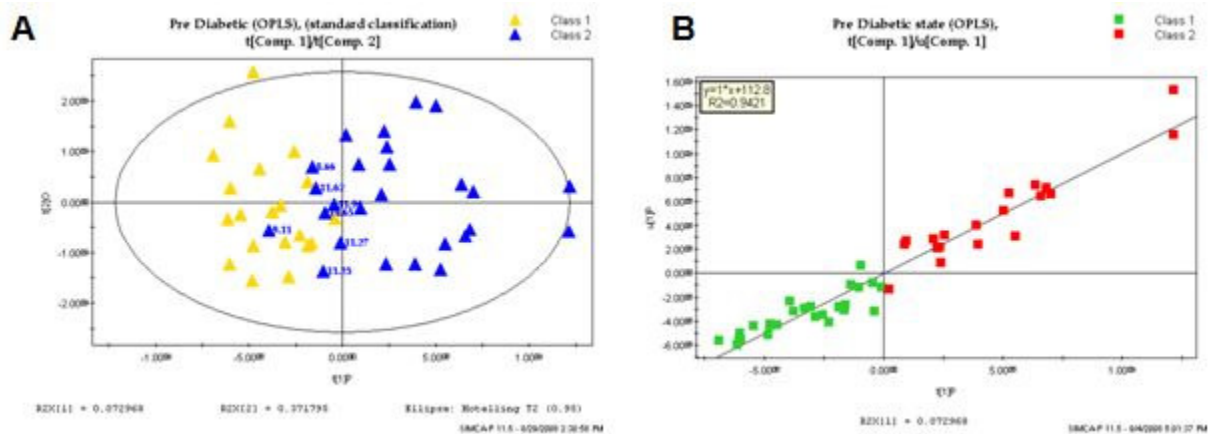


Figure 7.5: A) OPLS-DA analysis: Score scatter plot with the conventional value of ISI (low and high). The last step was to define the list of masses characteristic for the low and high value of ISI. From the inspection of the figure 6.6 we can see the most relevant masses, known and unknown compounds will forming this list. B) OPLS-DA analysis with re-classification of the samples: with the re-classification of the spectra (the samples with a value of ISI ranging from [8.5; 14.65] are re-classified as low ISI level), we reach a very good relationship in confront with the first PLS model, because the last OPLS-DA model has rotated the solution to put the all Y-related variation into the first component.

At the end of the statistical analysis a PLS-DA was developed without the transition group making a prevision and a validation of all analysis.

Partial least squares method was used as discriminant (Sjöström, et al., 1986) because of the existence of collinearity between the variables of the measurement space. The use of PLS-DA is justified because of its analogy to the regression models corresponding to the theoretically and statistically well-known test for the discrimination between two classes (Stahle, et al., 1987). González-Arjona et al. have described a detailed review on PLS-DA and stated the equivalence between PLS-DA and procureses discriminant analysis (González-Arjona, et al., 1999). Barrer and Rayens point out the mathematical structure of PLS-DA showing its theoretical relation to canonical correlation analysis. A similar result is stated by Nocairi (Nocairi, et al., 2005). The masses that differentiate the

two groups (see figure 7.6) are investigated with the MasSTRIX (the masses came from last OPLS-DA model). From a list of 4500 m/z, characteristic for the high ISI level, only 12% were assigned with MasSTRIX. The others are still unknown. Also from the m/z characteristic for the group of samples belonging to the low ISI level, the percent of assigned masses are increased to 25%.

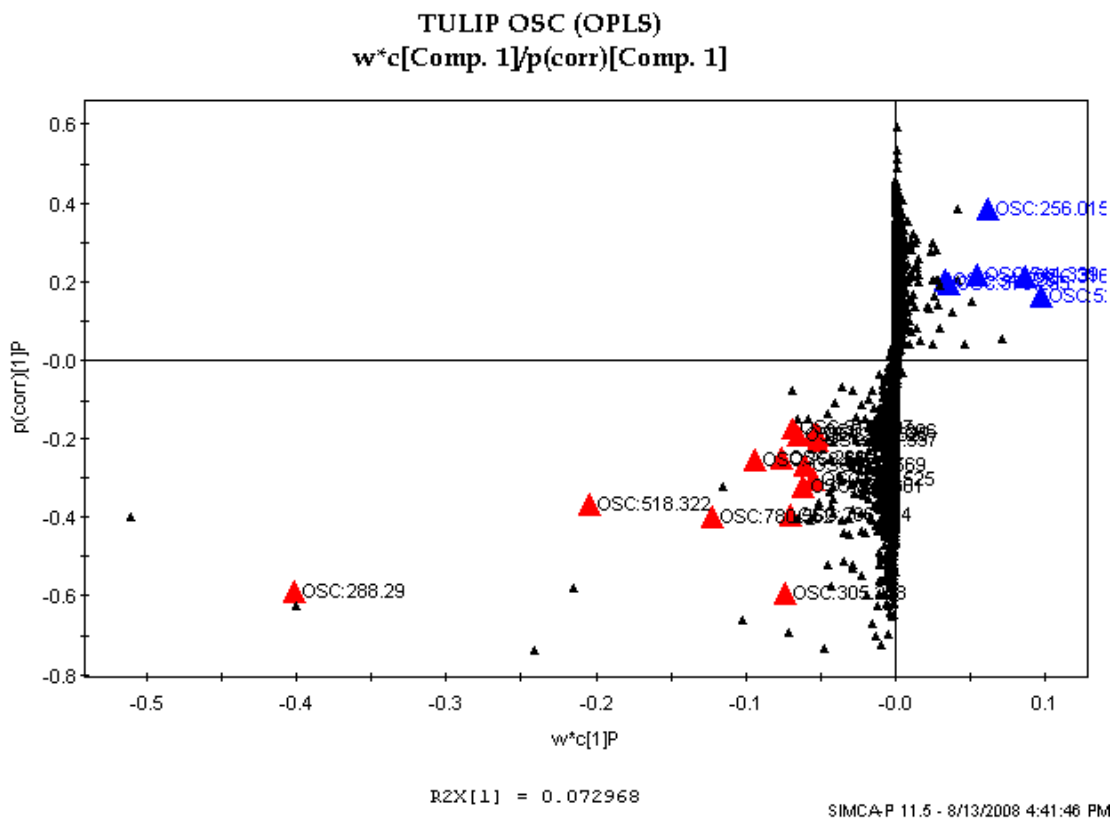


Figure 7.6: Loading plot with the most significant masses, in red and blue triangle the masses which were found with MasTrix. For the statistic analysis these play an important role but in MasTrix are sometimes difficult to give a plausible meaning. Many efforts must be done also in this direction.

The list of compound identified and present in the organism, are listed in table 7.2. They are different between the two main groups. We must consider that a single mass peak may be annotated by more than one metabolite, either if different structures (isomers) with the same sum formula exist, or if two compounds lie within the error range.

<b>LOW ISI Level</b>	<b>IC*</b>	<b>HIGH ISI Level</b>	<b>IC</b>
Arachidonic acid metabolism	62	C21-Steroid hormone metabolism	20
C21-Steroid hormone metabolism	28	Biosynthesis of steroids	19
Androgen and estrogen metabolism	24	Bile acid biosynthesis	11
Alpha-Linolenic acid metabolism	23	Arachidonic acid metabolism	7
Biosynthesis of unsaturated fatty acids	17	Alkaloid biosynthesis I	6
Neuroactive ligand-receptor interaction	15	Tyrosine metabolism	4
Linoleic acid metabolism	14	Naphthalene and anthracene degradation	4
Porphyrin and chlorophyll metabolism	10	Terpenoid biosynthesis	4
Bile acid biosynthesis	9	Phenylpropanoid biosynthesis	4
Galactose metabolism	7	Ascorbate and aldarate metabolism	3
Fatty acid biosynthesis	7	Androgen and estrogen metabolism	3
Biosynthesis of steroids	7	Pyrimidine metabolism	3
Drug metabolism - cytochrome P450	7	Phenylalanine metabolism	3
Prostate cancer	7	Tryptophan metabolism	3
Fructose and mannose metabolism	6	Citrate cycle (TCA cycle)	2
Sphingolipid metabolism	6	Glyoxylate and dicarboxylate metabolism	2
Starch and sucrose metabolism	5	Reductive carboxylate cycle (CO <sub>2</sub> fixation)	2
Retinol metabolism	5	Retinol metabolism	2
PPAR signaling pathway	5	Drug metabolism - cytochrome P450	2
Caffeine metabolism	4	Biosynthesis of unsaturated fatty acids	2
Bisphenol A degradation	4	Neuroactive ligand-receptor interaction	2
Tryptophan metabolism	4	Pentose and glucuronate interconversions	1
Alkaloid biosynthesis I	4	Fatty acid biosynthesis	1
Glycolysis / Gluconeogenesis	3	Ubiquinone biosynthesis	1
Ascorbate and aldarate metabolism	3	Glutamate metabolism	1
Fatty acid metabolism	3	Alanine and aspartate metabolism	1
Phenylalanine metabolism	3	Valine, leucine and isoleucine biosynthesis	1
Terpenoid biosynthesis	3	Phenylalanine, tyrosine and tryptophan biosynthesis	1
Alkaloid biosynthesis II	3	Glycerophospholipid metabolism	1
ABC transporters - General	3	Sphingolipid metabolism	1
Pentose phosphate pathway	2	Nicotinate and nicotinamide metabolism	1
Phenylalanine, tyrosine and tryptophan biosynthesis	2	Pantothenate and CoA biosynthesis	1
Streptomycin biosynthesis	2	Biotin metabolism	1
Phenylpropanoid biosynthesis	2	Porphyrin and chlorophyll metabolism	1
Aminoacyl-tRNA biosynthesis	2	Aminoacyl-tRNA biosynthesis	1
Fc epsilon RI signaling pathway	2	Metabolism of xenobiotics by cytochrome P450	1
Small cell lung cancer	2	Prostate cancer	1
Non-small cell lung cancer	2		
Inositol metabolism	1		

Fatty acid elongation in mitochondria	1
Ubiquinone biosynthesis	1
Tyrosine metabolism	1
Novobiocin biosynthesis	1
Inositol phosphate metabolism	1
Nicotinate and nicotinamide metabolism	1
Biotin metabolism	1
Calcium signaling pathway	1
Phosphatidylinositol signaling system	1
mTOR signaling pathway	1
Hedgehog signaling pathway	1
VEGF signaling pathway	1
Gap junction	1
Long-term depression	1
Insulin signaling pathway	1
GnRH signaling pathway	1
Melanogenesis	1
Adipocytokine signaling pathway	1
Type II diabetes mellitus	1
Basal cell carcinoma	1
Asthma	1

Table 7.2: Pathway found in the MassTrix, the masses submitted were characteristic for the two groups: HIGH and LOW values, in the last one were included also the transition samples. \* IC=Identified compound (present in the organism).

The most relevant pathways that differentiate the first group from the second (see figure 7.7) are:

- Arachidonic acid metabolism
- Biosynthesis of unsaturated fatty acids
- Linoleic acid metabolism
- Bile acid biosynthesis

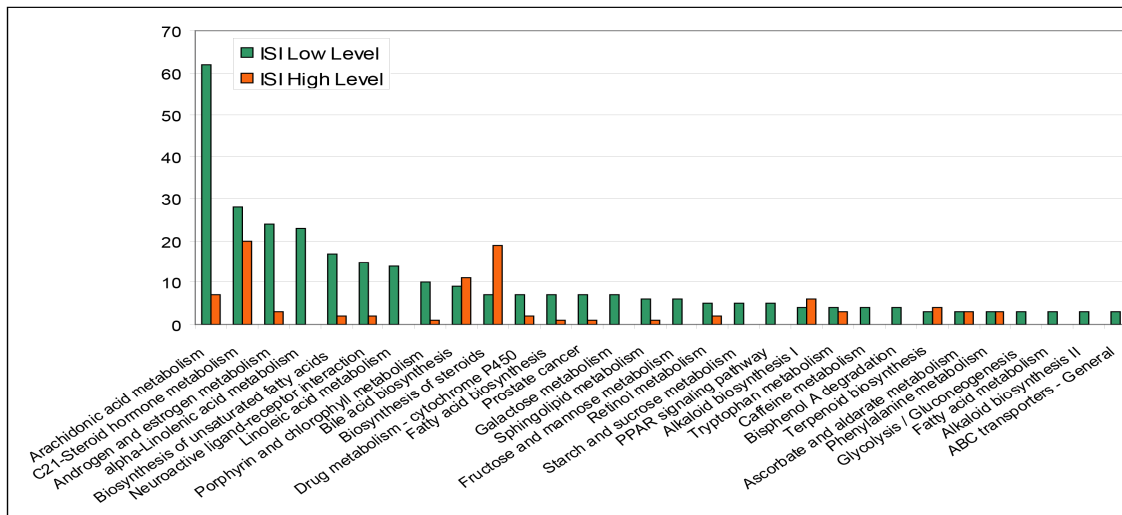


Figure 7.7: general schema with the most relevant pathway found.

The masses relative to the most important pathways were successively used to verify if they are effectively the biomarkers that differentiate the two levels of ISI. Using only these masses a new PLS-DA model (see figure 7.8) is built and the classification is performed according to the ISI level.

The assigned  $m/z$  coming from the four pathways gave the 65% of the recognition ability. This percentage is justified by considering that only 37% of the initial list of discriminant masses is taken in account. For the same reason the score scatter plot (see figure 7.7) reveals the two groups not well distinguished even though the two classes seem to cluster homogeneously.

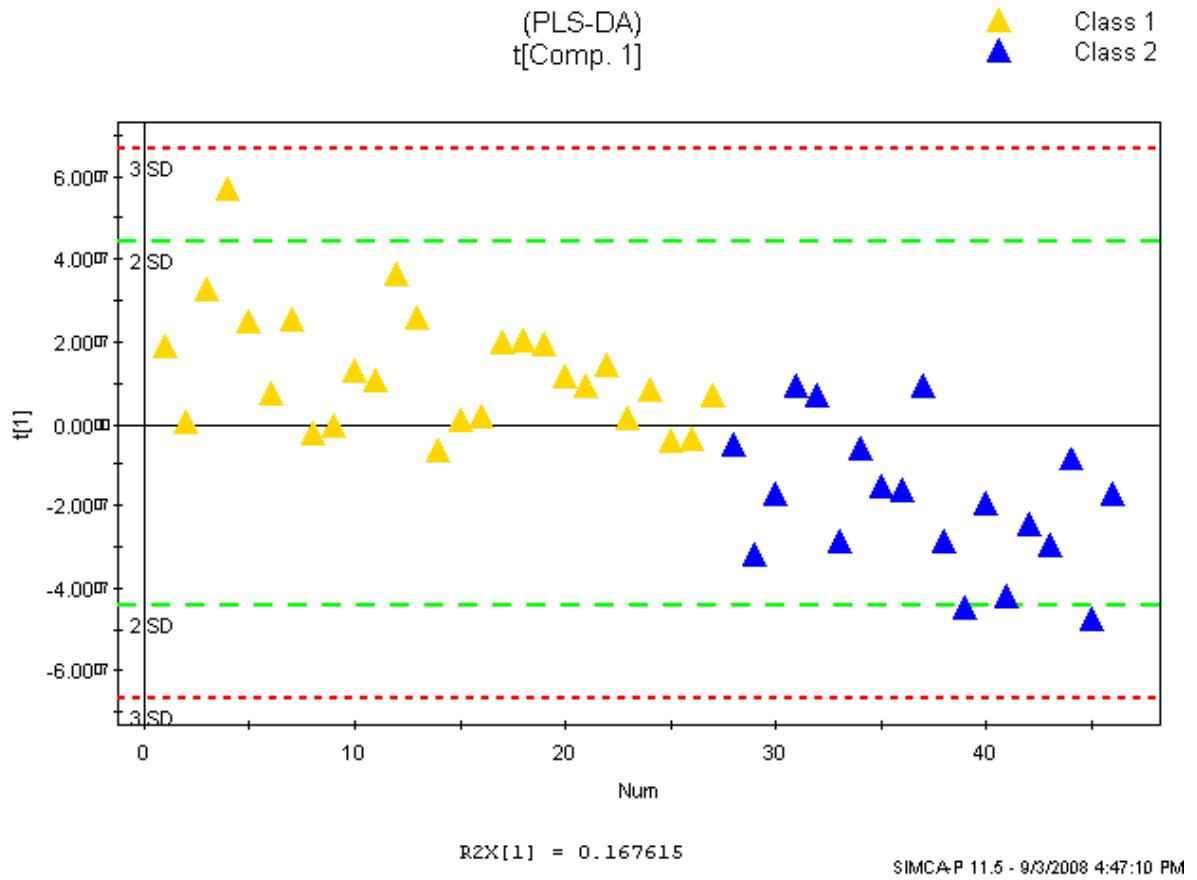


Figure 7.8: PLS-DA model, built up only with the assigned masse coming from the pathways: Arachidonic acid metabolism, Biosynthesis of unsaturated fatty acids, Linoleic acid metabolism, Bile acid biosynthesis.

## 7.5 Conclusion

In both metabolomics surveys, the application of multivariate analysis shows a great potential to map early biochemical changes in disease and hence provide a unique opportunity to develop predictive biomarkers that can trigger earlier interventions. These studies open the way for new approaches, offering suggestions for further investigations and debates. The technology is completely new and very complex because applied to human diseases. It confirms once more the need for an interdisciplinary collaboration to investigate more in detail the phenomenon. There are still limitations on this study. Indeed, high-resolution mass spectra reflect the isomer filtered complement of the entire space of molecular structures (Hertkorn, et al., 2007). An annotation such as the one proposed here thus associates experimental accurate mass (within an experimental error) with a limited number of bulk chemical formula (isomers), derived from the unique elementary composition space and restricted by the choice of the organism (and its annotated genome) (Suhre, et al., 2008). The differentiation between isomers and the final metabolite identification can only be done on a case-by-case basis in further identification steps, using classical analytical chemistry approaches involving metabolite orthogonal separation, spectroscopy and further spectrometry together with chemical synthesis (Chen, et al., 2008). An educated interpretation of the resulting pathway in the light of the genome of the organism thus remains the golden rule (Suhre, et al., 2008). Moreover, it has to be noticed that quantitative exploitation of the results is often strongly influenced by the stochastic ionization process in electrospray and the undefined amount of isobaric constituents of the target molecule.

The multivariate analyses have successfully classified the samples and also identified peaks that differ most significantly between the different groups. The discovery, interpretation, and presentation of multivariate spatial patterns play an important role for scientific understanding of complex problems.





# Chapter 8

## 8 CONCLUSIONS AND OUTLOOK

An in-house-developed strategy is presented that mainly conjunct Chemometric tools with high resolution mass spectrometry analysis. The new approach provides high potential in analyzing and extrapolating information from high complex datasets and was applied for different applications to evaluate its universal characteristic.

This strategy at first was applied to study the geometabolome of biogeographic isolations of the extremophilic bacterium *Salinibacter ruber* (chapter 4). Using this procedure we showed that strains of the cosmopolitan extremophilic bacterium *Salinibacter ruber*, isolated from different sites in the world, could be distinguished by means of characteristic metabolites, and that these differences can be correlated to their geographical isolation site distances. Only weighing the

relative intensity of each individual peak and treating the data by using multivariate analysis, revealed statistically significant differences between the different samples.

The same strategy was applied to describe and characterize the high complexity and diversity of metabolites in wines (chapters 5 and 6). It was possible to identify families of metabolites that could discriminate the species of wood of the barrels and their geographical origin. Based on the fact that several studies have revealed the influence of oak wood on the organoleptic properties of wines matured in oak barrels (Waterhouse, et al., 1994), (Jarauta, et al., 2005). This work opens up new ways to comprehend the mechanisms responsible for undesired evolutions (untimely oxidation) or even recording of environmental changes such as climatic modifications over decades.

The last challenge is to apply this technology to detect and monitor the health condition. In the last chapter this strategy was applied to biomedical and health diagnostics. It was possible to isolate possible metabolites as marker for COPD and prediabetic patients, also with the information coming from complementary databases. Preliminary results are presented and the study still requires many efforts and implementation to refine the strategy due to the high complexity of the human mechanism but the proof of principal is done to show the applicability of ICR-FT/MS combined to our data evaluation pipeline.

This is becoming more evident as ‘omics’ research is moving toward modeling biochemical networks through systems biology. However our strategy has two fundamental limits mainly originated from the principles of analysis. High-resolution mass spectra reflect the isomer filtered complement of the entire space of molecular structures (Hertkorn, et al., 2007). The annotated mass associates experimental accurate mass (within an experimental error) with a limited number of bulk chemical formula (isomers), derived from the unique elementary composition space and restricted by the choice of the organism (and its annotated genome). The differentiation between isomers and the final metabolite identification can only be done on a case by case basis in further identification steps, using other classical analytical chemistry approaches involving separation technologies like chromatography and electrophoresis, spectroscopy and further spectrometry together with chemical synthesis (Suhre, et al., 2008). An educated interpretation of the resulting pathway in the light of the genome of the organism thus remains the golden rule (Suhre, et al., 2008). Moreover the quantitative exploitation of the results is often strongly influenced by the stochastic ionization

process in electrospray and the undefined amount of isobaric constituents of the target molecule.

One solution of these limitations is the combination of different analytical techniques (separation/spectrometry/spectroscopy) which generates such a bright view of the data from a sample that evaluation is challenging to a statistician. This needs to combine automatic sample preparation before ultrahigh pressure liquid chromatographic (UHPLC) and/or capillary electrophoretic separation (CE) coupled to high resolution tandem mass spectrometry (Q/TOF) via various atmospheric pressure ionization modes (ESI, APCI, APLI, APPI). The separated compounds can then be analyzed in a high dimensionality by adding in addition a broad range of at-line and off-line nuclear magnetic resonance spectroscopy (NMR) methods.

This thesis presents an absolutely novelty in the use of electrospray ICR-FT/MS data analysis, especially it underlines the ability to scratch the surface in terms of potential applications dealing new hypothesis and future developments.

The processes are still technically complex, indeed there is a lack in terms of databases; for example many of the biomarkers found are not yet classified. The future challenges will be done in this direction because a considerable work in this scope still remains, also in the developing the chemical and computational technologies that provides a basis for this field. All these aspects reflect the fact that Metabolomics is a young field, especially when the technology applied belongs to the newest generation (ICR-FT/MS 12T).

Storing metadata with all information about the analytical technique and data-processing details are important to be able to reproduce the experimental conditions and compare results obtained in different time conditions and laboratories.

A possible strategy of data investigation is presented here and it is a complete excursus starting from the raw data to still a list of biomarkers, with also the possibility to identify part of them. Metabolomics raw data processing was probably the most challenging and time consuming step in data analysis because it requires automated data processing solutions.

The rule of statistical (in particular multivariate analysis) and machine-learning algorithms tools is crucial, with their ability to extrapolate from high complex datasets useful information, considering the biology system in which the survey is collocated. Moreover, it is an extensive task to find significant information in a large and “high complex” amounts of data. For the fact that the data are “high complex” and they contain already as much relevant information as possible.

The role of the databases is also extremely important, because it gives a biological meaning to the surveys. This work appears like a puzzle in which each piece is fundamental for the further development and all together give the complete meaning of the measurements. Many steps need further improvements but the structure itself is completed.

## 9 BIBLIOGRAPHY

- (n.d.). Retrieved from [www.atsjournals.org](http://www.atsjournals.org).
- Aharoni, A., & al., e. (2003). Terpenoid metabolism in wild type and transgenic *Arabidopsis* plants. *Plant cell.*, *15*, 2866-2884.
- Aharoni, A.; Ric de Vos, C.H.; Harrie, A.V.; Maliepaard, C.A.; Kruppa, G.; Bino, R.; Goodenowe, D.B.;. (2002). Nontargeted metabolome Analysis by use of Fourier transform ion cyclotron mass spectrometry. *A journal of integrative biology* , *6*, 217-234.
- Altmaier, E., Ramsay, S., Graber, A., Mewes, H., Weinberger, K., & Suhre, K. (2008). Bioinformatics analysis of targeted metabolomics--uncovering old and new tales of diabetic mice under medication. *J. Endocrinology*, *149*, 3478-3489.
- Amann, R., Ludwig, W., & Schleifer, K. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *J. Microbiol Rew.*, *59*, 143-169.
- Anthony, M., Sweatman, B., Beddell, C., Lindon, J., & Nicholson, J. (1994). Pattern recognition classification of the site of nephrotoxicity based on metabolic data derived from proton nuclear magnetic resonance spectra of urine. *Mol. Pharmacol.*, *46*, 199.
- Antón, J., Oren, A., Benloch, S., Rodríguez-Valera, F., A., & R., R.-M. R. (2002). *Salinibacter ruber* gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds. *Appl. Environ. Microbio.*, *52*, 485-491.
- Antón, J., Rosselló-Mora, R., Rodríguez-Valera, F., & Amann, R. (2000). Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Appl Environ Microbio.*, *66*, 3052-3057.
- Atkinson, A. J., Magnuson, W. G., Colburn, W. A., De Gruttola, V., DeMets, D., Downing, G., et al. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol Ther*, *69*, 89-95.

- Aussenac, J. e. (2001). Purification method for the isolation of monophosphate nucleotides from Champagne wine and their identification by mass spectrometry. *J. Chrom.*, *907*, 155-164.
- Baker, M. (2005). In biomarkers we trust? *Net. Biotechnol.* , *23*, 297-304.
- Bakker, J., & Timberlake, C. F. (1997). Isolation, Identification, and Characterization of New Color-Stable Anthocyanins Occurring in Some Red Wines. *J. Agric. Food Chem.*, *45*, 35-43.
- Balashov, S., Imasheva, E., Boichenko, V., Antón, J., Wang, J., & Lanyi, J. (2005). Xanthorhodopsin: a proton pump with a light-harvesting antenna. *Science* *309*, 2061.
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *J Chemometrics* *17*(3), 166-173.
- Barrera-Garcia, V. D., Gougeon, R. D., Di Majo, D., Carmen, D. A., Voilley, A., & Chassagne, D. (2007). *55*, 7021.
- Barrera-Garcia, V. D., Gougeon, R. D., Voilley, A., & Chassagne, D. (2006). Sorption Behavior of Volatile Phenols at the Oak Wood/Wine Interface in a Model System. *54*, 3982-3989.
- Bender, D. (2005). Perspective—the promise of metabolomics. *J Sci Food Agric*, *85*, 7-9.
- Bernhard, D., Moser, C., & Backovic, A. G. (2006). Cigarette smoke - an aging accelerator? *Exp.Geront*, *2*.
- Bhalla, R.; Narasimhan, K.; Swarup, S.;. (2005). Metabolomics and its role in understanding cellular responses in plants. *Plant cell Rep* , *24*, 562-571.
- Bilello, J. (2005). The agony and ecstasy of ‘omic’ technologies in drug development. *Curr.Mol Med* , *5*, 39-52.
- Bisson, L. F., Waterhouse, A. L., Ebeler, S. E., Andrew Walker, M., & Lapsley, J. T. (2002). The present and future of the international wine industry. *Nature* *418*, 696-699.
- Bloemen, K., Verstraelen, S., Van Den Heuvel, R., Witters, H., Nelissen, I., & Schoeters, G. (2007). The allergic cascade: Review of the most important molecules in the asthmatic lung. *Immunology letter*, *113*, 6-18.
- Boutilier, K., Ross, M., Podtelejnikov, A., Orsi, C., Taylor, R., Taylor, P., et al. (2005). Comparison of different search engines using validated MS/MS test datasets. *Anal. Chim. Acta*, *534*, 11-20.
- Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Series B, J. of royal statistical society* *26*, 211-246.
- Breitling, R., Pitt, A., & Parrett, M. (2006). Precision mapping of the metabolome. *TRENDS in Biotechnology*, *24*, 543-548.

- Brenner, S.; Johnson, M.; Bridgham, J.; (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* , 18, 630-634.
- Brindle, J. T., Antti, H., Holmes, E., Tranter, G., Nicholson, J., Bethell, W., et al. (2002). Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics. *Nat Med* 8, 1439-1445.
- Brindle, J., Nicholson, J., Schofield, P., Grainger, D., & Holmes, E. (2003). Application of chemometrics to 1H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. *Analyst*, 128, 32-36.
- Brown, P., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat Genet* , 21, 33-37.
- Burns, J., Crozier, A., & Lean, M. E. (2001). Alcohol consumption and mortality: is wine different from other alcoholic beverages? *Nutr. Met. Cardiovascular Dis.*, 11, 249-258.
- Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J., Holmes, E., & Trygg, J. (2006). OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemometrics*
- Cadahia, E., Varea, S., Munoz, L., Fernandez de Simon, B., & Garcia-Vallejo, M. J. (2001). *Agric. Food Chem.*, 49, 3677.
- Cavalier-Smith, T. (2007). Concept of a bacterium still valid in prokaryote debate. *Nature*, 446, 257.
- Ceres, I. (2008). <http://www.ceres-inc.com/techno/platforms/metab.html>.
- Chassagne, D., Guilloux-Benatier, M., & Alexandre, H. A. (2005). *Agric. Food Chem.*, 91, 39.
- Chatonnet, P., Boidron, J., & Pons, M. (1989). *Connaissance Vigne Vin*, 23, 223.
- Chatonnet, P., Dubourdieu, D., & Boidron, J. (1998). 49, 79.
- Chee, M.; Yang, R.; Hubbell, E.; (1996). Accessing genetic information with high-density DNA arrays. *Science*, 274, 610-614.
- Chen, J., Zhao, X., Fritsche, J., Yin, P., Schmitt-Kopplin, P., Wang, W., et al. (2008). Practical approach for the identification and isomer elucidation of biomarkers detected in a metabonomic study for the discovery of individuals at risk for diabetes by integrating the chromatographic and mass spectrometric information. *Analytical Chemistry*, 80, 1280-1289.
- Cheynier, V. e. (2006). Structure and Properties of Wine Pigments and Tannins. *Am. J. Enol. Vitic.*, 57, 298-305.

- Cho, J., & Tiedje, J. (2000). Biogeography and degree of endemism of fluorescent *Pseudomonas* strains in soil. *Appl. Environ. Microbiol.*, 66, 5448-5456.
- Comisarow, M., & Marshall, A. (1974). Fourier Transform Ion Cyclotron Resonance Spectroscopy. *Chemical Physics Letters*, 25, 282.
- Comisarow, M., & Marshall, A. (1975). Resolution-Enhanced Fourier Transform Ion Cyclotron Resonance Spectroscopy. *Chemical Physics Letters*, 62, 293.
- Comisarow, M., & Marshall, A. (1976). Theory of Fourier transform ion cyclotron resonance mass spectroscopy. I. Fundamental equations and low-pressure line shape. *Chemical Physics Letters*, 64, 110.
- Conde, C. e. (2006). Pathways of glucose regulation of monosaccharide transport in grape cells. *Plant physiology*, 141, 1563-1577.
- Cooley, J., & Tukey, J. (1965). An Algorithm for Machine. Computation of Complex Fourier Series. *Mathematics computation*, 19, 297.
- Cooper, H. J., & Marshall, A. G. (2001). Electrospray Ionization Fourier Transform Mass Spectrometric Analysis of Wine. *J. Agric. Food Chem.*, 49, 5710-5718.
- Corcelli, A., Lattanzio, V., Mascolo, G., Babudri, F., Oren, A., & Kates, M. (2004). Novel sulfonolipid in the extremely halophilic bacterium *Salinibacter ruber*. *Appl. Environ. Microbiol.*, 70, 6678-6685.
- Corder, R. e. (2006). Red wine procyanidins and vascular health. *Nature*, 444, 566.
- Craig, A., Cloarec, O., Holmes, E., Nicholson, J., & Lindon, J. (2006). Scaling and normalization effects in NMR spectroscopic metabonomics data set. *Anal. Chem.* 78
- Cui, J. e. (2002). Cardioprotection with white wine. *Drugs under experimental and clinical research*, 28, 1-10.
- Curtu, A. L., Gailing, O., Leinemann, L., & Finkeldey, R. (2007). *Plant Biol.*, 9, 116.
- Das, D. K. (1999). Cardioprotection of red wine: role of polyphenolic antioxidants. *J. Drugs under experimental and clinical research*, 25, 115-120.
- Davidov, E., Holland, J., Marple, E., & Naylor, S. (2003). Advancing drug discovery through systems biology. *Drug Discov. Today*, 8, 175-183.
- de Wit, R., & Bouvier, T. (2006). 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Envir. Microbiol.*, 8, 755-758.
- DeFronzo, R. A. (2004). Pathogenesis of type 2 diabetes mellitus. *Med. Clin. North. Am.*, 88.4, 787-835.



- Deming, S. (1986). Chemometrics: an overview. *Clin Chem* , 10, 1702-1706.
- Dettmer, K., Aronov, P., & Hammock, B. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry rew.*, 26, 51-78.
- Dixon, R.A.; Gang, D.R.; Charlton, A.J.;. (2006). Perspective-applications of metabolomics in agriculture. *J Agric Food Chem* , 54, 8984-8994.
- Doco, T., Quellec, N., Moutounet, M., & Pellerin, P. (1999). Polysaccharide Patterns During the Aging of Carignan noir Red Wines. *J. Enol Vitic.*, 50, 25-32.
- Doussot, F., De Jeso, B., Quideau, S., & Pardon, P. J. (2002). *Analisis* 50, 5955.
- Doussot, F., Pardon, P., Dedier, J., & De Jeso, B. (2000). *Agric. Food Chem.*, 28, 960.
- Dunne, V., Bhattachayya, S., Besser, M., Rae, C., & Griffin, J. (2005). *Metabolites from cerebrospinal fluid in aneurysmal subarachnoid haemorrhage correlate with vasospasm and clinical outcome: a pattern-recognition 1H NMR study.* *NMR Biomed.*, 18, 24
- Eckel, R. H., Grundy, S., & Zimmet, P. (2005). The metabolic syndrome. *Lancet.*, 365.9468, 1415-1428.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling.* Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall.
- Eriksson, L., Antti, H., Gottfries, J., Holmes, E., Johansson, E., Lindgren, F., et al. (2004). Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Analytical and bioanalytical Chem*, 380.3, 419-429.
- Eriksson, L., Johansson, E., & Wold, S. (1997). *Quantitative Structure-activity Relationships in Environmental Sciences-VII.* SETAC, Pensacola.
- Eriksson, L., Johansson, E., & Wold, S. *Quantitative Structure-activity Relationships in Environmental Sciences-VII.* Pensacola, USA: SETAC, In Schuurmann, G. and Chen, F.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). Umeå, Sweden: Umetrics AB.
- Eriksson, L., Johansson, E., Kettaneh-Wold, N., & Wold, S. (2001). Multi-and Megavariate Data Analysis-Principles and Applications. 94-97.
- Eriksson, L., Toft, M., Johansson, E., Wold, S., & Trygg, J. (2006). Separating Y-predictive and Y-orthogonal variation in multi-block spectral data. 20, 352-361.

- Fan, X., Bai, J., & Shen, P. (2005). Diagnosis of breast cancer using HPLC metabonomics fingerprints coupled with computational methods. *Conf. Prot. IEEE Eng. Med. Biol. Soc.*, 6, 6081-6084.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17, 368-376.
- Fenchel, T. (2003). Biogeography for bacteria. *Science*, 301, 925-926.
- Fenchel, T., & Finlay, B. (2006). The diversity of microbes: resurgence of the phenotype. 361, 1965-1973.
- Feuillat, F., & Keller, R. (1997). *Am. J. enol. Vitic.* 48, 502.
- Feuillat, F., Keller, R., Sauvageot, F., & Puech, J.-L. (1999). Characterization of French Oak Cooperage (*Quercus robur* L., *Quercus petraea* Liebl.). *Am. J. enol. Vitic.* 50, 513-518.
- Feuillat, M. B., Keller, R., Huber, F., Leaute, B., & Puech, J. (2003). 109, 19.
- Feuillat, M. (2003). *Compte rendu d'étude préliminaire sur l'expérimentation "Tonnellerie 2000"*. Dijon.
- Feuillat, M. e. (2003). La variabilité du bois de merrain : subir ou agir. 109, 19-22.
- Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Mol Biol.*, 48, 155-171.
- Fiehn, O., Kopka, J., & Dormann, P. (2000). Metabolite profiling for plant functional genomics. *Nature Biotech.*, 18(11), 1157-1161.
- Flamini, R., Panighel, A., Perchiazzi, N., & Ongarato, S. (2005). Monitoring of the principal carbonyl compounds involved in malolactic fermentation of wine by solid-phase microextraction and positive ion chemical ionization GC/MS analysis. *J. Mass spectrometry*, 40, 1558-1564.
- Garde-Cerdan, T., & Ancin-Azpilicueta, C. (2006). Review of quality factors on wine ageing in oak barrels. *Trends Food Sci. Techn.*, 17, 438-447.
- Garde-Cerdan, T., & Ancin-Azpilicueta, C. (2006). *Trends Food Sci.* 17, 438.
- Geladi, P., & Kowalski, B. (1986). Partial Least-Squares regression: A Tutorial. *Analytica chim. Acta*, 185, 1-17.
- Gerner, C.; Teufelhofer, O.; Parzefall, W.;. (2006). New approaches to toxicologic mechanisms by the application of genomics, proteomics and metabolomics. *Toxicol Lett* , 164, 3-4.
- Gibney, M.J.; Walsch, M.; Brennan, L.;. (2005). Metabolomics in human nutrition: opportunities and challenges. *Am J Clin Nut* , 82, 497-503.
- Glabasnia, A., & Hofmann, T. (2007). *J. Agric. Food Chem.*, 55, 4109.
- Godchaux, W., & Leadbetter, E. (1984). Sulfonolipids of gliding bacteria. Structure of the N-acylaminosulfonates. *J. Biol. Chem.*, 259, 2982-2990.

- Goes da Silva, F. (2005). Characterizing the Grape Transcriptome. Analysis of Expressed Sequence Tags from Multiple Vitis Species and Development of a Compendium of Gene Expression during Berry Development. *Plant Physiology*, 139, 574-597.
- González-Arjona, D., López-Pérez, G., & González, A. (1999). Performing procrustes discriminant analysis with HOLMES., *Talanta*, 49, 189.
- Goodacre, R. (2005). Biomarker discovery using metabolomics and explanatory machine learning. *J Med Gen* , 42, 16.
- Gougeon, R. D. (submitted). Expressing forest origins in the chemical composition of cooperage oak woods and corresponding wines by FTICR-MS., *J. Chem. Eur*
- Green, J. L., Bohannan, B., & Whitaker, R. J. (2008). Microbial Biogeography : from taxonomy to traits., *Science*, 320, 1039-1043.
- Green, J., & Bohannan, B. (2006). Spatial scaling of microbial biodiversity., *Trends Ecol. Evol.*, 21, 501-517.
- Green, J., Bohannan, B., & Whitaker, R. (2008). Microbial Biogeography: From Taxonomy to Traits. *Science* , 320, 1039-1043.
- Green-Tringe, S., & Rubin, W. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.*, 6, 805-814.
- Gross, R. W., & Han, X. (2007). Lipidomics in diabetes and the metabolic syndrome. *Methods Enzymol.*, 433, 73-90.
- Guchu, E., Diaz-Maroto, M., Diaz-Maroto, I., Vila-Lameiro, P., & Perez-Coello, M. (2006). *J. Agric. Food Chem.*, 54, 3062.
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52, 696-704.
- Hale, M. D., McCafferty, K., Larmie, E., Newton, J., & Swan, J. (1999). *J Enol Vitic.*, 50, 495.
- Halket, J., & al., e. (2003). Deconvolution gas chromatography mass spectrometry of urinary organic acids. Potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom* , 13, 279-284.
- Han, X., Holtzman, D., McKeel, D., Kelley, J., & Morris, J. (2002). Substantial sulfatide deficiency and ceramide elevation in very early Alzheimer's disease: potential role in disease pathogenesis. *Neurochem.*, 82, 809-818.
- Harrigan, G. (2006). Metabolic profiling—IBC's Inaugural Meeting—using metabolomics to accelerate drug discovery and development, November 2005, Durham. NC. *Idrugs* , 9, 28-31.

- Hartmann, M., Baumbach, J., & Nolte, J. (2006). Metabolomics—a new approach for the detection of colon cancer? *Ann Oncol* , 17, 36.
- Hasegawa, M., Kishino, H., & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol. Evol.*, 22, 160-174.
- Hayasaka, Y., Wilkinson, K. L., Elsey, G. M., Raunkjaer, M., & Sefton, M. A. (2007). *J Agric. Food Chem.*, 55, 9195.
- Hertkorn, N., Meringer, M., Gugisch, R., Ruecker, C., Frommberger, M., & Perdue, E. (2007). High-precision frequency measurements: indispensable tools at the core of molecular-level analysis of complex systems Analytical Bioanalytical Chemistry. *Anal. Bioanal. Chem.* 389, 1311-1327.
- Herve du Penhoat, C., Michon, V., Peng, S., Viriot, C., Scalbert, A., & Gage, D. (1991). *J. Chem. Soc.*, 1653.
- Hochberg, Z. (2006). Metabolomics of the obese. *Int J Obesity* , 30, 4.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian J. Statistics*, 6, 65-70.
- Holmes, E., & Antti, H. (2002). Chemometric contributions to the evolution of metabolomics: mathematical solutions to characterising and interpreting complex biological NMR spectra. *Analyst*, 127, 1549.
- Holmes, E., Tsang, T., Huang, J., Leweke, F., & Koethe, D. (2006). Metabolic profiling of CSF: evidence that early intervention may impact on disease progression and outcome in schizophrenia. *PLoS Med.*, 3, 327.
- Horváth, I., Hunt, J., & Barnes, P. (2005). Exhaled breath condensate: methodological recommendations and unresolved questions. *Euro Respir. J.*, 26, 523-548.
- Hue, A. Y., & Durand, B. M. (1977). Occurrence and significance of humic acids in ancient sediments. *Fuel*, 56, 73-80.
- Huges-Martiny, J., Bohannan, B., Brown, J., Colwell, R., Fuhrman, J., & Green, J. (2006). Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.*, 4, 102-112.
- Jackson, E. (1991). *A User's Guide to Principal Components*. New York: Wiley-Interscience.
- Jaillon, O. e. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449, 463-467.
- Jang, M. e. (1997). Cancer Chemoprotective Activity of Resveratrol, a Natural Product Derived from Grapes. *Science*, 275, 218-220.
- Jarauta, I., Cacho, J., & Ferreira, V. (2005). Concurrent Phenomena Contributing to the Formation of the Aroma of Wine during Aging in Oak Wood: An Analytical Study. 53, 4166-4177.

- Jeandet, P., Clément, C., & Conreux, A. (2007). *Macromolecules and Secondary Metabolites of Grapevine and Wine*. Paris, France.
- Jenkins, H. e. (2005). A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotech.*, 22, 1601-1605.
- Jonsson, P., & al., e. (2005). Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyt.*, 130.5, 701-707.
- Kaddurah-Daouk, R., Kristal, B., & Weinshilboum, R. (2008). Metabolomics: A Global Biochemical Approach to Drug Response and Disease. *Annual Rev. of Pharmacology and Toxicology*, 48, 653-683.
- Kaddurah-Daouk, R., McEvoy, J., Baillie, R., Lee, D., & Yao, J. (2007). Metabolomic mapping of atypical antipsychotic effects in schizophrenia. *Mol. Psychiatry*.
- Kell, D., Brown, M., Davey, H., Dunn, W., Spasic, I., & Oliver, S. (2005). Metabolic footprinting and systembiology: the medium is the message. *Nature*.
- Kemsley, E. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics Intel Lab System*, 33, 47-61.
- Kenny, L., Dunn, W., Ellis, D., Myers, J., & Baker, P. (2005). Novel biomarkers for pre-eclampsia detected using metabolomics and machine learning. *Metabolomics*, 1, 277.
- Kharitonov, S., & Barnes, P. (2001). Exhaled Markers of Pulmonary Disease. *Am J. Respir. Crit. Care Med*, 163, 1693-1722.
- Kind, T., & Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7, 234.
- Knowler, W. C., & al., e. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N. Engl. J. Med*, 346.6, 393-403.
- Konstantinidis, K., & Tiedje, J. (2005). Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol*, 187, 6285-6264.
- Koop, R. (2003). Combinatorial biomarkers: From early toxicology assays to patient population profiling. *Drug Discov. Today* , 73, 284-291.
- Kuwatsuka, S., Tsutsuki, K., & Kumada, K. (1978). Chemical studies on soil humic acid I. Elementary composition of humic substances. *Soil Sci. Plant Nutr*, 24, 337-347.
- Lavine, B., & Workman, J. (2004). Fundamental Review of Chemometrics. *Anal. Chem.*, 76, 3365.

- Lee, K. R., Lin, X., Park, D., & Eslava, S. (2003). Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics*, 3, 1680-1686.
- Lenz, E. M., & Wilson, I. (2007). Analytical strategies in metabonomics. *J. Proteome. Res.*, 6.2, 443-458.
- Lewin, D.A.; Weiner, M.P.;. (2004). Molecular biomarkers in drug development. *Drug Discov. Today* , 9, 976-983.
- Li, X., Fekete, A., Englmann, M., Götz, C., Rothballer, M., Frommberger, M., et al. (2006). Development and application of a method for the analysis of N-acylhomoserine lactones by solid-phase extraction and ultra high pressure liquid chromatography. *J. Chromatogr.*, 1134, 184-193.
- Lindon, J. C., & al., e. (2004). Metabonomics technologies and their applications in physiological monitoring, drug safety assessment and disease diagnosis. *Biomarkers*, 9.1, 1-31.
- Lindon, J. C., Nicholson, J. K., & Holmes, E. (2007). *The Handbook of Metabonomics and Metabolomics*. Amsterdam: Elsevier.
- Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C., Craig, A., Pearce, J. T., et al. (2005). Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotech.*, 23, 833-838.
- Lindon, J., Holmes, E., & Nicholson, J. (2007). *Metabonomics in pharmaceutical R&D* (Vol. 274). FEBS J.
- Lindon, J.C. et al.;. (2005). Summary recommendations for standardization and reporting of metabolic analyses. *Nature Biotechnology* , 23, 833-838.
- Lu, X., & al., e. (2008). LC-MS-based metabonomics analysis. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.*, 866.1-2, 64-76.
- Lucio, M., & Schmitt-Kopplin, P. (2006). Modelling the binding of triazine herbicides to humic substances using capillary electrophoresis. *Environ. Chem.*, 4, 15-20.
- Ludwig, W., & Klenk, H. (2001). *Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics*. In *Bergey's Manual of Systematic Bacteriology*. New York: Springer.
- Ludwig, W., Strunk, O., Westram, R., Richter, L., & Meier, H. Y. (2004). ARB: a software environment for sequence data. *Nucleic Acid Res.*, 32, 1363-1371.
- Lund, S. T., & Bohlmann, J. (2006). The Molecular Basis for Wine Grape Quality - A Volatile Subject. *Science*, 311, 804-805.
- Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, Å., Pettersen, J., et al. (1998). *Chemometr. Intellig. Lab. Syst.*, 42, 3-40.

- Maki, K. C., Abaira, C., & Cooper, R. (1995). Arguments in favor of screening for diabetes in cardiac rehabilitation. *J. Chardiopulm. Rehabil.*, 15.2, 97-102.
- Malhi, H., & Gores, G. (2006). Cancer therapy—back to metabolism. *Cancer Biol Ther* , 5, 986-987.
- Malins, D., Hellstrom, K., Anderson, K., Johnson, P., & Vinson, M. (2002). Antioxidant-induced changes in oxidized DNA. *Proc. Nat. Acad. Sci. USA*, 99, 5937-5941.
- Mämmela, P., Savolainen, H., Lindroos, L., Kangas, J., & Vartiainen, T. (2000). *J. Chrom.*, 891, 75.
- Manly, B. (1997). Randomization, Bootstrap and Monte Carlo Methods in Biology. *J. of animal Ecology*, 67, 162.
- Markley, J., Anderson, M., Cui, Q., Eghbalnia, H., Lewis, I., Hegeman, A., et al. (2007). New bioinformatics resources for metabolomics. *Pacific Symposium on biocomputing* , 12, 157-168.
- Marmot, M. G., Rose, G., Shipley, M. J., & Thomas, B. J. (1981). Alcohol and mortality: a U-shaped curve. *Lancet*, 1, 580-583.
- Marshall, A. (2004). Accurate mass measurement: taking full advantage of nature's isotopic complexity. *Physic*, 347, 503-508.
- Marshall, A., & Guan, S. (1996). Advantages of High Magnetic Field for Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Rapid Commun. Mass Spect.* , 10, 1819-1823.
- Martens, H., & Næs, T. (1990). Multivariate Calibration. *Moleculare Spectroscopy*, 46, 1541.
- Martens, H., Dijkstrahuis, G. B., & Byrne, D. V. (2000). Power of experimental designs, estimated by Monte Carlo simulation. *J. Chemometrics*, 14, 441-462.
- Masson, E., Baumes, R., Le Guerneve, C., & Puech, J. (2000). *Agric. Food Chem.*, 48, 4306.
- Masson, G., Guichard, E., Fournier, N., & Puech, J. (1995). *J. Enol. Vitic.*, 46, 424.
- Matsuda, M., & DeFronzo, R. (1999). Insulin sensitivity indices obtained from oral glucose tolerance testing: comparison with the euglycemic insulin clamp. *Diabetes Care*, 22.9, 1462-1470.
- Maturrano, L., Santos, F., Rosselló-Mora, R., & Antón, J. (2006). Microbial diversity in Maras salterns, a hypersaline environment in the peruvian andes. *Appl. Environ. Microbiol.*, 72, 3887-3895.
- Matuszewski, B.K.; Constanzer, M.L.; Chavez Eng, C.;. (2003). Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC- MS/MS. *Anal Chem* , 75, 3019-3030.

- Meija, J. (2000). Mathematical tools in analytical mass spectrometry. *Analytical and Bioanalytical Chemistry*
- Mesbah, N., Abou-El-Ela, S., & Wiegel, J. (2007). Novel and unexpected prokaryotic diversity in water and sediments of the alkaline, hypersaline lakes of the Wadi An Natrun. *Microb. Ecol.*
- Mielke, P. W., & Berry, H. (2001). *Permutation methods: A distance function approach*. New York: Springer.
- Monagas, M., Bartolome, B., & Gomez-Cordoves, C. (2005). Updated Knowledge About the Presence of Phenolic Compounds in wine. *Critical Rev. in Food Science and Nutrition*, 45, 85-118.
- Mongay, C., Pastor, A., & Olmos, C. (1996). Determination of carboxylic acids and inorganic anions in wines by ion-exchange chromatography. *J. Chrom.*, 736, 351-357.
- Mongodin, E., Nelson, K., Daugherty, S., Deboy, R., & Wister, J. K. (2005). The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc. Natl. Accad. Sci.*, 102, 18147-18152.
- Monteiro, S. e. (2001). The Wide Diversity of Structurally Similar Wine Proteins. *Agric. Food Chem.*, 49, 3999-4010.
- Mosedale, J. R., Charrier, B., & Janin, G. (1996). 69, 111.
- Mosedale, J. R., Puech, J.-L., & Feuillat, F. (1999). The influence on wine flavor of the oak species and natural variation of heartwood components. *Am. J. Enol. Vitic.*, 50, 503-512.
- Mosedale, J., & Ford, A. (1996). *J. Sci. Food Agric.*, 70, 273.
- Nguyen, D., & Rocke, D. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, 18, 1625-1632.
- Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Molecular System Biology*, 2, 52.
- Nicholson, J. K., Holmes, E., & Wilson, I. D. (2005). Gut microorganisms, mammalian metabolism and personalized health care. *Nature Rev. Microb.*, 3, 431-438.
- Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29, pp. 1181-1189.
- Nocairi, H., Qannari, E., Vigneau, E., & Bertrand, D. (2005). Discrimination on latent components with respect to patterns. *Data Anal.*, 49, 139.



- Ochman, H., & Davalos, L. (2006). The nature and dynamics of bacterial genomes. *Science*, 311, 1730-1733.
- Odunsi, K., Wollman, R., Ambrosone, C., Hutson, A., & McCann, S. (2005). Detection of epithelial ovarian cancer using 1H-NMR-based metabonomics. *Int. J. Cancer*, 113, 782-788.
- Oliver, S. (1998). Systematic functional analysis of the yeast genome. *TIBTECH*, 16, 447.
- Onzález-Arjona, D., López-Pérez, G., & González, A. (1999). Performing procrustes discriminant analysis with HOLMES. *Talanta*, 49, 189.
- Ortiz, M., Sáez, J., & Palacios, J. (1993). Typification of alcoholic distillates by multivariate techniques using data from chromatographic analyses. *Analyst*, 118, 801.
- Ott, M., & Vriend, G. (2006). Correcting ligands, metabolites, and pathways. *BMC Bioinformatics*, 7, 517.
- Paige, L., Mitchell, M., Krishnan, K., Kaddurah-Daouk, R., & Steffens, D. (2006). A preliminary metabolomic analysis of older adults with and without depression. *Int. J. Geriatr. Psychiatry*, 22, 418-423.
- Papke, R., Ramsing, N., Bateson, M., & Ward, D. (2005). Geographical isolation in hot spring cyanobacteria. *Environ. Microbiol.*, 5, 650-659.
- Parra, V., Arrieta, A., Fernandez-Escudero, J. A., Iniguez, M., Saja, J. A., & Rodriguez-Mendez, M. L. (2006). *Anal. Chim. Acta.*, 563, 229.
- Peña, A., Valens, M., Santos, F., Buczolits, S., Antón, J., & Kämpfer, P. (2005). Intraspecific comparative analysis of the species *Salinibacter ruber*. *Extremophiles*, 9, 151-161.
- Perdue, E. M., & Lytle, C. (1983). Distribution model for binding of protons and metal ions by humic substances. *Env. Sci and Tech.*, 17, 654.
- Perdue, E. M., Hertkorn, N., & Kettrup, A. (2007). *Anal. Chem.*, 79, 1010.
- Prida, A., & Puech, J. (2006). *Agric. Food Chemist.*, 54, 8115.
- Puech, J., Feuillat, F., & Mosedale, J. (1999). *J. Enol. Vitic.*, 50, 469.
- Quideau, S. e. (2005). The Chemistry of Wine Polyphenolic C-Glycosidic Ellagitannins Targeting Human Topoisomerase II. *Chem. Eur. J.*, 11, 6503-6513.
- Quideau, S., & Feldman, K. S. (1996). *Chem. Rev.*, 96, 475.
- Quideau, S., Jourdes, M., Lefeuvre, D., Montaudon, D., Saucier, C., Glories, Y., et al. (2005). *Chem. Eur.*, 11, 6503.
- Quideau, S., Jourdes, M., Saucier, C., Glories, Y., Pardon, P., & Baudry, C. A. (2003). *Angewand. Chem. Int.*, 42, 6021.

- Raamsdonk, L., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M., et al. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol.* , 19, 45-50.
- Ramette, A., & Tiedje, J. (2006). Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *J. Microb. Ecol.*
- Ren, B.; Robert, F.; Wyrick, J.J.;. (2001). Genome-wide location and function of DNA binding proteins. *Science* , 290, 2306-2309.
- Renaud, S., & De Lorgeril, M. (1992). Wine, alcohol, platelets and the French paradox for coronary heart disease. *Lancet*, 339, 1523-1526.
- Reuter, J. H., & Perdue, E. M. (1984). A chemical structural model of early diagenesis of sedimentary humus/protokerogens. *Mit. Geol. Pald. Inst. Univ. Hamburg*, 56, 249-262.
- Rist, M.J.; Wenzel, U.; Daniel, H.;. (2006). Nutrition and food science go genomics. *TIBTECH* , 24, 172-178.
- Robertson, D. (2005). Metabonomics in toxicology: a review. *Toxicol Sci* , 85, 809-822.
- Robertson, D. (2005). Metabonomics in toxicology: a review. *Toxicol Sci.*, 85, 809-822.
- Rocke, D., & Durbin, B. (2003). Approximate Variance-stabilizing Transformations for Gene-expression Microarray Data. *Bioinformatics*, 19(8), 966-972.
- Roessner, U., Willmitzer, L., & Fernie, A. R. (2001). High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant. Physiol.*, 127, 749-764.
- Roessner, U., Willmitzer, L., & Fernie, A. R. (2001). Metabolite profiling allows comprehensive phenotyping of genetically or environmentally modified plant system. *Plant Cell* , 13, 11-29.
- Rolan, P., Atkinson, A., & Lesko, L. (2003). Use of biomarkers from drug discovery through clinical practice: report of the Ninth European Federation of Pharmaceutical Sciences Conference on Optimizing Drug. *Clin Pharmacol Ther* , 73, 284-291.
- Rossello´-Mora, R., Lucio, M., Pená, A., Brito-Echeverría, J., López-López, A., Valens-Vadell, M., et al. (2008). Metabolic evidences of biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. *ISME*, 2, 242-253.
- Rozen, S., Cudkowicz, M., Bogdanov, M., Matson, W., & Kristal, B. (2005). Metabolomic analysis and signatures in motor neuron disease. *Metabolomics*, 1, 101-108.

- Ruf, J.-C. (2003). Overview of epidemiological studies on wine, health and mortality. *Drugs under experimental and clinical research*, 29, 173-179.
- S. Kuwatsuka, K. Tstsuki, K. Kumada. (1978). Chemical studies on soil humic acid I. Elementary composition of humic substances. *Soil Sci. Plant Nutr.*24, 337-347.
- Sabatine, M., Liu, E., Morrow, D., Heller, E., & McCarroll, R. (2005). Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, 112(25), 3868-3875.
- Saito, N.; Robert, M.; Kitamura, S.;. (2006). Metabolomics approach for enzyme discovery. *J Proteome Res* , 5, 1979-1987.
- Sajbidor, J. (1997). Effect of some environmental factors on the content and composition of microbial membrane lipids. *Crit. Rev. Biotechnol.*, 87-103.
- Sauvageot, F., & Feuillat, F. (1999). *J. Enol. Vitic.*, 50, 447.
- Savaglio, S., & Carbone, V. (2000). Scaling in athletic world record. *Nature*, 404, 244.
- Schafer, S., & al., e. (2007). Lifestyle intervention in individuals with normal versus impaired glucose tolerance. *Eur. J. Clin. Invest.*, , 37.7, 535-543.
- Schlotterbeck, G.; Ross, A.; Dieterle, F. (2006). Metabolic profiling technologies for biomarker discovery in biomedicine and drug development. *Pharmacogenomics* , 7, 1055-1075.
- Schmitt Kopplin, P., & al., e. (2007). Ion Cyclotron resonance Fourier transform Mass spectrometry for non targeted metabolomics of molecular interactions in the rhizosphere. *Soil Biology* , 11.
- Singh, O. (2006). Proteomics and metabolomics: the molecular make-up of toxic aromatic pollutant bioremediation. *Proteomics* , 6, 5481-5492.
- Singh, O., & Nagaraj, N. (2006). Transcriptomics, proteomics and interactomics: unique approaches to track the insights of bioremediation. *Brief. Funkt. Genomic Proteomic*, 4, 355-362.
- Sjödin, K., Schroeder, L. M., Eidmann, H. H., Norin, T., & Wold, S. (1989). Attack rates of scolytids and composition of volatile wood constituents in healthy and mechanically weakened pine trees. *Scand. J. Forest. Res.*, 4, 379-391.
- Sjöström, M., Wold, S., & Södestrom, B. (1986). *PLS discriminant plots. In Pattern Recognition in Practice II*. Amsterdam: Elsevier.
- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, 78.

- Soleas, G., Diamandis, E. P., & Goldberg, D. M. (1997). Wine as a Biological Fluid: History, Production, and Role in Disease Prevention. *J. of clinical lab. Anal.*, 11, 287-313.
- Sommer, H., Thomas, H., & Hipple, J. (1951). The Measurement of  $e / M$  by Cyclotron Resonance. *Phys. Rev.*, 82, 697.
- Sória-Carrasco, V., Valens-Vadell, M., Peña, A., Antón, J., Amann, R., Castresana, J., et al. (2007). Phylogenetic position of *Salinibacter ruber* based on concatenated protein alignments. *Syst. Appl. Microbiol.*, 30, 171-179.
- Spillman, P. J., Sefton, M. A., & Gawel, R. (2004). *Aust. J. Grape Wine Res.*, 10, 216.
- Stahle, L., & Wold, S. (1987). Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J. Chemometrics*, 1, 185-196.
- Staley, J. (2006). The bacterial species dilemma and the genomic-phylogenetic species concept. *Phil. Trans. R. Soc. B.*, 361, 1899.
- Suhre, K., & Schmitt-Kopplin, P. (2008). MassTRIX: mass translator into pathways. *Nucleic Acids Research*, 194, 1-4.
- Taylor, J. E., Hyde, K. D., & Jones, E. B. (2000). *J. Biogeogr.*, 27, 297.
- Towey, J., & Waterhouse, A. (1996). *J. Enol. Vitic.*, 47, 163.
- Trigg, J. (n.d.). PhD Thesis.
- Trygg, J. (2002). O2-PLS for qualitative and quantitative analysis in multivariate calibration. *J. Chemometrics*, 16(6), 283-293.
- Trygg, J. (2008). Orthogonal Projections to Latent Structures Discriminant Analysis Modeling on in Situ FT-IR Spectral Imaging of Liver Tissue for Identifying Sources of Variability. *Anal. Chem.*, 20: 18714965.
- Trygg, J., & Wold, S. (2003). O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemometrics*, 17, 53-64.
- Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *J. Chemometrics*, 16(3), 119-128.
- Trygg, J., Gullberg, J., Hohansson, A. I., Jonsson, P., & Moritz, T. (2006). Chemometrics in metabolomics: an introduction. *Biotechnology in agriculture and forestry*, 57, 117.
- Tuomilehto, J., & al., e. (2001). Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance. *N. Engl. J. Med.*, 344.18, 1343-1450.
- Uetz, P.; Giot, L.; Cagney, G.;. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* , 403, 623-627.

- Underwood, B., Broadhurst, D., Dunn, W., Ellis, D., & Michell, A. (2006). Huntington disease patients and transgenic mice have similar procatabolic serum metabolite profiles. *J. Brain*, *129*, 877-886.
- Unne, V., Bhattachayya, S., Besser, M., Rae, C., & Griffin, J. (2005). Metabolites from cerebrospinal fluid in aneurysmal subarachnoid haemorrhage correlate with vasospasm and clinical outcome: a pattern-recognition 1H NMR study. *NMR Biomed.*, *18*, 24-33.
- Van de Peer, Y., & De Wachter, R. (1994). TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.*, *10*, 569-570.
- Van den Berg, R., Hoefsloot, H., Westerhuis, J., Smilde, A., & Van der Werf, M. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, *7*, 142-157.
- Van Der Greef, J., Martin, S., Juhasz, P., Adourian, A., & Plasterer, T. (2007). The art and practice of systems biology in medicine: mapping patterns of relationships. *J. Proteome Res.*, *6*, 1540-1559.
- Villas-Bôas, S., Koulman, A., & Lane, G. (2007b). Analytical methods from the perspective of method standardization. In: Topics in Current Genetics: Metabolomics. M.C. Jewett and J. Nielsen. In *Metabolomics*. Springer-Verlag, Berlin.
- Villas-Bôas, S., Noel, S., & Lane, G. (2006). Extracellular metabolomics: a metabolic footprinting approach to assess fiber degradation in complex media. *Anal Biochem*, *Anal. Biochem.*, *349*, 297-305.
- Villas-Bôas, S., Roessner, U., & Hansen, M. (2007a). *Metabolome Analysis: An Introduction*. Hoboken, New Jersey: Wiley-interscience.
- Visser, S. A. (1983). Application of van Krevelen's graphical-, statistical method for the study of aquatic humic material. *Environ. Sci. Techno.*, *17*, 412-417.
- Vogels, J. T., Tas, A. C., Van den Berg, F., & Van der Greef, J. (1993). A new method for classification of wines based on proton and carbon-13 NMR spectroscopy in combination with pattern recognition techniques. *Chemometrics and intelligent lab. Systems*, *21*, 249-258.
- Von Hippel, A., & Alger, R. (1949). A Precise Method of Determining the Faraday by Magnetic Resonance. *Phys. Rev.*, *76*, 1877.
- Vong, R., Geladi, P., Wold, S., & Esbensen, K. (1988). Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). *J. Chemometrics*, *2*, 281-286.
- Vriend, G., & Ott, M. (2006). Correcting ligands, metabolites, and pathways. *BMC Bioinformatics*, *7*, 517.

- Wang, C., Kong, H., Guan, Y., Yang, J., & Gu, J. (2005). Plasma phospholipid metabolic profiling and biomarkers of type 2 diabetes mellitus based on high-performance liquid chromatography/electrospray mass spectrometry and multivariate statistical analysis. *Anal. Chem.*, *77*, 4108-4116.
- Want, E., Nordstro, A., Morita, H., & Siuzdak, G. (2007). From exogenous to endogenous: The inevitable imprint of mass spectrometry in metabolomics. *J. Proteome Res.*, *6*.
- Wapstra, A., Audi, G., & Thibault, C. (2003). The AME2003 atomic mass evaluation (I). Evaluation of input data, adjustment procedures. *Nuclear Phys. A.*, *729*, 129-336.
- Ward, D., Cohan, F., Bhaya, D., Heidelberg, J., Köhl, M., & Grossman, A. (2007). Genomics, environmental genomics and the issue of microbial species. *Heredity*, 1-13.
- Waterhouse, A. L., & Towey, J. P. (1994). Oak Lactone Isomer Ratio Distinguishes between Wine Fermented in American and French Oak Barrels. *J. Agric. Food Chem.*, *42*, 1971-1974.
- Weckwert, W. (2003). Metabolomics in Systems. *Annual Review of Plant Biology*, *54*, 669-689.
- Weisberg, S. (1985). *Applied Linear Regression* (Vol. QA278.2 .W44 ). New York: Wiley.
- Westerhuis, J., Hoefsloot, H., Smit, S., Vis, D., Smilde, A., van Velzen, E., et al. (2008). Assessment of PLS-DA cross validation. *Metabolomics*, *4*, 81.
- Whitaker, R., Grogan, D., & Taylor, J. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*, *301*, 976-978.
- Whitfield, P.D.; German, A.J.; Nobel, P.J.M.;. (2004). Metabolomics: an emerging post-genomic tool for nutrition. *Brit J Nut* , *92*, 549-555.
- Wink, M. (1988). Plant breeding: importance of plant secondary metabolites for protection against pathogens and herbivores. *Theor Appl Genet* , *75*, 225-233.
- Wishart, D. (2005). Metabolomics: the principles and potential applications to transplantation. *Am J Transplant* , *5*, 2814-2820.
- Wold, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics*, *20*, 397-405.
- Wold, S., & al., e. (1998). Orthogonal signal correction of near-infrared spectra. *Chemometrics Intel. Lab. Syst.*, *44*, 175-185.
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Component Analysis. . *Chemometrics Intel. Lab. Syst.*, *2*, 37-52.
- Wold, S., Sjöström, M., & Eriksson, L. (2006). PLS in Chemistry. *The encyclopedia of computational chemistry*, 1998-2020.

- Wu, Z., Rodgers, R. P., & Marshall, A. G. (2004). Two-and Three-Dimensional van Krevelen Diagrams: A Graphical Analysis Complementary to the Kendrick Mass Plot for Sorting Elemental Compositions of Complex Organic Mixtures Based on Ultrahigh-Resolution Broadband Fourier Transform Ion Cyclotron R.M.M. 2004, *Anal. Chem.*, 2511-2516.
- Yang, J., Xu, G., Zheng, Y., Kong, H., & Pang, T. (2004). Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases. *J.Chromatogr. B. Anal. Tech. Biomed. Life Sci.*, 813, 59-65.
- Yoon, K. H., & al., e. (2006). Epidemic obesity and type 2 diabetes in Asia. 368.9548, *Lancet*, 1681-1688.
- Yuan, K., Kong, H., Guan, Y., Yang, J., & Xu, G. (2007). A GC-based metabonomics investigation of type 2 diabetes by organic acids metabolic profile. *Chromatogr. B. Analyt. Tech. Biomed. Life Science*. 850, 236