Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA

SCIENZE degli ALIMENTI

Ciclo XXI

(CHIM03)

NMR, METABONOMICS AND MOLECULAR PROFILES: APPLICATIONS TO THE QUALITY ASSESSMENT OF FOODSTUFFS

Presentata da: Gianfranco Picone

Coordinatore Dottorato

Prof. Claudio Cavani

Relatore

Prof. Francesco Capozzi and Dr.ssa Elena Babini

Esame finale anno 2009

To those who believe in me
and to me!

DIPARTIMENTO DI SCIENZE DEGLI ALIMENTI

Assessment for admission to the final examination for the degree of PhD in Food Science (XXI Cycle)

Dr. Gianfranco PICONE

The Ph.D. student focused his research work on the application of Nuclear Magnetic Resonance (**NMR**) spectroscopy to the analysis and to the quality control of foods, based on a metabonomic approach. This method can bring to several results:

1. classification of food: examples in this way are the classification of traditional foods and the discrimination between wild type and genetically modified foods based on their differences;
2. metabolites characterization: with the aim to describe foods from the content of metabolites;
3. identification of metabolites: with the purpose to identify those metabolites (biomarkers) able to explain the reasons of differences among groups. This can lead to better understand the possible effects of technological and biotechnological processes on foods, to assess food quality and to fight foods frauds and adulterations.

Moreover, the metabonomic approach combines the techniques of high resolution NMR with pattern recognition technology to rapidly evaluate the metabolic status of the biological matrix; in fact, when using spectroscopic techniques for chemical quantification, data obtained must be processed using univariate data analytical methods (**chemometrics**).
The research has been carried out using as food matrices: i) table grapes (*Vitis vinifera*) wild type with its genetically modified derivative and ii) fish from the bream family *Sparidae*, Gilthead Seabream (*Sparus aurata*).
The obtained results permitted, in the first case (wild type grape compared to the modified line) to evaluate the global screening of the metabolites' profile, showing changes not immediately expected as effect of the gene modification. In the second case, the results allowed to characterize the Gilthead Seabream from a metabolites point of view and to verify changes in their content due to different aquaculture systems.

In order to improve his knowledge on NMR spectroscopic data analysis, Dr. Gianfranco Picone has spent three months at the University of Copenhagen (Faculty of Life Science) and in order to improve his knowledge on the study of fish physiology, with special regard to the metabolism of lipids, he has spent three months at the Universidad de Murcia (Departamento de Biologia).
The results of this research activity will be subject of scientific international publications.

The Board unanimously agrees that Dr. Gianfranco PICONE is qualified to sit the final exam for the doctorate degree in Food Science.

Bologna, February 11st 2009

Coordinator of PhD Course in Food Science

Prof. Claudio CAVANI

ALMA MATER STUDIORUM • UNIVERSITA' DI BOLOGNA
VIALE G. FANIN , 44 – 40127 BOLOGNA - ITALIA - TEL. 0039 051 2096761 - FAX 0039 051 2096764
E-mail: DipScAlim.Dipartimento.SegretarioAmministrativo@unibo.it

ABSTRACT

Nuclear Magnetic Resonance (NMR) is a branch of spectroscopy that is based on the fact that many atomic nuclei may be oriented by a strong magnetic field and will absorb radiofrequency radiation at characteristic frequencies. The parameters that can be measured on the resulting spectral lines (line positions, intensities, line widths, multiplicities and transients in time-dependent experiments) can be interpreted in terms of molecular structure, conformation, molecular motion and other rate processes. In this way, high resolution (HR) NMR allows performing qualitative and quantitative analysis of samples in solution, in order to determine the structure of molecules in solution and not only. In the past, high-field NMR spectroscopy has mainly concerned with the elucidation of chemical structure in solution, but today is emerging as a powerful exploratory tool for probing biochemical and physical processes. It represents a versatile tool for the analysis of foods. In literature many NMR studies have been reported on different type of food such as wine, olive oil, coffee, fruit juices, milk, meat, egg, starch granules, flour, etc using different NMR techniques.

Traditionally, univariate analytical methods have been used to explore spectroscopic data. This method is useful to measure or to select a single descriptive variable from the whole spectrum and , at the end, only this variable is analyzed. This univariate methods approach, applied to HR-NMR data, lead to different problems due especially to the complexity of an NMR spectrum. In fact, the latter is composed of different signals belonging to different molecules, but it is also true that the same molecules can be represented by different signals, generally strongly correlated. The univariate methods, in this case, takes in account only one or a few variables, causing a loss of information.

Thus, when dealing with complex samples like foodstuff, univariate analysis of spectra data results not enough powerful. Spectra need

to be considered in their wholeness and, for analysing them, it must be taken in consideration the whole data matrix: chemometric methods are designed to treat such multivariate data.

Multivariate data analysis is used for a number of distinct, different purposes and the aims can be divided into three main groups:

- data description (explorative data structure modelling of any generic n-dimensional data matrix, PCA for example);
- regression and prediction (PLS);
- classification and prediction of class belongings for new samples (LDA and PLS-DA and ECVA).

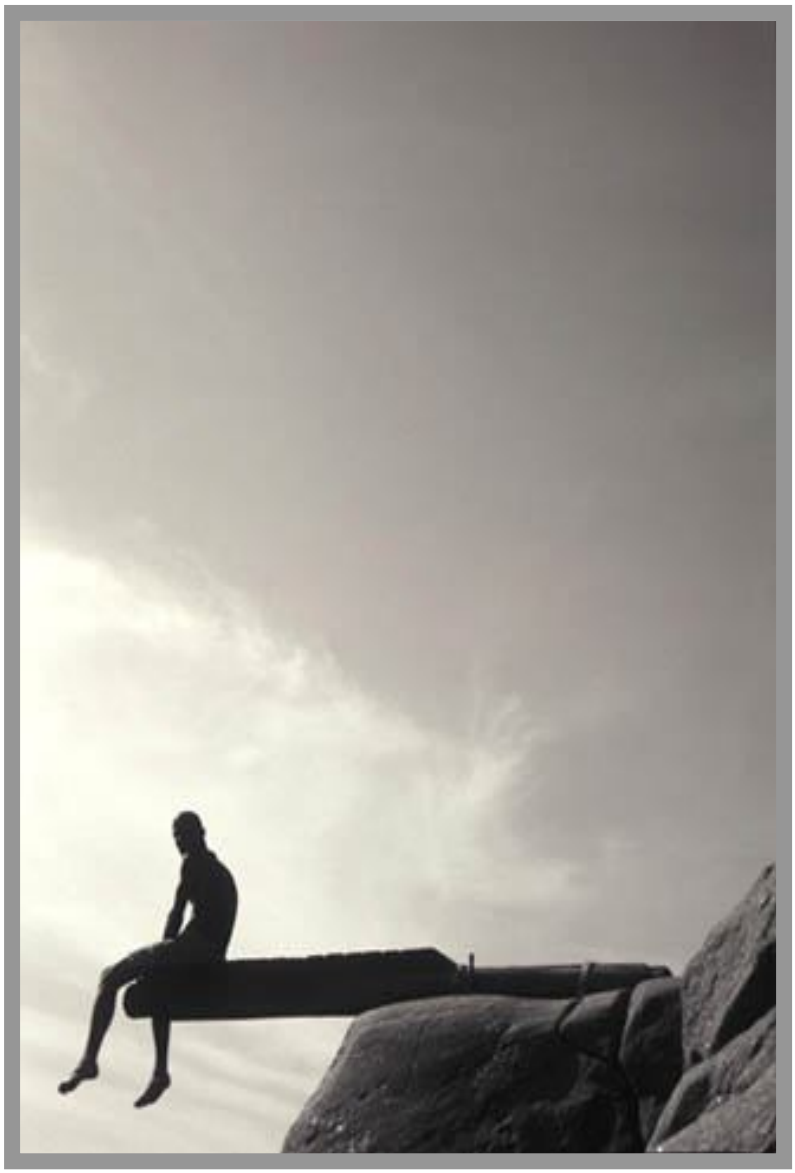
The aim of this PhD thesis was to verify the possibility of identifying and classifying plants or foodstuffs, in different classes, based on the concerted variation in metabolite levels, detected by NMR spectra and using the multivariate data analysis as a tool to interpret NMR information.

It is important to underline that the results obtained are useful to point out the metabolic consequences of a specific modification on foodstuffs, avoiding the use of a targeted analysis for the different metabolites. The data analysis is performed by applying chemometric multivariate techniques to the NMR dataset of spectra acquired. The research work presented in this thesis is the result of a three years PhD study. This thesis reports the main results obtained from these two main activities:

A1) Evaluation of a data pre-processing system in order to minimize unwanted sources of variations, due to different instrumental set up, manual spectra processing and to sample preparations artefacts;

A2) Application of multivariate chemiometric models in data analysis.

CHAPTER ONE
- INTRODUCTION -



1 INTRODUCTION

To define the total quality of food becomes really hard because involves several aspects including both social-economical (S. U. O'Hara et *al.*, 2001) and scientific-technological ones, like organoleptic and sensory attributes (N. A. Abumrad, 2005), food safety (A. Alimelli et *al.*, 2007), nutritional value (I. Andersona et *al.*, 1992), functionality (A. Becalsky et *al.*, 2003), service and stability (A. E. Bender, 1987), psychological factors (A. J. Blood et *al.*, 1999). Thus, when the term "quality" is defined, it must be taken in consideration the "consumers-producers" needs together with the "science-technological" contributes. So, at the end, the total food quality (**TFQ**) is a sum of different characteristics and they can be represented by a **subjective** and **objective** concepts.

From a subjective concept point of view, "the consumer is the main instrument of appraisal of the sensory properties of the product, immediately perceived by the human senses" (A. M. Giusti et *al.*, 2007). Thus, quality has become of utmost importance to society especially in the last decades during which consumers have become more conscious of quality and organizations are judged more on their overall quality performance instead of their financial performance alone. During this period, the most drastic change in quality thinking is the change from production-oriented to customer-oriented concepts. This moving is also strictly connected to the entrance of large countries like China, India, South America in to the global marketing. A consequence of this entrance is the introducing of own traditions, new raw materials, new transformation technologies, and different fashions to gastronomic preparations (D. Roger et *al.*, 2000). This new territories moving towards the Western countries brought the consumers to manifest a new awareness of the influence of food on mood, efficiency at work and life-style but above all on health. This especially holds after the latest events, like the detection of the industrial chemical melamine in

contaminated chinese formula milk (G. S. Becker, 2008). Together with the consumers, on the other side of the chain, there are the producers who are demanding more protectionism from supranational organizations, especially for local specialties.

From an objective concept point of view, characteristics not related to the human senses like nutritional quality (E. M. Mojduszka et *al.*, 1999), safety of use (B-H Cho et *al.*, 2008), shelf-life (M. Bill et *al.*, 1998) and technological proprieties of foods (M. Servilli et *al.*, 2004) are function of their composition. In fact, the capability to be transformed and the nutritional value (K. Dewettinck et *al.*, 2008) are strictly related to the chemical composition and concentration of certain metabolites and as consequence to the quality (objectively intended); the presence, the absence or even the bio-technology modification of particularly metabolites can hardly altered the inner composition, increasing or not the biological/nutritional value of food; a typical example in this way is genetically modified food (P. Tenbült et *al.*, 2008). For this reason, considering the whole total **molecular profile** can be the right tool for the evaluation of food's chemical-nutritional quality (D. Barber et *al.*, 2008).

1.1 THE MOLECULAR PROFILE OF FOOD

Considering the objective concepts of TFQ, foods are complex structure derived from plants, carcasses of animals, and single-cell organisms. In this structure water (R. R. Ruan et *al.*, 1998), saccharides (V. M. F. Lai et *al.*, 2004), proteins (M. A. Lluch et *al.*, 2001), lipids (P. Benatti et *al.*, 2004), and mineral compounds (N. Michał, 2007) constitute the main building materials (Table 1.1) and are responsible for the nutritional value and most sensory properties of foodstuffs (Z. E. Sikorski, 2007). All these nutrients are highly correlated with each other in their effect and many display interactive and synergistic effects and provide the human body with the necessary building material and source of energy, as well

as elements and compounds indispensable for metabolism like colorants, flavor compounds, vitamins, prebiotics, probiotics and additives.

Table 1.1: typical Products as Rich Sources of the Main Food Components (Table from Z. E., Sikorski (2007). *Chemical and Functional Properties of Food Components*, *Third Edition*. Ed., CRC Press Taylor & Francis Group, Boca Raton (FL), London, New York, Introduction

Water	Saccharides	Proteins	Lipids	Minerals	Vitamins
Juices	Saccharose	Soybean	Oils	Nuts	Fish
Fruits	Honey	Beans	Lard	Fish products	Vegetables
Milk	Cereals	Meat	Butter	Cereals	Meat
Vegetables	Chocolate	Fish	Pork	Vegetables	Fruits
Jellies	Potato	Wheat	Chocolate	Dairy products	Cereals
Lean Fish	Manioc	Cheese	Nuts	Fruits	Fish liver
Lean Meat	Fruits	Eggs	Egg yolk	Meat	Yeast

These constituents, present in lower quantities, also contribute to perform and to complete the foodstuff's structure. Some of them are nutritionally essential and they are presented in the raw materials like vitamins. In addition, others either serve no role in human metabolism or for which the role has not yet been elucidated (C. J. Seal et *al.*, 2007).

Although some of these nutrients are not indispensable they can be utilized by the body, including most free amino acids, or impart desirable sensory properties to food products (E. A. Decker et *al.*, 2000). Other minor components are useless or even harmful if present in excessive amounts. A variety of compounds are added in-

tentionally during processing, to be used as preservatives, antioxidants, colorants, flavorings, sweeteners, and emulsifying agents, or to fulfill other technological purposes (E. S. Zdzisław et *al.*, 2007).

The nutritional value of foods depends, thus, primarily on the levels of nutrients present in the raw materials, from freshness, especially in the case of numerous species of vegetables, fruits, and seafood and then from components added during the transformation processing. In this last case, the industry processing may also increase the biological value of food by inducing chemical and genetic changes (A. M. Giusti et *al.*, 2007) whose extent depends on the chemical properties of food components, on the conditions of storage and on the parameters of freezing, salting, drying, smoking, marinating, frying, cooking, and other methods of preservation or processing.

On the other side, there are also nutritionally undesirable side effects, such as destruction of essential food components as a result of heating, chemical treatment, oxidation and interactions among all the components. In facts, these compounds undergo various biochemical and chemical changes during postharvest storage and processing of raw materials.

Last but not the least, nutritionally alteration due to metabolic profile changes can be induced by biotechnology (genetic) modification. The Convention on Biological Diversity (CBD 2000) defines biotechnology as “any technological application that uses biological systems, living organisms or derivatives thereof, to make or modify products or processes for specific use”. It thus includes activities such as traditional food fermentations, waste treatment, drug development, fish farming and crop development.

Food biotechnology has been defined as “the application of biological techniques to food crops, animals and microorganisms to improve the quality, quantity, safety, ease of processing and production economics of food. It thus includes the traditional manufactur-

ing processes used for bread, beer, cheese and various fermented milk products" (IFST 2004).

1.2 MOLECULAR PROFILING TECNIQUES FOR FOOD QUALITY AND SAFETY ASSESMENT

To ensure healthy and tasty food, to increase consumer trust and provide competivity to the food industry and a competitive food sector, more information about food quality is required. Also these kind of information can be represented by a **subjective** and **objective** concepts. In this last case, the heterogeneous food material and its production is complex so the research process itself is further difficult. If to define the term "quality and safety" mean to involve chemical, biochemical, agronomic and technologic aspects, also researcher techniques must take in consideration them. It does not exist an unic "standard of measure", but several of them are combined together to define the entire concept of **TFQ**. Thus, the profile methods can be completed only if it has been taken in consideration the following aspects:

1. **Proteomics**
2. **Metabonomics**
3. **Chemometrics (data analysis)**

In case of biotechnology (genetic) modification it also important and fundamental take in consideration the following fourth aspect

4. **Genomics (detection of altered gene expression)**

Even if genomics is listed as the fourth elements to consider in the entire concept of the TFQ, it represents the beginning of the metabolic pathways (Figure 1.1), along which is possible to take in consideration all the aspects that define the concept of quality in food.

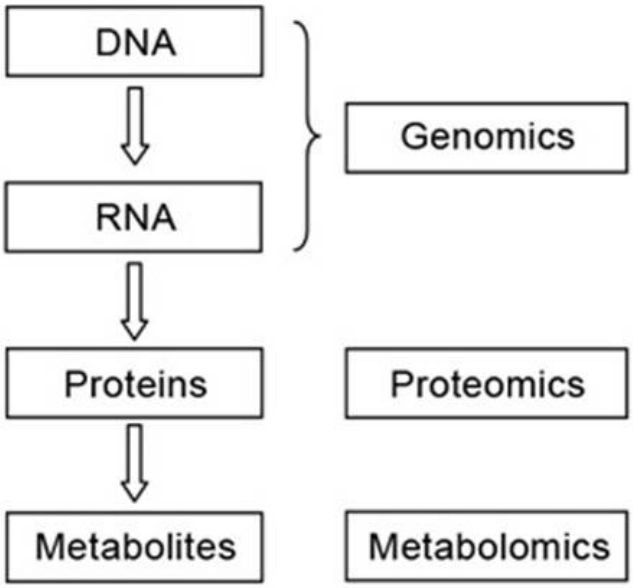


Figure 1.1: evolution of Genomics to Metabolomics through Proteomics. **Genomics:** mapping the entire genome (DNA and RNA sequences). **Proteomics**: identifying, sequencing and characterizing the functional protein network (the "proteome" or protein complement generated from the genome). **Metabolomics**: the comprehensive analysis of the whole metabolome under a given set of conditions (Figure is adapted from R. M., Adibhatla, J., Hatcher and R., Dempsey (2006). *Lipids and Lipidomics in Brain Injury and Diseases.* AAPS Journal, 8, E314-E321)

1.2.1 Genomics (Detection of altered gene expression)

The Webster's New World™ Medical Dictionary (A. a. v. v., 2008) defines **genomics** as "the study of genes and their function. Genomics aims to understand the structure of the genome, including the mapping genes and sequencing the DNA [...]". Genomics includes:

1. **functional genomics**: the characterization of genes and their mRNA and protein products;
2. **structural genomics**: the dissection of the architectural features of genes and chromosomes;
3. **comparative genomics**: the evolutionary relationships between the genes and proteins of different species

This new discipline is born "from a marriage of molecular and cell biology with classical genetics" (V. A. McKusick et *al.*, 1987), and, at the end, it represents the beginning of the information to define

the molecular profile a biological system: from men, to animals across plants and foods. While in medical research, genomics had been a fast evolution, in nutrition and especially in food research it is almost a brand new approach. In fact, in the past, nutrition and food research focused the attention almost exclusively on measuring the presence or absence of metabolites or their concentration. However, as there are important levels of biological regulation beyond gene and protein expression (protein–protein interactions and alterations of protein activity by metabolic intermediates), the final stage along the line, from gene to mRNA, to protein to function, is, therefore, the analysis of the pattern and the concentrations of the metabolites that flow between the proteins, organelles, cells and organs (M. J. Rist et *al.*, 2006). In this way nutrition and food scientists will analyze metabolites but this time the analysis of the metabolome will comprise the sum of all detectable low- and intermediate-molecular-weight compounds rather than individual metabolites. Anyway, this kind of approach is not unilateral, but measuring metabolites' levels can give functional elucidation of genes' role and it becomes a powerful tool for the comprehension of the metabolic pathways into bacterial systems, yeast and plants – both wild type and transgenic organisms. In the first case (**bacterial systems and yeasts**), the functional genomics is above all adopted for the genetic screening of a large range of microorganisms (O. P. Kuipers, 1999); in fact, genomics of food microbes generates valuable knowledge that can be used for metabolic engineering, improving cell factories and development of novel preservation methods. This method can provide an understanding of how microorganism genes respond to environmental influences (i.e. by their expression) in different situations or ecologies, and should therefore allow adaptation of conditions to improve food technological processes. Functional genomics can, for example, shed light on common genetic mechanisms which enable microorganisms to use certain sugars during fermentation, as well as on genetic differences allowing some strains to perform better than others. It holds great potential for defining and modifying elusive metabolic

mechanisms used by microorganisms. Moving from the gene to the protein level, it should also be mentioned that proteomics and is also a very active area of research which offers potential for improving fermentation technologies (FAO, 2004).

In the second case (**plants**) genomics is used to refers the study of the way genes and genetic information are organized within the genome, the methods of collecting and analyzing this information, and how this organization determines their biological functionality (H. Campos-de Qiroz, 2002). Plant genomics is reversing the previous paradigm of identifying genes behind biological functions and instead focuses on finding **biological functions behind genes**. This can leads to create a data bank of useful genes that can be employed to improve plants such as crops and trees. It also reduces the gap between phenotype and genotype and helps to comprehend not only the isolated effect of a gene, but also the way its genetic context and its genetic networks, it interacts with, can modulate its activity. In this way is possible to understand gene-gene iterations, but above all how environmental or biotechnological changes on a gene can influenced the whole genetic context and networks and as consequence the molecular/metabolic profiling.

1.2.2 Proteomics

Proteomics is the study of the entire set of proteins present in a cell, organism or tissue under defined conditions (K. Belhajjame et *al.*, 2005) with the aim of understanding the behavior of these proteins under varying environments and conditions. This kind of approach, at the beginning, has been used in medical environment; it was developed to identify proteins associated with a disease and then potential new drugs for the treatment of disease. The study, in summary, try to design drugs to interfere with the action of a protein on the base of its 3D structure with the help of computer technique which attempts to fit millions of small molecules to the three-dimensional structure of a protein. This approach is called

"**virtual ligand screening**" (G. Klebe, 2005) and it is represented in Figure 1.2.

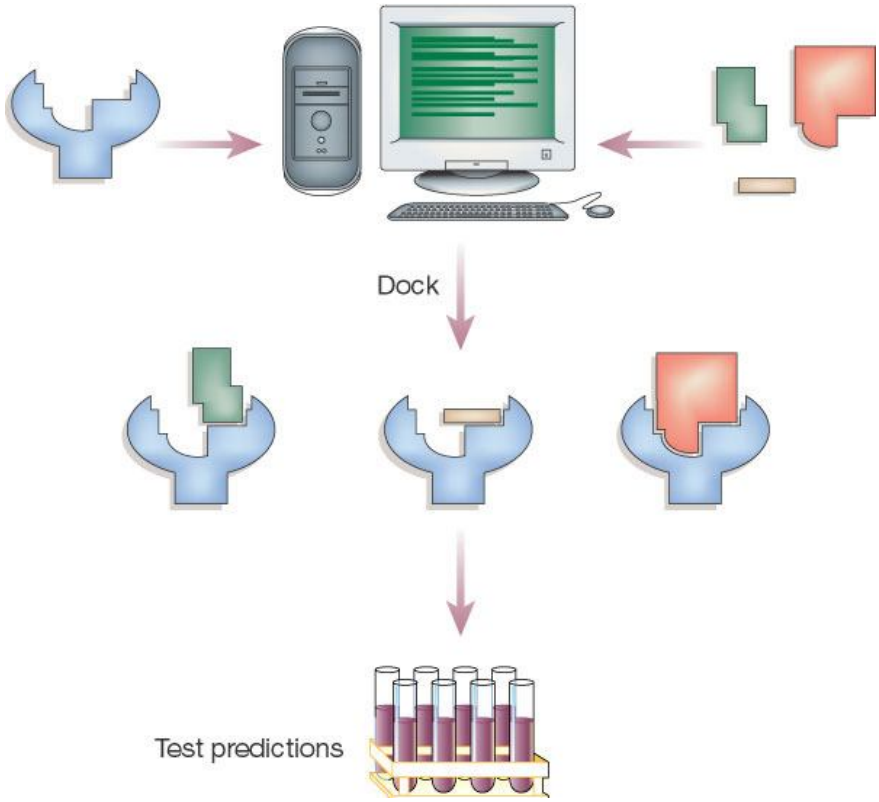


Figure 1.2: large libraries of available, often purchasable, compounds are docked into the structure of receptor targets by a docking computer program. Each compound is sampled in thousands to millions of possible configurations and scored on the basis of its complementarity to the receptor. Of the hundreds of thousands of molecules in the library, tens of top-scoring predicted ligands (hits) are subsequently tested for activity in an experimental assay (Figure is taken, with concession from the author, from the following article: B. K., Shoichet (2004). *Virtual screening of chemical libraries*. Nature 432, 862-865)

The computer rates the quality of the fit to various sites in the protein, with the goal of either enhancing or disabling the function of the protein, depending on its function in the cell. A good example of this is the identification of new drugs to target and inactivate the HIV-1 protease (W. Wang et *al.*, 2001). Nowadays, the potential of proteomics is adopted also in nutrition and food research, where it is increasingly being recognized. Proteomics is employed to address questions of nutrition and health, based on the concept that foods and drinks affect individual consumers differently (B., de Roos et *al.*, 2008). Questions like "why does a food well tolerated by one individual cause violent gastric discomfort in another?",

"are there biomarkers for food preference?", "what genes are activated by specific foods in procuring health and wellness?" can find an answer in protein markers that indicate such predisposition before disease symptoms arise. This is the basis of new nutritional-discovery tools, also indicated by the term **nutrigenomics** (R. M. Elliott et *al*., 2006). The main focus of nutrigenomics is on how diet regulates gene function (transcription and translation) and metabolism (i.e. diet → gene interactions) and its aim is to find new specific nutritional diet to inactivate proteins involved in diseas.

In food chemistry, proteomic techniques offer a new very promising approach to identify protein in food matrix and to study protein–protein interactions in raw and processed foods, as well as interactions between proteins and other food components (M. Carbonaro, 2004). In Table 1.2 are listed specific applications of proteomics in analysis of food quality.

Table 1.2: examples of application of proteomics to food quality assessment (From: M., Carbonaro (2004). *Proteomics: present and future in food quality evaluation*. Trends in Food Science & Technology 15, 209–216)

Application study	Reference
• Comparison of meat species	• P. Roncada et *al*., 2002
• Post mortem changes in porcine meat	• R. Lametsch et *al*.., 2001
• Determination of wheat quality	• D. M. Gottlieb et *al*., 2002
• Analysis of wheat kernel amphiphilic proteins	• N. Amiour et *al*., 2002
• Glutenin subunit mapping	• R. Cozzolino et *al*., 2001
• Metabolic pathways in rice	• A. Koller et *al*., 2002
• Tomato protein expression under heat stress	• Y. Iwahashi et *al*., 2000
• Identification of hazelnut 11 S allergen	• K. Beyer et *al*., 2002
• Markers of sesame seed allergens	• K. Beyer et *al*., 2002
• Immunological analysis of shrimp allergens	• C. J. Yu et *al*., 2003
• Map of commercial bovine milk	• M. Galvani et *al*., 2001
• Collection of bioactive peptides of b-casein	• P. G. Righetti et *al*., 1997
• Reference map of fat globule membrane proteins	• S. Quaranta et *al*., 2001
• Bioavailability of milk proteins	• M. Carbonaro et *al*., 2003

In **food quality**, proteomics researches focus the attention especially on the alteration of food due to specific conditions. It is largely applied especially in **meat science**. In fact, because muscle is mainly composed of water and proteins, it is evident that prote-

ome analysis can give much information on structures and functions of proteins involved in several mechanisms which determine meat quality. Furthermore, as meat quality is strictly connected to the meat tenderness, understanding the factors involved in the development of tenderness, like the post-mortem glycolysis and proteolysis (X. Jia et *al*., 2006), in muscle tissues is a major concern for the beef-producing industry and for consumers requests.

Like meat science, also **cereal science** finds proteomics as a useful instruments to understand physiological and technological function of proteins of wheat kernels (N. Amiour et *al*., 2002). This approach provides some very useful information about proteic components linked to bread wheat quality and, therefore, to predict the quality of bread. Particularly interesting is the application of proteomics **in food allergy prevention** (K. Beyer, et *al*., 2002): this kind of study allows to identify several allergenic proteins in samples, indicating the potential of proteomic approaches to survey food samples with regard to the occurrence of allergens. This method has been applied systematically to identify and profile proteins expressed above all during seed development (K. R. Kottapalli et *al*., 2008) or in the mature seed of model plant species like soybean (M. Hajduch et *al*., 2005), rapeseed (M. Hajduch et *al*., 2006), Medicago (K. Gallardo et *al*., 2003), Arabidopsis (K. Gallardo et *al*., 2002), wheat (D. J. Skylas et *al*., 2001) and barley (T. Majoul, et *al*., 2003).

In all food cases, the main technological approach (**experimental proteomics**) currently applied provides at least five steps (H. A. Kuiper et *al*., 2003) including one or often two-dimensional gel electrophoresis (M. Berth et *al*., 2007) followed by excision of protein spots from the gel, digestion into fragments by specific proteases, analysis by mass spectrometry (J**.** S**.**, Anderson et *al*., 2000) as represented in Figure 1.3 and subsequent bioinformatics tools, to match the resulted fragments with information about known proteins.

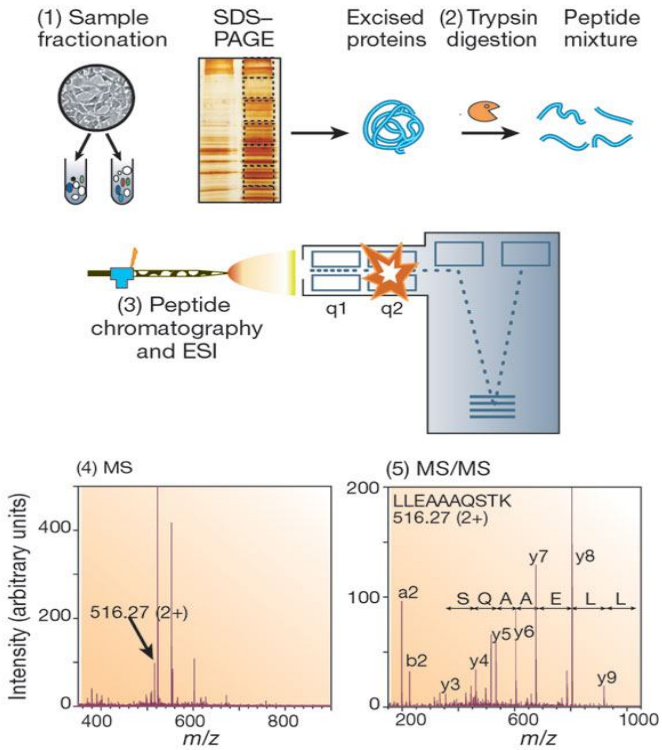


Figure 1.3: the typical proteomics experiment consists of five stages: **1**, the proteins to be analyzed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. This often includes a final step of one-dimensional or two dimensional gel electrophoresis. Therefore, proteins are degraded enzymatically to peptides in stage **2**, usually by trypsin, In stage **3**, the peptides are separated by one or more steps of HPLC and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and, in stage **4**, a mass spectrum of the peptides eluting at this time point is taken. The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or 'MS/MS' experiments ensues (stage **5**). These consist of isolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of the tandem or MS/MS spectrum. The MS and MS/MS spectra are typically acquired for about one second each and stored for matching against protein sequence databases. The outcome of the experiment is the identity of the peptides and therefore the proteins making up the purified protein population (Figure is taken, with concession from the author**,** from the following article: R., Aebersold and M., Mann (2003). *Mass spectrometry-based proteomics*. Nature 422, 198-207)

1.2.3 Metabonomics

Metabonomics is the study of metabolites and their role in various physiological states; it is a novel methodology arising from the post-genomics era and has extensive biomedical application. It can be expressed as "the quantitative measurement of the dynamic multi-parametric metabolic response of living systems to pathophysiological stimuli or genetic modifications" (J. K. Nicholson et *al.*, 1999) and, as a holistic approach, it detects, quantifies and catalogues the time related metabolic processes of an integrated bio-

logical system, ultimately, relates such processes to the trajectories of the physiological and pathophysiological events (H. Tang, 2006); however, a metabonomic study can provide significant results only if the metabolic changes in a target group is significantly different from the biological variation of the relative control group (I. B. Abdel-Farid et *al.*, 2007). Different from **metabolomics**, that focuses on high-throughput characterization of small molecule metabolites in biological matrices (D. S. Wishart, 2008), in order to have a complete set (**molecular profile**) of the molecules characterizing the matrices, metabonomics try to find out how the molecular profile responds (changes) to certain external factors. Complementary to proteomics and genomics, metabonomics has been first widely applied to a wide range of problems in diverse biomedical research areas with the aim to understand the metabolites's behavior under certain exogenous conditions. Different applications of the metabonomic approach have been documented, especially in toxicity screening (Z. Xiaoyu et *al.*, 2006), drug metabolism (H. C. Keun, 2006) and functional genomics (C. L. Gavaghan et *al.*, 2002). In summary, "**Metabonomics thus allows real-world, medical observations to be related to data from all the other 'omics' technologies, which are less directly related to actual biological outcomes than metabolism is** (Figure 1.4)" (J. K., Nicholson et *al.*, 2008).

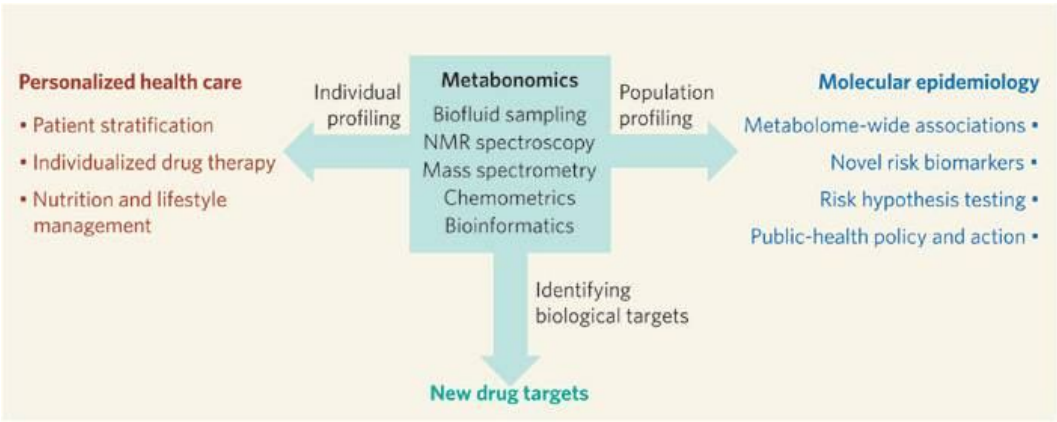


Figure 1.4: three broad areas that might benefit from metabonomics. (Figure is taken, with concession from the author, from the following article: J. K., Nicholson and J. C., Lindon (2008). *Systems biology: Metabonomics*. Nature 455, 1054-1056)

As nutrients and non-nutrients, amongst many exogenous factors, in food have also important effects on the biochemistry of human and associated micro flora, as well as proteomics, metabonomics finds an increasing development also in nutritional science (K. S. Solanky et *al.*, 2005; M. J. Gibney et *al.*, 2005) and in recent time in food chemistry (C. Castro et *al.*, 2007). From a nutritional point of view, the extension of this approach to human nutrition offers enormous potential (evaluation of different biofluids in nutritional metabonomics, of non nutrient chemicals and large-bowel metabolites) and it opens the door to studying different aspects of ''molecular nutrition'' more strictly connected to food quality, including:

1. food component analysis;
2. food quality/authenticity detection;
3. food consumption monitoring;
4. physiological monitoring in food intervention or diet challenge studies.

In fact, traditionally food component analysis involves identifying and classifying food constituents into very broad categories such as proteins, fats, carbohydrates, fiber, vitamins, trace elements, solids and/or ash. However, with the advent of metabolomics and metabonomics, foods and beverages are now being analyzed with considerably more chemical detail (D. S. Wishart, 2008) with hundreds or even thousands of distinct chemical identities being detected and/or identified in certain foods. From this perspective (**metabolomics researcher**) most foods can essentially be viewed as complex chemical mixtures consisting of various metabolites and chemical additives in a solid, semi-solid or liquid matrix as shown in the following examples:

1. the milk metabolome contains more than 200 different oligosaccharides (M. R. Ninonuevo et *al.*, 2006);

2. the edible plant metabolome likely consists of more than 10,000 different detectable compounds with more than 2000 nutrient metabolites and more than 8000 non-nutrient phytochemicals having already been identified (T. M. Ehrman et *al.*, 2007).

These examples want to underline the high capacity from one side of metabolomics to identify hundreds of metabolites in highly complex biological matrix; moreover, on the other side, the capability of metabonomics, for this intrinsic characteristic, to assess perturbations in metabolic pathways. In this way, metabonomics becomes a powerful tool to examine underlying biology and mechanisms of disease, for drug development and toxicity studies, for assessing the effects of environmental changes and even for the development of control methods for both food adulteration and food quality (D. S. Wishart, 2008). For instance, orange juice can be blended with lower-cost grapefruit juice without any obvious changes to flavor or color. However, the presence of grapefruit juice in a presumptively pure orange juice product can have serious consequences to individuals on certain medications. In particular, the detection of adulterated or contaminated food products (**frauds**), often very difficult to detect via taste or color, is possible for the presence of characteristic chemicals or certain concentrations of chemicals in certain types of juices, extracts and oils (N. Ogrinc et *al.*, 2003). Food quality assessment also impacts food quality control. In fact, metabonomic techniques may find their greatest use in the food industry in monitoring quality control or batch-to batch product reproducibility. Both in human nutrition and in food chemistry, the metabonomics approach is widely developing above all for the improvement of advanced analytical technologies (I. D. Wilson, 2007). Different platform have been employed to investigate on the metabolites' profile such as:

- Nuclear Magnetic Resonance (**NMR**);
- Gas Chromatography Mass Spectrometry (**GC-MS**);
- Liquid Chromatography Mass Spectrometry (**LC-MS**).

In Table 1.3 are listed the main advantages and disadvantages that can be detected using above listed analytical techniques, whilst in Figure 1.5 are illustrated the sensitivities of the techniques.

Table 1.3: comparison of different metabolomics technologies (From D. S. Wishart (2008). *Metabolomics: applications to food science and nutrition research*. Trends in Food Science & Technology, 19, 482–493)

Metabonomics Technologies		
Technology	**Advantages**	**Disadvantages**
NMR	• Quantitative • Non-destructive • Fast (2-3 min/sample) • Requires no derivitization • Requires no separation • Detects all organic classes • Allows ID of novel chemicals • Robust, mature technology • Can be used for metabolite imaging (fMRI) • Large body of software and databases for metabolite ID • Compatible with liquids and solids	• Not very sensitive • Expensive instrumentation • Large instrument footprint • Cannot detect or ID salts and inorganic ions • Cannot detect non-protonated compounds • Requires larger (0.5 mL) samples
GC MS	• Robust, mature technology • Relatively inexpensive • Quantitative (with calibration) • Modest sample size need • Good sensitivity • Large body of software and databases for metabolite ID • Detects most organic and some inorganic molecules • Excellent separation reproducibility	• Sample not recoverable • Requires sample derivitization • Requires separation • Slow (20-30 min/sample) • Cannot be used in imaging • Novel compound ID is difficult
LC MS	• Superb sensitivity • Very flexible technology • Detects most organic and some inorganic molecules • Minimal sample size requirement • Can be used in metabolite imaging (MALDI) • Can be done without separation (direct injection) • Has potential for detecting largest portion of metabolome	• Sample not recoverable • Not very quantitative • Expensive instrumentation • Slow (20-30 min/sample) • Poor separation resolution and reproducibility (vs. GC) • Less robust instrumentation than NMR or GC-MS • Limited body of software and databases for metabolite ID • Novel compound ID is difficult

In particular, the NMR technique has the advantage to supply detailed information on the molecular structure of the biological material observed, reflecting at the end the metabonomic status of a biological living system, without losing important information on the system.

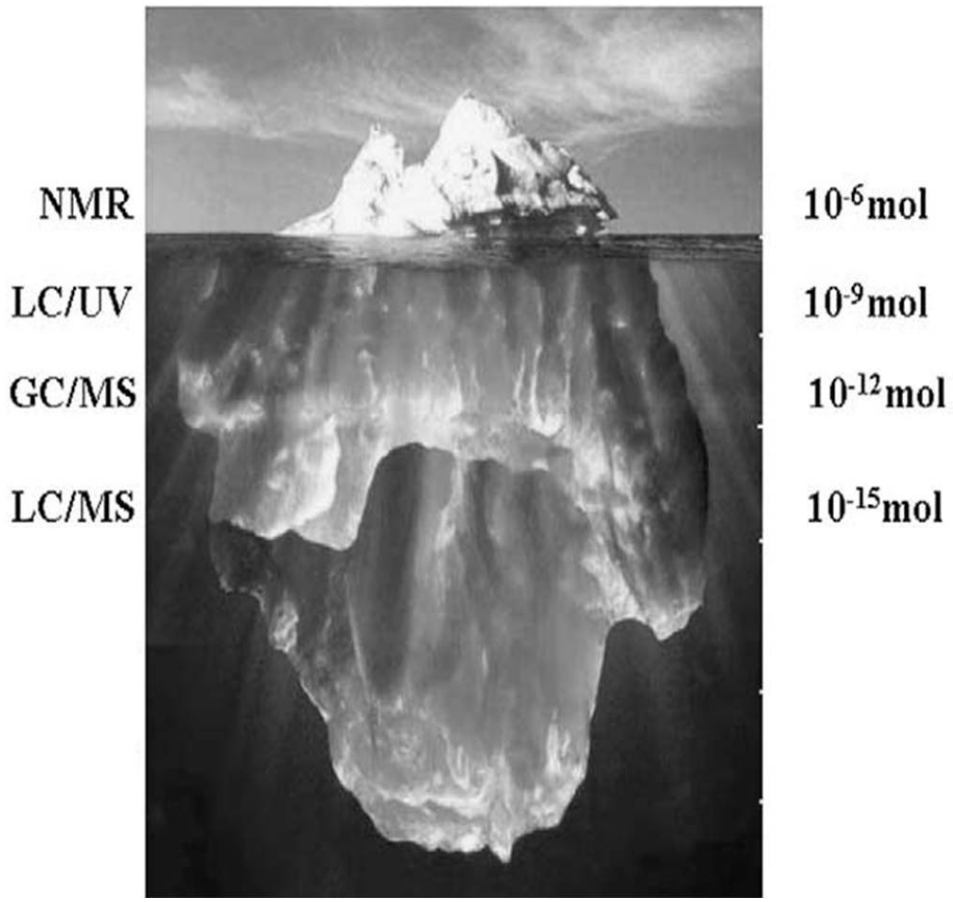


Figure 1.5: a comparison of the relative sensitivities of various metabolomic tools. NMR has rapid analysis times but suffers from lower sensitivity thus allowing visualization only of the more concentrated metabolites (i.e. the tip of the iceberg). GC/MS and HPLC/MS provide good selectivity and sensitivity (Figure is adapted from L. W., Sumner, P., Mendes and R. A., Dixon (2002). *Plant metabolomics: large-scale phytochemistry in the functional genomics era.* Review. Phytochemistry, 62, 817–836)

NMR is generally used to detect hydrogen atoms in metabolites (^{1}H NMR); thus, in a typical sample (biological fluid in medical research or organic extract in food research) all hydrogen-containing molecules (almost all metabolites) will give an ^{1}H NMR spectrum, as long as they are present in concentrations above the detection limit (NMR characteristics are going to be discussed in par 1.3).

The NMR spectrum results to be in this way the superposition, commonly called also **fingerprint** (I. F. Duarte et *al.*, 2002; F. M. Amaral et *al.*, 2005), of the spectra of all of the metabolites in the sample. An advantage of NMR is that the sample doesn't require any physical or chemical treatment prior to the analysis, but only the solution conditions such as the temperature, pH and salt concentration have to be adjusted so as to closely mimic a given physiological fluid, especially in the case of proteins' study (K. Wüthrich, 2002). On the other side, the MS studies usually require the metabolites to be separated from the sample before detection, by using HPLC (I. D. Wilson et *al.*, 2005) or the metabolites can be chemically modified to make them volatile, so to be used in GC–MS (R. Goodacre et *al.*, 2004).

1.2.4 Data analysis (chemometrics methods)

As has been told in previous paragraph, a metabonomics study can provide significant results only if the metabolic changes in a target group is significantly different from the biological variation of the relative control group (I. B. Abdel-Farid et *al.*, 2007). This means that the spectra of samples of interest are compared with those from controls, so that the spectral features caused by external factors can be determined.

This approach leads to ignore the importance of the concentration of the metabolites. This is in part true: precise metabolite concentrations are not always necessary to formulate hypotheses about the mechanism of changes, especially when it talks about diseases; so metabonomics analysis becomes a way to **clusterized** or **classify** (Y. Wang et *al.*, 2004) on the base of differences (in term of presence or not of metabolites) between control and target samples.

But, however, if only a few metabolites turn out to be important then knowledge of their concentrations might be instructive. Thus, after their identifications, they can subsequently be measured and used like **bio-chemical markers**. In this way metabonomics (also

called **quantitative metabolomics** or **targeted profiling**) becomes a **predictive** tool to predict property of interest (typically adherence to a performance standard) and in the same a time an instrument that can provide elements for the distinction among samples that show same characteristics (O. Moreira Sampaio et *al.*, 2008).

This predictive tools, has its advantages and disadvantages, but, given the importance attached to bioactive compound identification, there is a growing preference for quantitative metabolomics in many areas of food science and nutrition research (M. J. Gibney et *al.*, 2005; S. Moco et *al.*, 2006).

Pattern recognition (A. K. Jain et *al.*, 2000) is the common name used to identify methods that lead to organize the metabonomics data in order to classify samples in classes and to predict or classify unknown samples in one of the known classes on the basis of its pattern of measurements (L. A. Berrueta et *al.*, 2007). It is mainly based on mathematical and statistical procedures (**multivariate analysis** or **chemometrics**) that allow to analyze the great amounts of information from even few samples (variables or features), produced by modern analytical instruments (like NMR), in relatively short time (I. F. Duarte et *al.*, 2004).

This leads to the availability of compressed multivariate data matrices from which is possible to extract the maximum useful information. Thus, multivariate analysis becomes a good tool to explore all data come from the "-omics" technologies, because, as it has been said in previous paragraphs, the "–omics" raw data are generally from spectroscopic measurements (M. A. Al-Holy et *al.*, 2006; M. Bunzel et *al.*, 2006; B. Cavaliere et *al.*, 2008) that give a lot of information.

At the end, from a spectroscopic measurement, for example an ^{1}H NMR spectrum, is possible to select one or more spectral regions or specific signals that often provide better results in classifications or prediction than the use of the entire spectra (Figure 1.6).

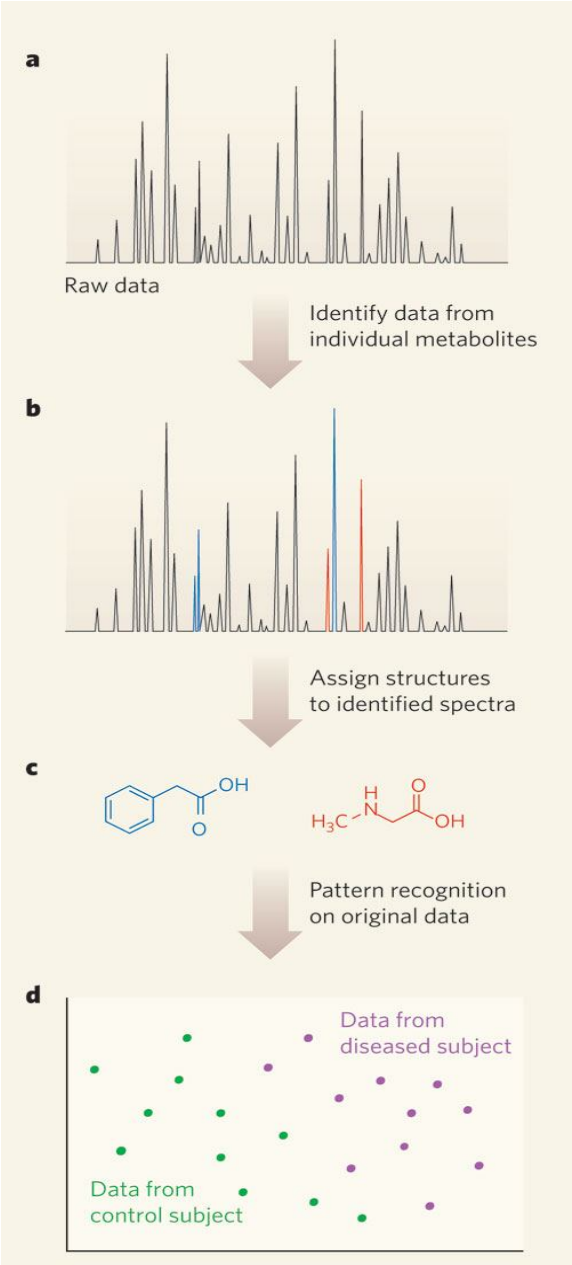


Figure 1.6: metabonomics data analysis: **a**, typical procedure might start with the NMR spectrum of a sample, which contains signals from hundreds of metabolites. **b**, The individual spectra for each metabolite are identified. **c**, This enables the structure of the metabolites to be determined. **d**, Pattern-recognition techniques can be used to work out how the spectra from two groups differ each other. (Figure is adapted from the following article: J. K., Nicholson and J. C., Lindon (2008). *Systems biology: Metabonomics*. Nature, 455, 1054-1056)

These statistical o multivariate statistical comparisons and feature identification techniques usually involve unsupervised clustering as **Principal Component Analysis** (**PCA)** (L. I. Smith, 2002)**,** or supervised classification as **Linear Discriminant Analysis** (**LDA**) (S. J. Dixon et *al*, 2007), **Partial Least Squares Discriminant**

Analysis (**PLS-DA**) (J. A. Westerhuis et *al.*, 2008) or **Partial Least Sqaures** (**PLS**) (D. M. Pirouz, 2006).

Formally, **PCA** is a way of identifying patterns in data, expressing them in such a way as to highlight their similarities and/or differences (M-A Rodríguez-Delgado et *al.*, 2002; B. M. Silva et *al.*, 2006). The advantage of this techniques is the capability to reduce multidimensional data set (a data matrix) into a new set of uncorrelated (i.e., orthogonal) variables by performing a covariance analysis (ANCOVA) between factors.

The PCA works by decomposing the X-matrix (containing the original data set) as the product of two smaller matrices, which are called **loading** and **score** matrices (Figure 1.7).

The loading matrix (**V**) contains information about the variables: it is composed of a few vectors (Principal Components, PCs) which are (obtained as) linear combinations of the original X-variables. The score matrix (**U**) contains information about the objects. Each object is described in terms of its projections onto the PCs, (instead of the original variables) (Eq. 1.1).

$$X = V^T * U \quad \text{(Eq.1.1)}$$

The information not contained in these matrices remains as "unexplained X-variance" in a residual matrix (E) which has exactly the same dimensionality as the original X-matrix.

The PCs, among many others, have two interesting properties:

1. they are extracted in decreasing order of importance. The first PC always contains more information than the second, the second more than the third and so on...

2. they are orthogonal to each other. There is absolutely no correlation between the information contained in different PCs.

In PCA, is possible to decide how many PCs should be extracted (the number of significant components, i.e. the dimensionality of the model).

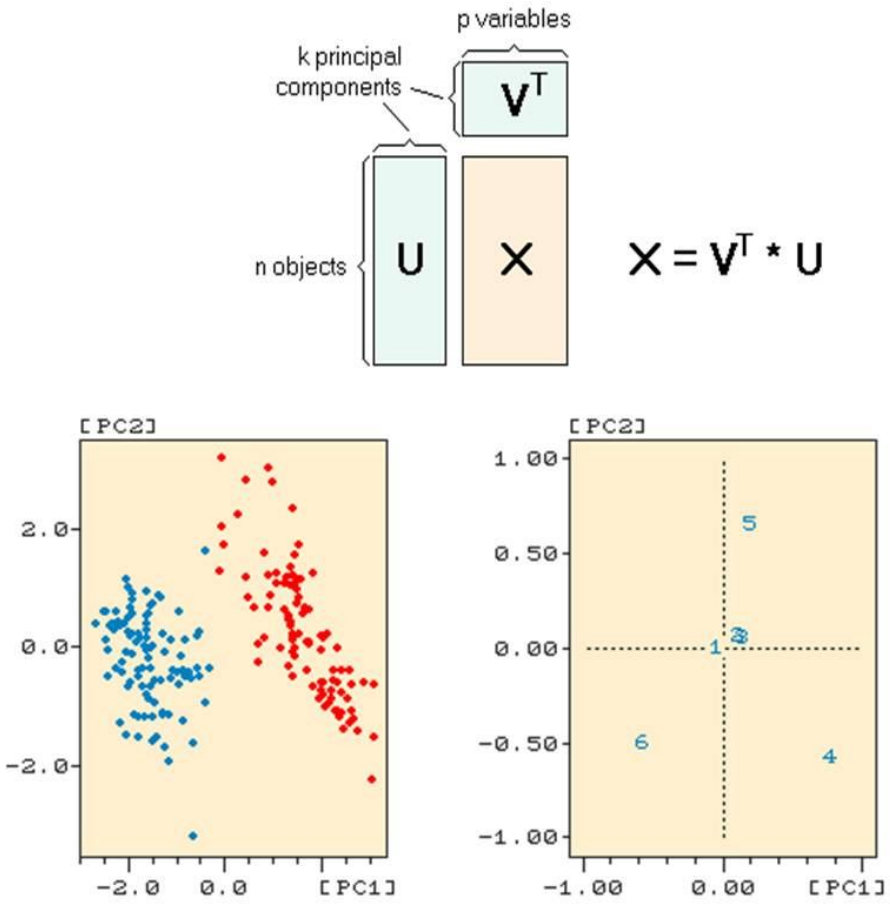


Figura 1.7: the two matrices V and U are orthogonal. The matrix V is usually called the loadings matrix and the matrix U is called the scores matrix. There are a few common plots which are always used in connection with PCA: 1) the scores/scores plot (left part of the Figure below)and 2) the corresponding loading/loading plot (right part of the figure below) (Figures is adapted from H., Lohninger (1999). *Teach/Me Data Analysis*. Springer-Verlag, Berlin-New York-Tokyo)

Each new PC extracted increases further the amount of information (variance) explained by the model. However, usually the first four of five PCs explain more than 90% of the X-variance. Anyway, there is not a simple nor unique criterion to decide how many PC to extract and two kinds of considerations should be taken into account. From a theoretical point of view, it is possible to use cross-validation techniques (H. Kubinyi et al., 1993) to decide the number of PCs to include. Since data patterns can be hard to find in data of high dimension, like spectroscopic ones, where most of the

time the information is redundant, PCA is a powerful tool for analyzing them. In ^{1}H NMR, redundancy means that some of the variables are correlated with one another, because they are measuring the same construct (different picks for the same molecule). Because of this redundancy, is possible to reduce the observed variables into a smaller number of artificial variables (**principal components** or **latent factors**) that are a linear combination of the original ones and will account for most of the variance in the observed variables, without much loss of information. In this way, by using a few components, each sample (spectrum) can be represented by relatively few numbers instead of by values for thousands of variables (spectral data points). Samples can then be plotted, making it possible to visually assess similarities and differences between samples and determine whether samples can be grouped (M. Ringnér, 2008). As a clustering technique, PCA is most commonly used to identify how one sample is different from another, which variables contribute most to this difference, and whether those variables contribute in the same way (i.e. are correlated) or independently (i.e. uncorrelated) from each other. In contrast to PCA, **LDA**, **PLS and PLS-DA** (O. E. De Noord, 1994; D. Kleinbaum et *al.*, 1988; H. Martens et *al.*, 1989) are supervised classification technique that can be used to enhance the separation between groups of observations by rotating PCA components such that a maximum separation among classes is obtained (S. Chevallier et *al.*, 2006).

The purpose of Discriminant Analysis is to classify objects (people, customers, foods, genes, things, etc.) into one of two or more groups based on a set of features that describe the objects (e.g. gender, age, income, weight, preference score, genotypes, metabolites' content etc.). In general, we assign an object to one of a number of predetermined groups based on observations made on the object. For example, we want to know whether a soap product is good or bad based on several measurements on the product such as weight, volume, people's preferential score, smell, color contrast etc. The object here is soap. The class category or the group

("good" and "bad") is what we are looking for (it is also called dependent variable). Each measurement on the product is called features that describe the object (it is also called independent variable). Thus, in discriminant analysis, the dependent variable (Y) is the group and the independent variables (X) are the object features that might describe the group. The dependent variable is always category (nominal scale) variable while the independent variables can be any measurement scale (i.e. nominal, ordinal, interval or ratio). If we can assume that the groups are linearly separable, we can use linear discriminant model (LDA). Linearly separable suggests that the groups can be separated by a linear combination of features that describe the objects (K. Fukunaga, 1990; R. Duda et *al.*, 2000; T. Hastie et *al.*, 2001). If only two features, the separators between objects group will become lines. If the features are three, the separator is a plane and the number of features (i.e. independent variables) is more than 3, the separators become a hyper-plane.

Partial Least Squares (PLS) is useful when a (very) large set of independent variables have to be predicted. It originated in the social sciences (H. Wold, 1966) but became popular also in all branches based on in chemometrics methods, including food science (A. Szydlowska-Czerniak, 2007; J. M. Poveda et *al.*, 2006). It is a multivariate regression method allowing to establish a relationship between one or more dependent variables (U) and a group of descriptors (T). T- and U-variables are modelled simultaneously to find the latent variables (LVs) in T that will predict the latent variables in U and at the same time account for the largest possible information present in T; Figure 1.8 gives a schematic outline of the method. The overall goal (shown in the lower box of Figure 1.8) is to use the factors to predict the responses in the population. This is achieved indirectly by extracting latent variables T and U from sampled factors and responses, respectively. The extracted factors T (also referred to as X-scores) are used to predict the Y-scores U,

and then the predicted Y-scores are used to construct predictions for the responses (D. T. Randall, 2004).

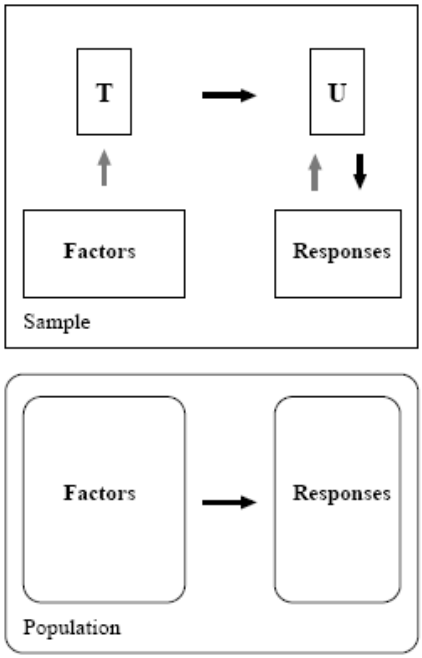


Figure 1.8: schematic outline of PLS method. Hence the PLS method is popular in industries that collect correlated data on many x-variables, known as predictors. For example, multivariate calibration in analytical chemistry; spectroscopy in chemometrics. The PLS method extracts orthogonal linear combinations of predictors, known as factors (**T** or **X-Scores**), from the predictor data that explain variance in both the predictor variables and the response (**U** or **Y-Scores**) variable(s) (Figure is adapted from D. T., Randall D. Tobias (2004). *An Introduction to Partial Least Squares Regression*. SAS Institute Inc., Cary, NC and S., Wold (1994). *PLS for Multivariate Linear Modeling QSAR: Chemometric Methods in Molecular Design. Methods and Principles* in *Medicinal Chemistry.* Van de Waterbeemd H (Editor) Verlag-Chemie)

So, in this case the latent variables are selected on the basis of explaining contemporarily both descriptors and predictors. These latent variables are similar to the principal components calculated from PCA - the first one accounts for the largest amount of information followed by the other components that account for the maximum residual variance. As for PCs, the last LVs are mostly responsible for random variations and experimental error. The optimal number of LVs, i.e. modelling information in X useful to predict the response Y but avoiding overfitting, is determined on the basis of the residual variance in prediction.

Cross-validation techniques are adopted for evaluating the predictive ability and select the optimal number of latent variables.

PLS was contrived to model continuous responses but it can be applied even for classification purposes by establishing an appropriate Y related to the belonging of each sample to a class. In this case it is called Partial Least Squares – Discriminant Analysis (PLS-DA).

In the case of proteomic data, one response variable for each group of samples is usually adopted. Each response variable is assigned a 1 value for the samples belonging to the corresponding class and a 0 value for the samples belonging to the other classes.

In general, a PLS analysis consists of the stages:

1. calculate a PLS model using a high number of factors (more than is likely to be required);

2. determine the number of factors to include in a fitted model by either:

- analysing information calculated during the process of extracting factors;
- calculating a prediction accuracy estimate based on, e.g., cross-validation;

3. fit the model with the determined number of factors by calculating parameter estimates of the linear regression;

4. given a set of predictors and responses used to fit a PLS model, and a suitable number of factors to use to calculate parameter estimates, estimate response values to new predictor data.

As it has been told at the beginning of the paragraph, chemometric approaches like PCA and PLS-DA, on their own, do not permit the direct identification or quantification of compounds. In the other approach to metabonomics (**quantitative metabolomics or targeted profiling**) the focus is on attempting to identify and/or quantify as many compounds in the sample as possible.

This usually done by comparing the spectroscopic data (obtained from sample's NMR or MS ones) spectroscopic data reference library obtained from pure compounds (N. J. Serkova et *al.*, 2007; A. M. Weljie, et *al.*, 2006; D. S. Wishart et *al.*, 2001). Once the constituent compounds are identified and quantified, the data are then statistically processed (using PCA or PLS-DA) to identify the most important biomarkers or informative metabolic pathways (A. M. Weljie et *al.*, 2006).

Depending on the objectives and instrumental capacity, quantitative metabolomics may be either targeted (selective to certain classes of compounds) or comprehensive (covering all or almost all detectable metabolites).

1.2.5 Conclusions

The application of the whole "**high throughput molecular profiling strategy"** (**HTMPs**, for convenience), that start from genomics through proteomics and metabolomics/metabonomics and ends with chemometrics (Figure 1.9), together with range of spectroscopic methods that were linked to pattern-recognition techniques, such as Nuclear Magnetic Resonance, is being applied with increasing frequency in all human sciences. In recent time, the strategy also took place both in the food industry and in the agriculture sector. Beyond them, the strategy is contributing to the improvement of food processing, food safety and quality assurance from "farm to fork" (C. Almeida et *al.*, 2006).

In the same time, the improvement of food quality and technology brought to innovate also the nutrition aspect of food. In fact, from a nutritional point of view, the HTMPs is allowing the development of functional food and personalized food products for individual and customized diet, based on genomic, proteomic and thus metabolic personal information, so to optimize health and prevent disease.

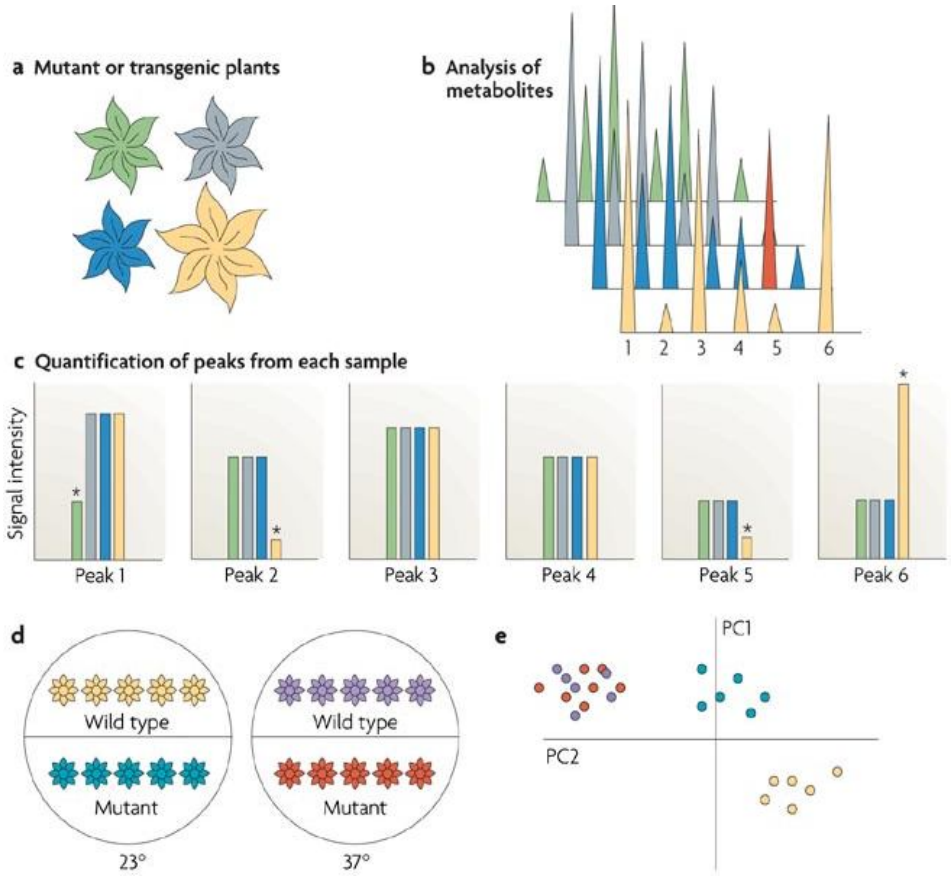


Figure 1.9: four genetically different plants are indicated by different colours in part **a**. Although there are clear differences in the sizes, these morphological distinctions cannot be used to make inferences about the status of metabolic pathways. Part **b** shows a schematic representation of the intensities of six peaks that were chromatographily separated and measured. The peaks are colour coded to correspond to the plants shown in part **a**. There is a seventh peak labelled in red in the blue plant chromatogram that represents a novel metabolite present only in this mutant. Quantitative comparison of each of the six normally detected compounds from the four plants (part **c**). The asterisk indicates statistically significant differences in amounts of specified compounds from each plant. In this example, wild-type and high-temperature tolerant mutant plants were continuously grown at 23°C or were shifted to 37°C (part **d**). Part **e** shows the results of principal components analysis of metabolites from these samples. The data reveal that the 23°C grown wild-type sample (yellow circles) is phenotypically distinct from the remaining samples. By contrast, the high-temperature tolerant mutant (turquoise and red) samples from both temperatures are similar to 37°C-treated wild-type plants (purple) (Figure is adapted from R. L., Last, A. D., Jones and Y., Shachar-Hill (2007). *Towards the plant metabolome and beyond*. Nature Reviews Molecular Cell Biology 8**,** 167-174)

1.3 NMR AS A METHOD FOR FOODS' MOLECULAR PROFILING

The application of nuclear magnetic resonance (NMR) spectroscopy to the analysis and quality control of foods has shown great development in the last few years. The increase of new application and

the attention to this technique by scientists, official control institutions and food industries can be attributed both to the high specifity and versatility of the NMR technique and to the improvement of NMR instrument performances and availability. In food chemistry, NMR becomes a specific analytical techniques to assess food quality and verify in an objective way the "food's history" (R. Sacchi et *al.*, 2007).

With the respect to the objective definition of food quality parameters (par. 1.1), all NMR techniques can give potential information (directly or indirectly) related to the molecular composition, the physical status of water and fat, the starch and protein in emulsions, the internal structure of solid foods and so on. The ability of high-resolution NMR to monitor in a non invasive way all abundant molecules present in a row materials or in a complex system is a major drive for NMR applications in food science. Thus, a lot of applications use the NMR spectra as a "**fingerprint**" of foods.

1.3.1 NMR Spectroscopy: the very basic principles

The principles of NMR spectroscopy are well known nowadays and it will be easily available in many textbook (I. P. Gerothanassis et *al.*, 2002) and for these reasons it will be not deeply discussed here but only the main parameters that are used in treating the NMR data in food analysis. NMR relevant parameters are:

1. chemical shift (δ) (R. J. Abraham et *al.*, 2005);

2. relaxation times.

Chemical shifts, generally referred to in terms of ppm, describe the dependence of nuclear magnetic energy levels on the electronic environment in a molecule. For proton the usual range falls between 0 and 12 ppm, as referred to the TMS (tetramethylsilane). Other nuclei such as ^{13}C, ^{31}P and ^{15}N have distinct advantages in terms of chemical shifts range in order of more than 100 ppm (Figure 1.10) but also a disadvantages due to their much weaker sensitivity.

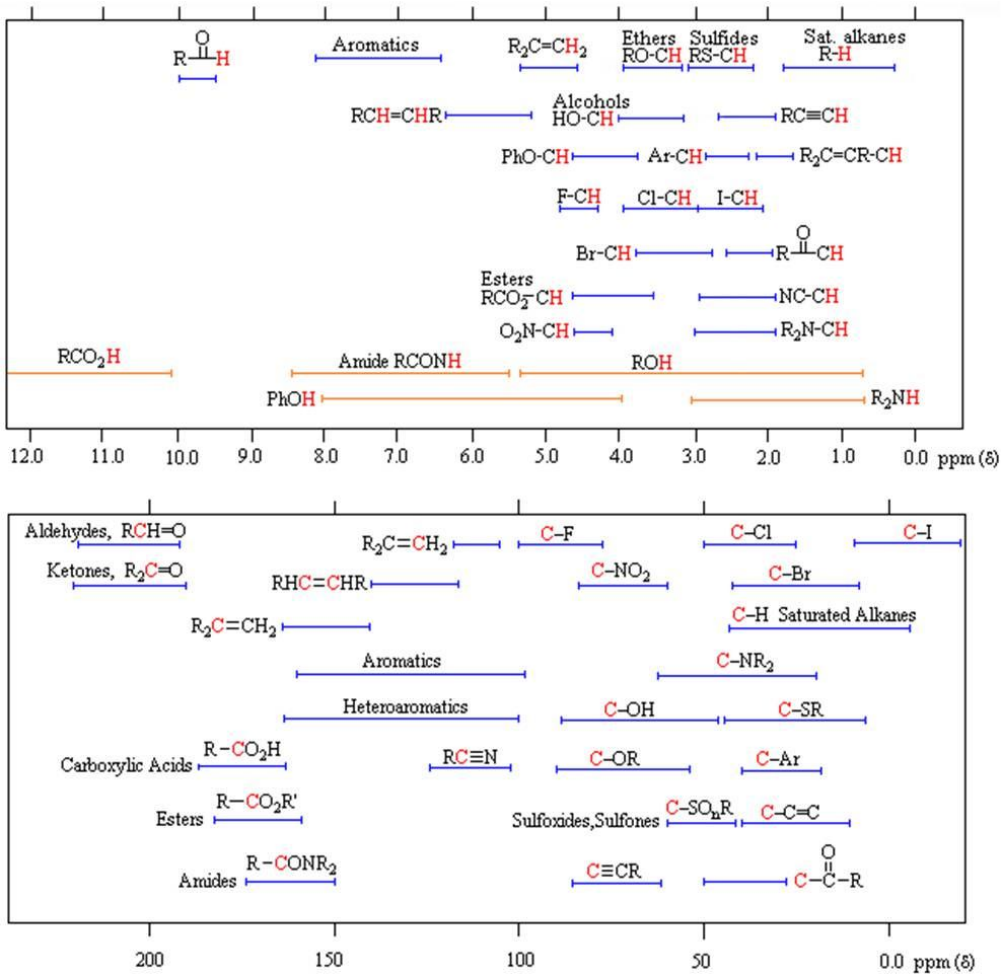


Figure 1.10: chemical shifts for ^1H (upper spectrum) and for ^{13}C (lower spectrum). The references are from samples in $CDCl_3$ solution. The δ (ppm) scale is relative to TMS at δ=0=ppm. For ^1H Spectrum: The broad ranges shown at the bottom of the chart (orange color) are typical of hydrogen bonded protons (OH and NH). These signals are concentration and temperature dependent. Note that in DMSO-d_6 solution, alcohol OH signals are shifted to lower field (δ 4.0 to 6.0ppm), and usually display vicinal coupling; for the ^{13}C spectrum: Conjugation of a double bond with a carbonyl group perturbs the carbon resonances of both groups. The β-carbon of the double bond is shifted to lower field by 20 to 30 ppm, and the carbonyl carbon is shifted to higher field by 5 to 15 ppm (Figure is adapted from www.cem.msu.edu)

In the NMR spectrum of an organic compound, peaks appear at the positions of absorption, also called the **positions of resonance**, also called the **positions of resonance or precession frequencies**, for different nuclei in the molecule. The exact chemical shift of a particular nucleus in a molecule gives information about how the atom with that nucleus is bonded in the molecule. The x-axis of the spectrum is called the delta scale (δ) with units of ppm and the y-axis is an intensity scale (Figure 1.11).

The height of the peak on the y-axis is proportional to the number of ^{1}H nuclei in the molecule with the same chemical shift. The intensity of a ^{13}C resonance is, however, not proportional to the number of carbons.

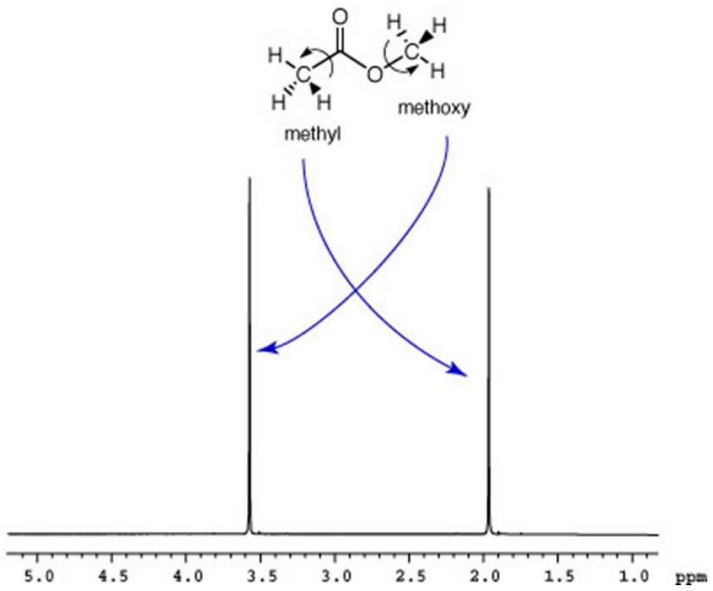


Figure 1.11: the ^{1}H NMR spectrum of methyl acetate shows two peaks: one at δ= 2.1 ppm for the three equivalent methyl protons and one at δ= 3.7 ppm for the three equivalent methoxy protons. The methoxy protons are said to resonate **downfield** from the methyl protons. Each of the three methyl protons are equivalent to each other and each of the three methoxy protons are equivalent to each because of rapid rotation about the carbon-carbon and carbon-oxygen single bonds. ^{1}H or ^{13}C nuclei are said to be **chemical shift equivalent** if they have exactly the same chemical shift (Figure is adapted from http://orgchem.colorado.edu)

The chemical shift is a relevant parameter especially in high-resolution NMR (**HR-NMR**). Among all the NMR techniques, HR-NMR has a very high sensitivity and it makes use of high-magnetic field that permit the observation of very detailed spectral parameters. These two factors are indeed quite important in assign component and measuring their intensities ratio. This leads the HR-NMR, for its inner nature, to be employed with the best results in food chemistry. Recent literature offers examples of interesting application of the method in differentiating foods obtaining in different processing conditions or technologies (A. Koller et *al.*, 2008), to assess adulterations of low-grade foods in mixtures (R. Sacchi et *al.*, 1997) and to characterize typical food products (M. D' Imperio et *al.*, 2007).

From a **qualitative** point of view, in term of linearity and selectivity, NMR is definitely the best analytical method from the linearity point of view since the intensity of resonance is strictly proportional to the number of nuclei resonating at a certain frequency. Selectivity is also good because NMR differentiates all the isotopes of elements and even for a given isotope is able to yield measurable differences in chemical shift for different chemical environmental. From a **quantitative** point of view, sensitivity depends on the signal-to-noise-ratio, which can be considered acceptable when it is higher than 10. Precision and accuracy can be determinate from means standard deviation on replicates.

Relaxation describe the physical pathways by which perturbed nuclei return to their original equilibrium state. In a three dimensional space, two types of relaxation time can exist: the **spin-lattice** or **longitudinal** that brings the nuclei aligned along the z axis (where the magnetic field is applied) and the **spin-spin** or **trasverse** that is concerned with the physical phenomena occurring within the xy plane (Figure 1.12), both characterized by a specific relaxation time ($\mathbf{T_1}$ and $\mathbf{T_2}$).

Assessment of proton relaxation behavior is a frequently used application of the Low Field NMR (**LF-NMR**) technique, so both T_1 and T_2 relaxation time are use as parameter for this method.

Unlike conventional NMR spectrometers equipped with strong superconductive magnets, LF-NMR instruments have a relatively weak permanent magnet. Consequently, the instrument cannot resolve different spectral components in the frequency domain, thus the LF-NMR technique can only deal with time-domain information. LF-NMR is a technique largely employed in food chemistry for the classical evaluation of fat content (X. Sun et *al.*, 1995), protein aggregation (L. Indrawati et *al.*, 2007), viscosity (M. H. Rahman et *al.*, 2002) and water (H. Lechert et *al.*, 1980; F. Mariette, 2008), in different foods. In the last case, water being a more mo-

bile molecules has a longer relaxation time and it easily monitored against less mobile larger molecules such as fats or proteins.

The application of the technique in this direction is rather important in food processing and aging (R. Ginaferri et *al.*, 2007), where it gives fast responses on the water/compounds relative ratio.

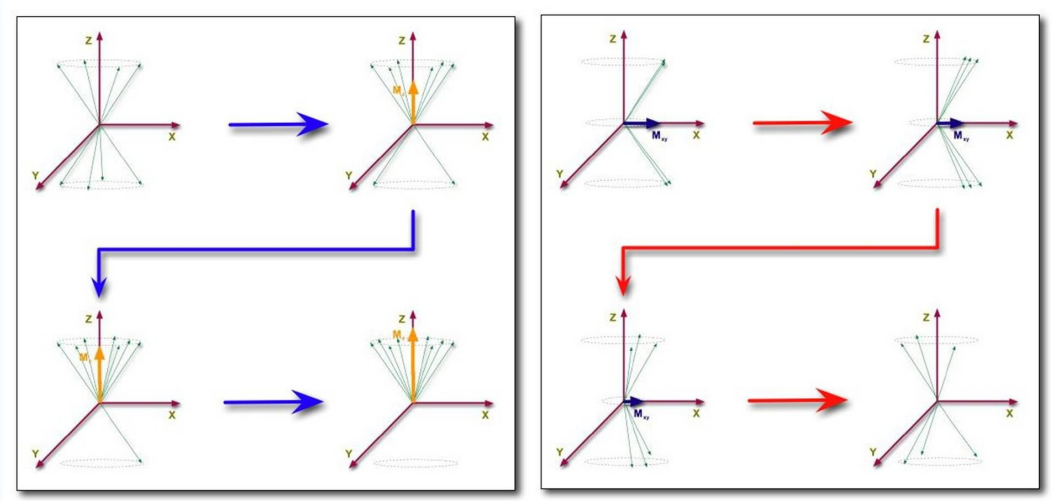


Figure 1.12: NMR Relaxation. On the left side a hypothetical situation is shown where four protons are aligned parallel to the external magnetic field, and four excited protons are aligned anti-parallel. The magnetic sum vector is zero at this stage. Excited protons then return progressively to their lower energy state. The result is a re-establishment of the longitudinal magnetization (**longitudinal** or spin-lattice relaxation) in an exponential manner characterized by a time constant, called $\mathbf{T_1}$. On the right side a second hypothetical situation is illustrated where three protons are aligned parallel and three anti-parallel, in phase coherence with each other. The other panels in the Figure illustrate a progressive loss of coherence that is characterized by this form of relaxation, where the transverse magnetization decreases to zero (**trasverse** relaxation). The transverse relaxation decreases in an exponential manner characterized by a time constant, called $\mathbf{T_2^*}$ - also called T_2 *Star* (Figure is taken, with concession from the author, from http://en.wikibooks.org)

Since the measure of T_2 is more rapid than the longitudinal ones (typical data acquisition times for pure aqueous solutions are about 3 and 40 min, respectively) and because transversal relaxation data may contain more information as well, sometimes only transversal relaxation data are reported.

1.3.2 Methods for metabolite approach by HR-NMR in food chemistry

In Table 1.4 are listed the main possible metabolic approaches based on the HR-NMR spectroscopy in food science.

Table 1.4: classification of metabolomic approaches (table is adapted from R., Goodacre, S., Vaidyanathan, W. B., Dunn, G. G., Harrigan and D. B., Kell (2004). *Metabolomics by numbers: acquiring and understanding global metabolite data*. TRENDS in Biotechnology, 22, 245–252)

Approach	Function
Metabolite profiling	Analysis focused on a group of metabolites, for example, a class of compounds such as carbohydrates, amino acids or those associated with a specific pathway (F. D. Gunstone, 1994)
Metabolomics	Comprehensive analysis of the whole metabolome under a given set of conditions (H. Vidarto et *al*., 2006)
Metabolic fingerprinting	Classification of samples on the basis of provenance of either their biological relevance or origin (C. Hyung-Kyoon et *al*., 2004)
Metabolic profiling	Often used interchangeably with 'metabolite profiling
Metabonomics	Measure of the fingerprint of biochemical perturbations caused by natural or technologist changes (C. Xiao et *al*., 2008)

All these approaches focus on the metabolic content of samples, moving the attention from the level of gene transcription (trascriptome) and protein modification (proteome) to the determination and evaluation (in term of concentration) of metabolites (molecules). But, as it has been told in par 1.2.3, while **metabolomics** leads to obtain a "complete set of metabolites/low-molecular-weight intermediates, which are context dependent, varying according to the physiology, developmental or pathological state of the cell, tissue, organ or organism" (S. G. Oliver, 2002), **metabonomics** measures the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification. Both metabolomics and metabonomics approaches can coexist in same researcher and one is strictly connected to the other in a

double sense: the metabolomic exploration (**metabolomics profile**) can help to understand the response of a biological system to extern stimuli (**metabonomics profile**) and *vice versa*.

In par 1.2.3 it has been also illustrated the range of sensitivity of spectroscopic techniques used in metabonomics researches; however, NMR-based metabonomics has proven to be particularly apposite for the rapid analysis of complex biological samples like food stuff. The NMR spectral results so generated yield a unique metabolic fingerprint for each complex biological mixture: if its status, such as in food processing, the unique metabolic fingerprint or signature reflects this change. This is the main characteristic that make the NMR one of the optimal technique for food analysis from a simple metabolites screening till the characterization of foods. The foods' molecular profile method, based o NMR technique, is schematically represented by the Figure 1.13, in which a red square points out the possible results that can be obtained from the method. Referred always to the food analysis, these results are:

1. **classification of food (metabonomics study)**: based on I. B. Abdel-Farid et *al.*, concept for whom metabonomic study can provide significant results only if the metabolic changes in a target group is significantly different from the biological variation of the relative control group. Examples in this way are the classification of traditional foods (L. Viggiani et *al.*, 2008; E. Schievano et *al.*, 2008) and discrimination between wild type and genetically modified foods (M. Defernez et *al.*, 2004) based on differences;

2. **metabolites characterization (metabolomics study)**: aim to describe foods from a metabolites' content point of view (G. Le Gall et *al.*, 2003; Z. Fang et *al.*, 2006; C. W. Wilson et *al.*, 1978);

3. **metabolites' identification (metabon(l)omics study)**: in order to identify those metabolites (biomarkers) able to explain

the reasons of differences among groups. This leads understand better possible effects on foods due to technologist and biotechnologist processes, to assess food quality (S. Yoo-Soo et *al.*, 2007) and to fight foods frauds and adulterations (K. G. Penman et *al.*, 2006).

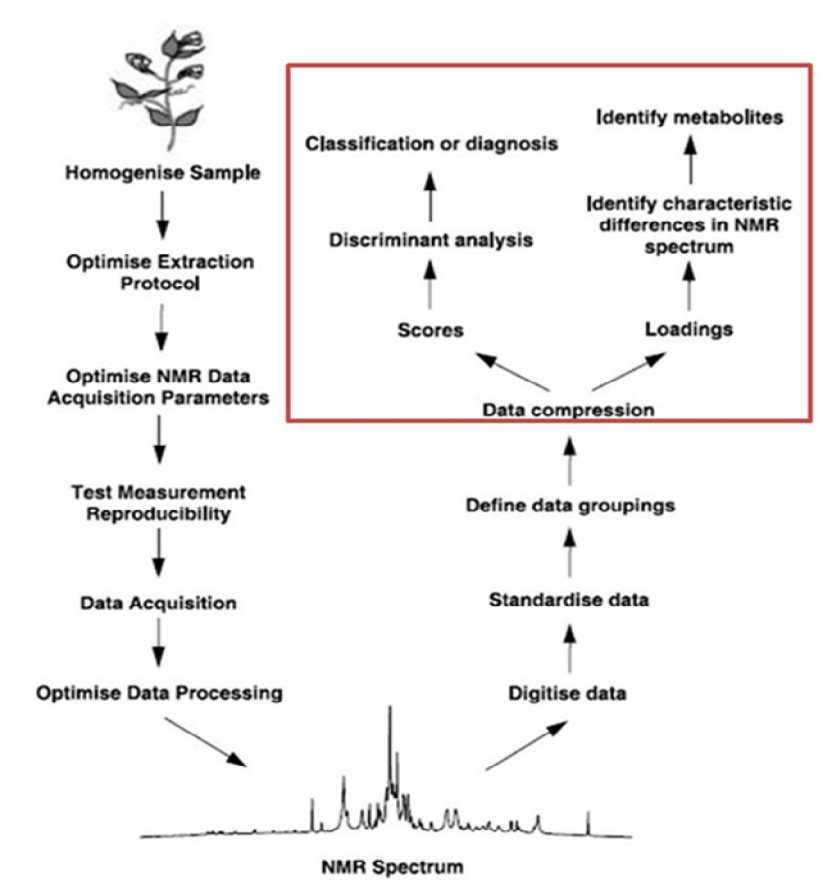


Figure 1.13: methods for metabolite profiling by NMR. In red the **molecular profiling** pathways (Figure is taken, with concession from the author, from A., Charlton, T., Allnutt, S., Holmes, J., Chisholm, S., Bean, N., Ellis, P., Mullineaux and S., Oehlschlager (2004). *NMR profiling of transgenic peas.* Plant Biotechnology Journal, 2, 27–35)

1.3.3 NMR and chemometrics

Once again, is import to point on evidence that metabon(l)omics combines the techniques of high resolution **NMR** with pattern recognition technology to rapidly evaluate the metabolic status of the biological matrix. When using spectroscopic techniques for chemical quantification, have been used univariate data analytical methods (par 1.2.4). this application is called **chemometrics**.

Actual definitions of chemometrics are:

"*the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analyzing chemical data*" (B. Kowalski, 1997) (Figure 1.14).

Figure 1.14: chemometrics definition (Figure is adapted from http://fullindex.net/)

"*The discipline that uses mathematical and statistical methods to obtain relevant information on material systems"* (I. Frank et *al.*, 1982) (Figure 1.15).

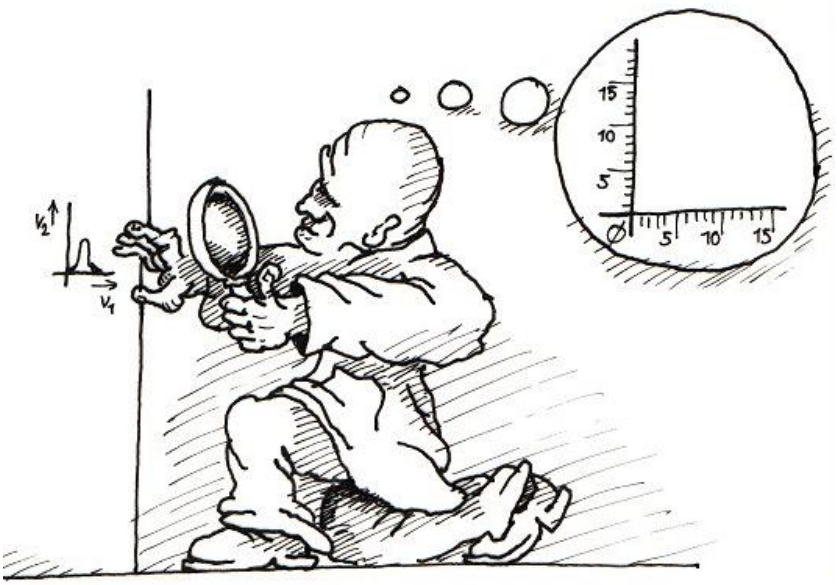


Figure 1.15: chemometrics definition (Figure is adapted from http://fullindex.net/)

The discipline of chemometrics originates in chemistry; thus, typical applications of chemometric methods are the development of quantitative structure activity relationships or the evaluation of analytical–chemical data. The data flood generated by modern analytical instrumentation (like spectroscopic technologies) is one reason that analytical chemists in particular develop applications of chemometric methods. While most other types of spectroscopic data can be subjected to chemometric analysis directly from the spectrometer, NMR data often need to be preprocessed in several ways in order to conform to the prerequisites for chemometric data analysis:

1. **Fourier transformation** (J. Keeler, 2002) (Figure 1.16).

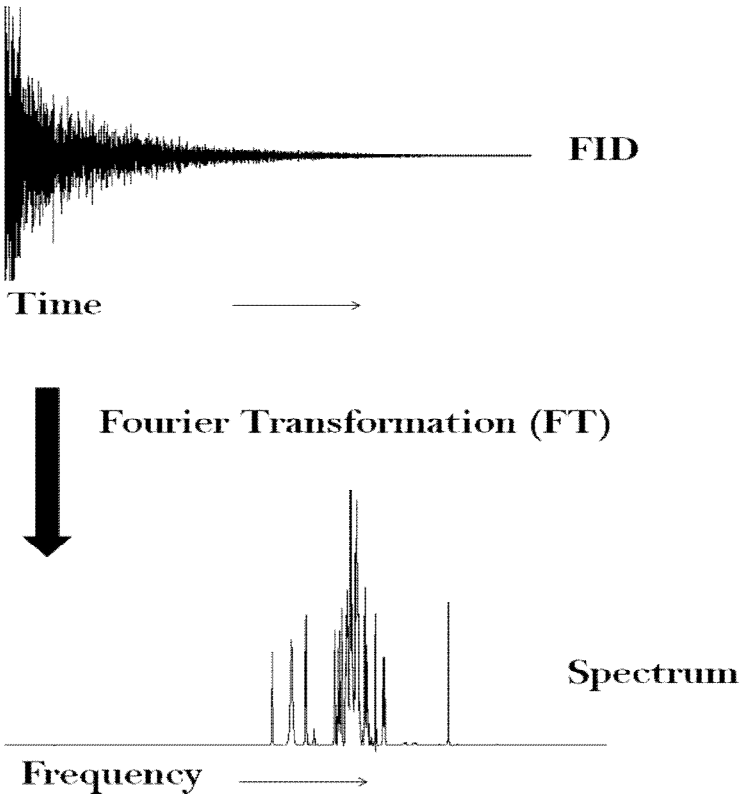


Figure 1.16: Fourier transformation of a FID obtained from a grape fruit (*Vitis vinifera*) sample's extract. The FT process takes the time domain function (the FID) and converts it into a frequency domain function (the spectrum). The ^{1}H-NMR spectra was recorded at T= 300K on a Varian Mercury-plus spectrometer, operating at ^{1}H frequency of 400 MHz; 2048 scans were acquired, with data collected into 16K data points with a spectral width of 16 ppm, a pulse angle of 60°, a recycle delay of 1.0 s, and acquisition time of 2.561 s with a constant receiver gain. Free induction decay (FID) was Fourier transformed, with the MestReC Software (http://www.mestrec.com/), by performing an exponential multiplication with a 1 Hz line broadening

In NMR spectroscopy a Fourier transformation (**FT**) is required to convert the time domain data (free induction decay or **FID**, an electrical signal oscillating at the NMR frequency), obtained from the spectrometer, to the frequency domain (NMR spectrum). Naturally, quantitative methods require that parameter settings for the Fourier transform (choice of zero-filling and apodization function) are equal for all samples to be evaluated, as they may influence the finer details in the spectra.

2. Phase errors (J. Keeler, 2002) (Figure 1.17)**.**

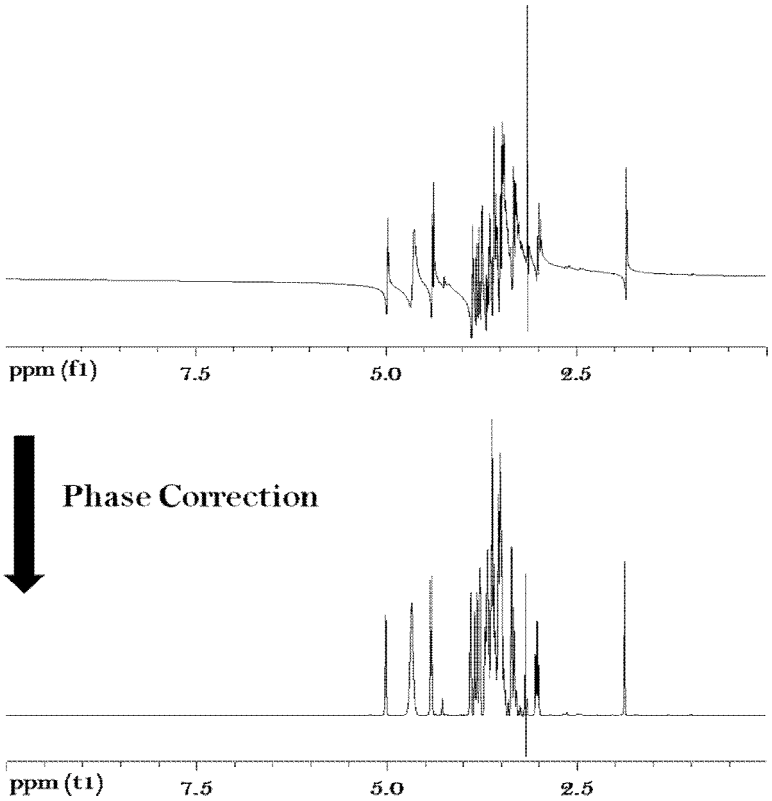
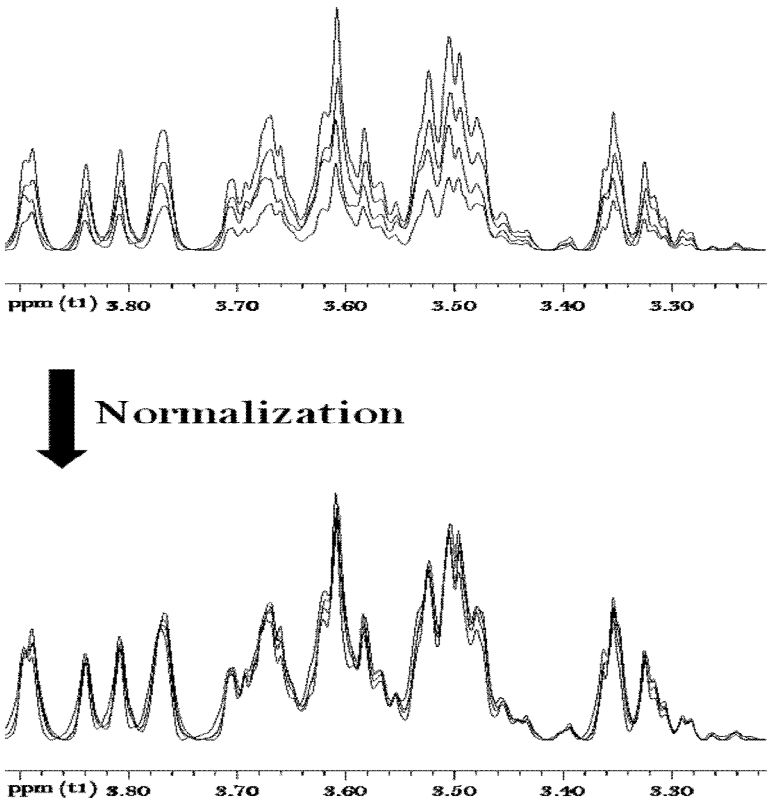


Figure 1.17: phase Correction of a spectrum obtained from a grape fruit (*Vitis vinifera*) sample's extract. The ^{1}H-NMR spectra was recorded at T= 300K on a Varian Mercury-plus spectrometer, operating at ^{1}H frequency of 400 MHz; 2048 scans were acquired, with data collected into 16K data points with a spectral width of 16 ppm, a pulse angle of 60°, a recycle delay of 1.0 s, and acquisition time of 2.561 s with a constant receiver gain. The phase error correction was made with the MestReC Software by performing a manual correction on both zero and first errors orders (http://www.mestrec.com/)

A difficult problem encountered with NMR data is the existence of phase errors of two orders: one and zero. In real experiment, after FT the spectrum line shapes are mixture of absorpitive and dispersive signals and are related to the delayed FID acquisition and is

commonly called first order phase. The delayed acquisition is a consequence of the minimum time required to change the spectrometer from transmit to receive mode, during this delay the magnetization vectors process according to their chemical shift frequencies. The zero order phase error arises because of the phase differences between the magnetization vectors and the receiver. Manual phase correction is usually implemented in the instrument software, but this process is very time consuming, especially for the large data sets that are often analyzed using chemometrics. More importantly, manually phase-correcting a series of spectra will lead to suboptimal results due to the subjective evaluation of the correction necessary for individual spectra.

3. Data normalization (Figure 1.18).



4.

Figure 1.18: normalization of a set of spectra obtained from a grape fruit (*Vitis vinifera*) samples' extract. In this Figure is shown part of the midfield ^{1}H NMR region (from 5.2 to 2.8 ppm). The ^{1}H-NMR spectra were recorded at T= 300K on a Varian Mercury-plus spectrometer, operating at ^{1}H frequency of 400 MHz; for each spectrum 2048 scans were acquired, with data collected into 16K data points with a spectral width of 16 ppm, a pulse angle of 60°, a recycle delay of 1.0 s, and acquisition time of 2.561 s with a constant receiver gain. The normalization was made with the MestReC Software (http://www.mestrec.com/)

Data normalization is an important step for any statistical analysis. The objective of data normalization is to allow meaningful comparisons of samples within the dataset.

It is a row operation that is applied to the data from each sample and comprises methods to make the data from all samples directly comparable with each other (A. Craig et *al*., 2006). In this way it is possible to minimize most of the differences introduced with the effect of variable dilution and spectral data acquisition and processing.

Normalization can be done using an internal “housekeeping” metabolite, for example an inner standard like TMS or, in our case, normalize each spectrum to (divide each variable by) the sum of the absolute value of all variables for the given sample. It returns a vector with unit area (area = 1) "under the curve" (B. M. Wise et *al.*, 2006).

5. Chemical shift variations: the last preprocessing problem to be mentioned here and which occurs only in HR-NMR spectroscopy is the chemical shift variations that may occur from sample to sample or even from peak to peak. The overall sample-to-sample variations are due to small variations in spectrometer frequency, whilst the peak to peak chemical shift variations are due to variations in, for example, pH. In this last case, a data reduction in the form called binning (A. Craig et *al.*, 2006) is a pragmatic solution to the problem.

6. Spectral binning (Figure 1.19) is a widely-used technique where the spectrum is subdivided into a number of regions, and the total area within each bin is used as an abstracted representation of the original spectrum.

A typical 64k NMR spectrum would be reduced using bin widths of 0.04 ppm, resulting in ~250 bin integral values.

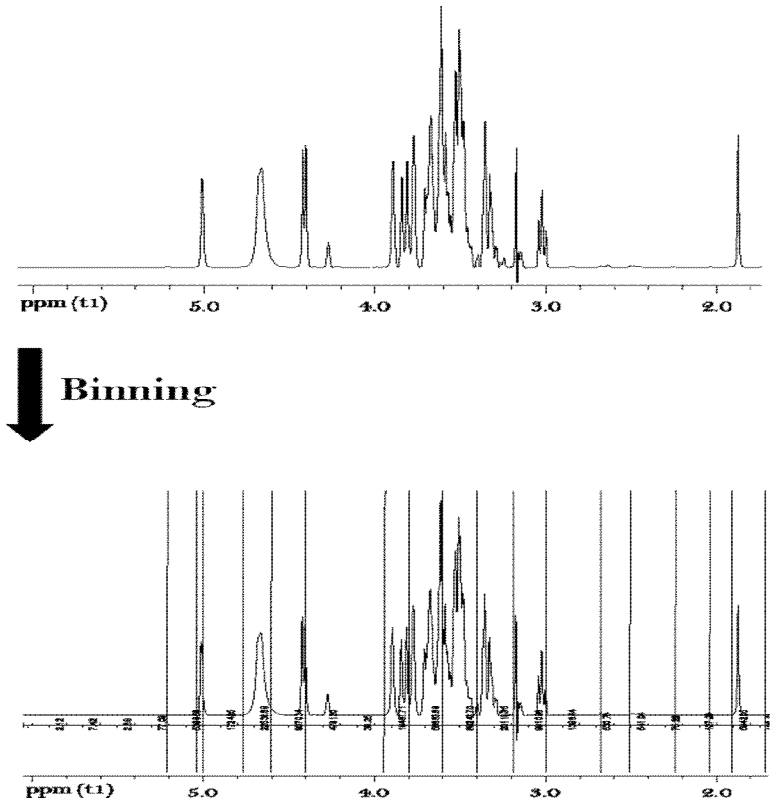


Figure 1.19: binning on a spectrum obtained from a grape fruit (*Vitis vinifera*) sample' s extract. In this Figure is shown the midfield ^{1}H NMR region (from 5.2 to 2.8 ppm) and part of the upfield ^{1}H NMR region (from 2.8 to 0.5 ppm). The ^{1}H-NMR spectra was recorded at T= 300K on a Varian Mercury-plus spectrometer, operating at ^{1}H frequency of 400 MHz; 2048 scans were acquired, with data collected into 16K data points with a spectral width of 16 ppm, a pulse angle of 60°, a recycle delay of 1.0 s, and acquisition time of 2.561 s with a constant receiver gain. The binning was made with the MestReC Software (http://www.mestrec.com/)

In order to better understand the consequences of pre-processing of NMR data, in Figures 1.20 and 1.21 are illustrated, as an example, the improvement of a PC analysis after the pre-processed methods described above on grape samples data set used in the present researcher work.

As it can be seen, in this case, normalization and binning approach (Figures 1.21 C and D) improve the results, allowing to separate two grape cultivar and among cultivar is possible to see a tendency to separate genotypes

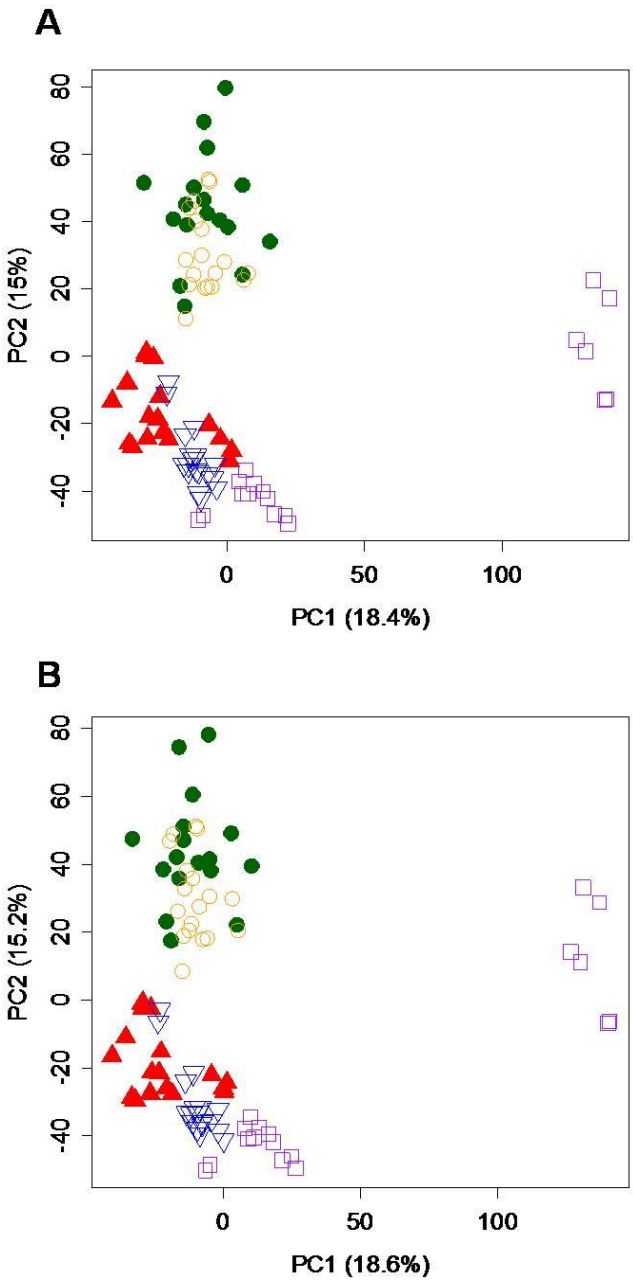


Figure 1.20: improvement of PC Analysis according to processed Spectral NMR Data (Part One) due to alignment . PCA on raw (**A**) and aligned (**B**) NMR spectral data set, including all the Thompson and Silcora samples. Samples are labelled as follows: ● Thompson Wild Type (T-WT), ○ Thompson Genetically Modified (T-GM1), ▲ Silcora Wild Type (S-WT), ▽ Silcora Genetically Modified 1 (S-GM1) and □ Silcora Genetically Modified 2 (S-GM2). According to the total variance (Tv), PC1 and PC 2 explain 18.6% and 15.2% respectively in the raw data PC plot.
The same values are even for the aligned data. As it can be seen, except for PC2 along which is possible to point out a separation among cultivars, genotypes are not differentiated (the Figures were prepared by using R program; see appendix A, chapter 4)

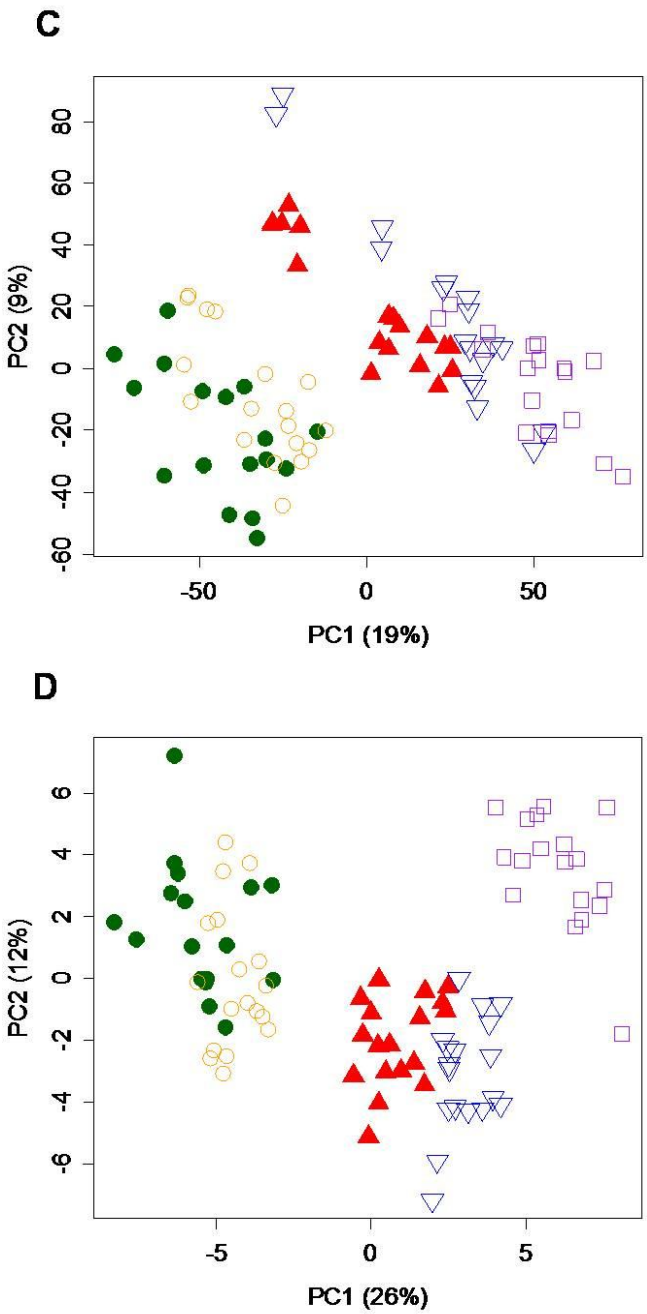


Figure 1.21 improvement of PC Analysis according to processed Spectral NMR Data (Part One) due to normalized (**C**) and binned (**D**) NMR spectral data set, including all the Thompson and Silcora samples. Samples are labelled as follows: ● Thompson Wild Type (T-WT), ○ Thompson Genetically Modified (T-GM1), ▲ Silcora Wild Type (S-WT), ▽ Silcora Genetically Modified 1 (S-GM1) and □ Silcora Genetically Modified 2 (S-GM2). According to the total variance (Tv), PC1 and PC 2 explain 19% and 9% respectively in the normalized data PC plot, whilst in binned data the values of both PCs are 26% and 12%. As it can be seen, both cultivars and genotypes, except for T-WT and T-GM1, are differentiated (the Figures were prepared by using R program; see appendix A, chapter 4)

REFERENCES

A. a. v. v (2008). Webster's New World™ Medical Dictionary, 3rd Edition

A. E., Bender (1987). *The nutritional aspects of food processing* in *Food technology international Europe* Turner Ed, London, England: Sterling Publications International, 273–275

A. J., Blood, R. J., Zatorre, P., Bermudez and A. C., Evans (1999). *Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions*. Nature Neuroscience, 2, 382–387

A. K., Jain, R. P. W., Duin and J., Mao (2000). *Statistical Pattern Recognition: A Review*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 4–37

A. M., Giusti, E., Bignetti and C., Cannella (2008). *Exploring New Frontiers in Total Food Quality Definition and Assessment: From Chemical to Neurochemical Properties*. Food Bioprocess Technology, 1, 130–142

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

A. Szydlowska-Czerniak (2007). *MIR spectroscopy and partial least-squares regression for determination of phospholipids in rapeseed oils at various stages of technological process*. Food Chemistry, 105, 1179–1187

A., Alimelli, G., Pennazza, M., Sintonico, R., Paolesse, D., Filippini, A., D'Amico et al. (2007). *Fish freshness detection by a computer screen photoassisted based gas sensor array.* Analytical Chimica Acta, 582, 320–328

A., Becalski, B. P.-Y., Lau, D., Lewis and S. W., Seaman (2003). *Acrylamide in foods: Occurrence, sources and modelling*. Journal of Agricultural and Food Chemistry, 51, 802–808

A., Charlton, T., Allnutt, S., Holmes, J., Chisholm, S., Bean, N., Ellis, P., Mullineaux and S., Oehlschlager (2004). *NMR profiling of transgenic peas*. Plant Biotechnology Journal, 2, 27–35

A., Craig, O., Cloarec, E., Holmes, J. K., Nicholson and J. C., Lindon (2006). *Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets*. Anal. Chem.,m78, 2262-2267

A., Koller, M. P., Washburn, B. M., Lange, N. L., Andon, C., Deciu, A., Spielmeyer, A., Wagner and G., Jahreis (2008). *Influence of thermal treatment of rapeseed on the canolol content.* Food Chemistry, 112, 944–948

B. K., Shoichet (2004). *Virtual screening of chemical libraries*. Nature, 432, 862–865

B. M., Silva, P. B., Andrade, R. C., Martins, R. M., Seabra and M. A., Ferreira (2006). *Principal component analysis as tool of characterization of quince* (*Cydonia oblonga Miller*) jam. Food Chemistry, 94, 504-512

B. M., Wise, N. B., Gallagher, R., Bro, J. M., Shaver, W., Windig and R., Scott Koch (2006). *PLS_Toolbox* Version 4.0 for use with MATLAB™. Eigenvector Research, Inc

B., Cavaliere, B., Macchione, G., Sindona and A. Tagarelli (2008). *Tandem mass spectrometry in food safety assessment: The determination of phthalates in olive oil.* Journal of Chromatography A, 1205, 137–143

B., de Roos and H. J., McArdle (2008). *Proteomics as a tool for the modelling of biological processes and biomarker development in nutrition research*. British Journal of Nutrition, 99, S66–S71

B., Kowalski (1997), in a formal CPAC presentation, December 1997

B-H., Cho and N. H., Hooker (2008). *Comparing food safety standards.* Food Control, 20, 40–47

C. J., Seal and K., Brandt (2007). *Nutritional quality of foods* in *Handbook of organic food safety and quality*, J., Cooper, U., Niggli and C., Leifert Ed., CRC Press Boca Raton, Boston, New York, Washington DC, chap. 3

C. J., Yu, Y. F., Lin, B. L., Chiang and L. P., Chow (2003). *Proteomics and immunological analysis of a novel shrimp allergen*. Journal of Immunology, 170, 445–453

C. L., Gavaghan, I. D., Wilson and J. K., Nicholson (2002). *Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA.* FEBS Letters, 530, 191–196

C. W., Wilson and P. E., Shaw. (1978) *Quantitative composition of cold-pressed grapefruit oil*. J. Agric. Food Chem., 26, 1432–1434

C., Almeida, I. F., Duarte, A., Barros, J., Rodrigues, M., Spraul and A. M., Gil (2006). *Composition of Beer by ^{1}H NMR Spectroscopy: Effects of Brewing Site and Date of Production*. J. Agric. Food Chem., 54, 700-706

C., Castro and C. A. Manetti (2007). *Multiway approach to analyze metabonomic data: a study of maize seeds development*. Analytical Biochemistry, 371, 194–200

C., Xiao, H., Dai, H., Liu, Y., Wang and H., Tang (2008). *Revealing the Metabonomic Variation of Rosemary Extracts Using ^{1}H NMR Spectroscopy and Multivariate Data Analysis*. J. Agric. Food Chem, 56, 10142–10153

CBD (2000). *Cartagena Protocol on Biosafety to the Convention on Biological Diversity*. United Nations Environment Programme. http://www.biodiv.org/biosafety/

D. J., Skylas, L., Copeland, W. G., Rathmell and C. W., Wrigley (2001). *The wheat-grain proteome as a basis for more efficient cultivar identification.* Proteomics, 1, 1542–1546

D. L., Massart, B. G. M., Vandeginste, L. M. C., Buydens, S., De Jong, P. J., Lewi and J., Smeyers-Verbeke (1997). *Handbook of Chemometrics and Qualimetrics.* Elsevier, Amsterdam, Part A, 207

D. M., Gottlieb, J., Schultz, M., Petersen, L., Nesic, S., Jacobsen and I., Sondergaard (2002). *Determination of wheat quality by mass spectrometry and multivariate data analysis*. Rapid Communication in Mass Spectrometry, 16, 2034–2039

D. M., Pirouz (2006). *An Overview of Partial Least Squares*. The Paul Merage School of Business University of California, Irvine

D. S., Wishart (2008). *Metabolomics: applications to food science and nutrition research*. Trends in Food Science & Technology, 19, 482–493

D. S., Wishart, L. M. M., Querengesser, B. A., Lefebvre, N. A.,Epstein, R., Greiner and J. B. Newton (2001). *Magnetic resonance diagnostics: a new technology for high-throughput clinical diagnostics.* Clinical Chemistry, 47, 1918–1921

D. T., Randall D. Tobias (2004). *An Introduction to Partial Least Squares Regression*. SAS Institute Inc., Cary, NC

D., Barber, R., Rodríguez and G., Salcedo (2008). *Molecular profiles: A new tool to substantiate serum banks for evaluation of potential allergenicity of GMO*. Food and Chemical Toxicology 46, 35–40

D., Kleinbaum, L., Kupper and K., Muller (1988). *Applied Regression Analysis and Other Multivariate Methods*, 2nd ed., Pws-Kent, Boston

D., Roger, C. S., Wilson, F., Sabban, N., Ishige and L. N., Magner (2000). *The History and Culture of Food and Drink in Asia* in *The Cambridge World History of Foods*. K. F., Kiple and K. C. Orneals Eds., Cambridge University Press, Vol II, part V.B., 1140–1191

E. A., Decker, S. A., Livisay and S., Zhou (2000). *Mechanism of endogenous skeletal muscle antioxidants: chemical and physical aspects*, in *Antioxidants in Muscle Foods. Nutritional Strategies to Improve Quality*, E. A., Decker, C. Faustman and C.J. Lopez-Bote Eds., Wiley-Interscience, New York, chap. 25

E. M., Mojduszka, J. A., Caswell, D. B., West and J. M., Harris (1999). *Changes in Nutritional Quality of Food Product Offerings and Purchases: A Case Study in the Mid-1990.s*. Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture. Technical Bulletin n° 1880

E. S., Zdzisław and B., Piotrowska (2007). *Food Components and Quality*, in *Chemical and Functional Properties of Food Components, Third Edition*, Z. E., Sikorski Eds., CRC Press Taylor & Francis Group, Boca Raton (FL), London, New York, chap. 1

E., Schievano, G., Pasini, G., Cozzi and S., Mammi (2008). *Identification of the Production Chain of Asiago d'Allevo Cheese by Nuclear Magnetic Resonance Spectroscopy and Principal Component Analysis*. J. Agric. Food Chem., 56, 7208–7214

F. D., Gunstone (1994). High resolution ^{13}C-NMR - a technique for the study of lipid structure and composition. Prog. Lipid Res, 33, 19–28

F. M., Amaral and M. S. B., Caro (2005). *Investigation of different pre-concentration methods for NMR analyses of Brazilian white wine.* Food Chemistry, 93 507–510

F. Mariette (2008). *NMR Relaxation of Dairy Products*. Modern Magnetic Resonance, 7, 1697-1701

FAO (2004). *Biotechnology applications in food processing: Can developing countries benefit?* Electronic forum on biotechnology in food and agriculture: conference 11

G. S., Becker (2008). *Food and Agricultural Imports from China*. CRS Report for Congress

G., Klebe (2007). *Virtual Ligand Screening: A Method to Discover New Drug Leads*. Supramolecular Structure and Function, 251–272

G., Le Gall, M. S., DuPont, F. A., Mellon, A. L., Davis, G. J., Collins, M. E., Verhoeyen and I. J., Colquhoun (2003). *Characterization and Content of Flavonoid Glycosides in Genetically Modified Tomato (Lycopersicon esculentum) Fruits*. J. Agric. Food Chem., 51, 2438–2446

H. A., Kuiper, E. J., Kok and K-H., Engel (2003). *Exploitation of molecular profiling techniques for GM food safety assessment*. Current Opinion in Biotechnology, 14, 238–243

H. C., Keun (2006). *Metabonomic modeling of drug toxicity*. Pharmacology & Therapeutics, 109, 92–106

H., Campos-de Quiroz (2002). *Plant genomics: an overview*. Biol Res, 35, 385–399

H., Kubinyi and U., Abraham (1993). *Practical Problems in PLS Analyses*. In *3D QSAR in Drug Design: Theory, Methods and Applications.* H. Kubinyi ed.; ESCOM, Leiden, 722

H., Lechert, W., Maiwald, R,. Kothe, and W-D., Basler (1980). *NMR-Study of water in some starches and vegetables*. Journal of Food Processing and Preservation, 3, 275–299

H., Lohninger (1999) *Teach/Me Data Analysis*. Springer-Verlag, Berlin-New York-Tokyo

H., Martens and T. Naes (1989). *Multivariate Calibration*, Wiley, London

H., Tang (2006). *Metabonomics: a holistic way of understanding nutritional biochemistry*. The 8th International Conference on the Applications of Magnetic Resonance in Food Science, Nottingham

H., Widarto, E., Van Der Meijden, A., Lefeber, C., Erkelens, H., Kim, Y., Choi, R., Verpoorte (2006). *Metabolomic Differentiation of Brassica rapa Following Herbivory by Different Insect Instars using Two-Dimensional Nuclear Magnetic Resonance Spectroscopy*. Chem Ecol. 32, 2417–2428

H., Wold (1966). *Estimation of principal components and related models by iterative least squares*. In *Multivariate Analysis*. P.R. Krishnaiaah (Ed.). New York: Academic Press, 391–420

H-K., Choi, M., Verberne, A. W. M., Leferber, C., Erkelens and R. Verpoorte (2004). Metabolic fingerprinting of wild type and transgenic tobacco plants by ^{1}H NMR and multivariate analysis technique. Phytochemistry, 65, 857–864

H-K., Choi, Y. H., Choi, M., Verberne, A. W. M., Lefeber, C., Erkelens, R., Verpoorte (2004). *Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique*. Phytochemistry, 65, 857-64

I. B., Abdel-Farid, H., Kyong Kim, Y., Hae Choi and R., Verpoorte (2007). *Metabolic Characterization of Brassica rapa Leaves by NMR Spectroscopy.* J. Agric. Food Chem, 55, 7936–7943

I. D., Wilson (2007). *"OMICS" - Overview on Metabonomics and Separation Techniques*. Encyclopedia of Separation Science, 1–8

I. D., Wilson, R., Plumb, J., Granger, H., Major, R., Williams, and E. M., Lenz (2005). *HPLC-MS-based methods for the study of metabonomics*. Journal of Chromatography B, 817, 67–76

I. F., Duarte, A., Barros, C., Almeida, M., Spraul and A. M., Gil (2004). *Multivariate Analysis of NMR and FTIR Data as a Potential*

Tool for the Quality Control of Beer. J. Agric. Food Chem., 52, 1031-1038

I. F., Duarte, A., Barros, P. S., Belton, R., Righelato, M., Spraul, E., Humpfer and A. M., Gil (2002). *High-Resolution Nuclear Magnetic Resonance Spectroscopy and Multivariate Analysis for the Characterization of Beer*. J. Agric. Food Chem., 50, 2475-2481

I. P., Gerothanassis, A., Troganis V., Exarchou and K., Barbarassou (2002). *Nuclear Magnetic Resonance (NMR) spectroscopy: Basic principles and phenomena, and their applications to chemistry, biology and medicine*. Chemistry Education: Research and Practice In Europe, 3, 229–252

I., Anderson and R., Oste (1992). *Loss of ascorbic acid, folacin and vitamin B12 and changes in oxygen content of UHT milk.* Introduction and methods. Milchwissenschaft, 47, 223–224

I., Frank and B., Kowalski (1982). *Chemometrics.* Anal. Chem., 54, 232R-243R

IFST (2004). *IFST Current Hot Topics: Genetic Modification and Food*. Institute of Food Science and Technology, UK

J. A., Westerhuis, H. C. J., Hoefsloot, S., Smit, D. J., Vis, A. K., Smilde, E. J. J., van Velzen, J. P. M., van Duijnhoven and F. A., van Dorsten (2008). *Assessment of PLS-DA cross validation*. Metabolomics, 4, 81–89

J. C., Cobas, F. J., Sardinia (2003). *Nuclear Magnetic Resonance Data Processing. MestRe-C: A Software Package for Desktop Computers*. Concepts in Magnetic Resonance Part A, 19, 80–96

J. K., Nicholson and J. C., Lindon (2008). *Systems biology: Metabonomics*. Nature, 455, 1054–1056

J. K., Nicholson, J. C., Lindon and E., Holmes (1999). *"Metabonomics": understanding the metabolic responses of living systems to patho-*

physiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica, 11, 1181–1189

J. M., Poveda, A., García, P. J., Martín-Alvarez and L., Cabezas (2006). *Application of partial least squares (PLS) regression to predict the ripening time of Manchego cheese*. Food Chemistry, 84, 29–33

J. S., Anderson and M., Mann (2000). *Functional genomics by mass spectrometry*. FEBS Lett, 480, 25–31

J., Keeler (2002). *Fourier transformation and data processing* in *Understanding NMR Spectroscopy*, J. Keeler, University of Cambridge, Department of Chemistry, chapter IV

K. G., Penman, C. W., Halstead, A. Matthias, J. J. De Voss, J. M. U., Stuthe, K. M., Bone and R. P., Lehmann (2006). *Bilberry Adulteration Using the Food Dye Amaranth*. J. Agric. Food Chem., 54, 7378–7382

K. J., Lee, K. H. Row and I-C., Jun (2008). *Preparative Separation of Isoflavones from Korean Soybean by HPLC.* Asian Journal of Scientifc Research 1, 288–292

K. R., Kottapalli, P., Payton, R., Rakwal, G. K., Agrawal, J., Shibato, M., Burow and N., Puppala (2008). *Proteomics analysis ofmature seed of four peanut cultivars using two-dimensional gel electrophoresis reveals distinct differential expression of storage, anti-nutritional, and allergenic proteins.* Plant Science, 175, 321–329

K. S., Solanky, N. J., Bailey, B. M., Beckwith-Hall, S., Bingham, A., Davis, E., Holmes, J. K., Nicholson and A., Cassidy (2005). *Biofluid ^{1}H NMR-based metabonomic techniques in nutrition research — metabolic effects of dietary isoflavones in humans*. The Journal of Nutritional Biochemistry, 16, 236–244

K., Belhajjame, S. M., Embury, H., Fan, C., Goble, H., Hermjakob, S. J., Hubbard, D., Jones, P., Jones, N., Martin, S., Oliver, C., Orengo, N. W., Paton, A., Poulovassilis, J., Siepen, R. D., Stevens,

C., Taylor, N., Vinod, L., Zamboulis and W., Zhu (2005). *Proteome Data Integration: Characteristics and Challenges.* Proceedings of the UK e-Science All Hands Meeting

K., Beyer, G., Grishina, L., Bardina, A., Grishin and H. A., Sampson, (2002). *Identification of an 11S globulin as a major hazelnut food allergen in hazelnut-induced systemic reactions*. Journal of Allergy and Clinical Immunology, 110, 517–523

K., Beyer, L., Bardina, G., Grishina and H. A., Sampson (2002). *Identification of sesame seed allergens by 2-dimensional proteomics and Edman sequencing: seed storage proteins as common food allergens*. Journal of Allergy and Clinical Immunology, 110, 154–159

K., Dewettinck, F., Van Bockstaele, B., Kühne, D., Van de Walle, T. M., Courtens and X., Gellynck (2008). *Nutritional value of bread: Influence of processing, food interaction and consumer perception*. Journal of Cereal Science 48, 243–257

K., Fukunaga (1990). *Introduction to statistical pattern classification*. USA: Academic Press

K., Gallardo, C., Job, S. P. C., Groot, M., Puype, H., Demol, J., Vandekerckhove and D., Job (2002). *Proteomics of Arabidopsis seed germination. A comparative study of wild-type and gibberellin-deficient seeds.* Plant Physiol, 129, 823–837

K., Gallardo, C., Le Signor, J., Vandekerckhove, R. D., Thompson and J., Burstin (2003). *Proteomics of Medicago truncatula seed development establishes the time frame of diverse metabolic processes related to reserve accumulation*. Plant Physiol, 133, 664–682

K., Wüthrich, (2002). *NMR Studies of structure and function of biological macromolecule*. Nobel Lecture, 235–267

L. A., Berrueta, R. M., Alonso-Salces and K., Héberge (2007). *Supervised pattern recognition in food analysis*. Journal of Chromatography A, 1158, 196–214

L. I., Smith (2002). *A tutorial on Principal Components Analysis*

L. W., Sumner, P., Mendes and R. A., Dixon (2002). *Plant metabolomics: large-scale phytochemistry in the functional genomics era.* Review. Phytochemistry, 62, 817–836

L., Iawati, R. L., Stroshine and G., Narsimhan (2007). *Low-field NMR: A tool for studying protein aggregation*. J Sci Food Agric, 87, 2207–2216

L., Indrawati, R. L., Stroshine and G., Narsimhan (2007). *Low-field NMR: A tool for studying protein aggregation*. Journal of the Science of Food and Agriculture, 87, 2207-2216

L., Viggiani and M. A., Castiglione Morelli (2008). *Characterization of Wines by Nuclear Magnetic Resonance: A Work Study on Wines from the Basilicata Region in Italy*. J. Agric. Food Chem.,56, 8273–8279

M. A., Al-Holy, M., Lin, A. G., Cavinato and B. A., Rasco (2006). *The use of Fourier transform infrared spectroscopy to differentiate Escherichia coli O157:H7 from other bacteria inoculated into apple juice*. Food Microbiol, 23, 162–168

M. A., Lluch, I., Hernando and I., Pérez-Munuera (2001). *Proteins in food structures* in *Chemical and Functional Properties of Food Proteins*, Z. E., Sikorski Eds., CRC Press Taylor & Francis Group, Boca Raton (FL), London, New York, chap. 2

M. A., Rahman, A. K. M. S., Islam, B. K., Bala and A., Khair (2002) *In vitro NMR Relaxation Study of Water Protons in. the Intracellular Water of Eggplant*. Pakistan Journal of Biological Sciences, 5, 88–90

M. A., Rahman, A. K. M. S., Islam, B. K., Bala and A., Khair (2002). *In vitro NMR Relaxation Study of Water Protons in the Intracellular Water of Eggplant*. Asian Network for Scientific Information, Pakistan, 88-90

M. J. Gibney, M., Walsh, L., Brennan, H. M., Roche, B., German and B., van Ommen, (2005). *Metabolomics in human nutrition: opportunities and challenges.* American Journal of Clinical Nutrition, 82, 497–503

M. J., Rist , U., Wenzel and H., Daniel (2006). *Nutrition and food science go genomic.* Trends in Biotechnology, 24, 172–178

M. R., Ninonuevo, Y., Park, H., Yin, J., Zhang, R. E., Ward, B. H., Clowers, et *al*. (2006). *A strategy for annotating the human milk glycome*. Journal of Agriculture and Food Chemistry, 54, 7471–7480

M. S., Branca, P. B., Andrade, R. C., Martins, R. M., Seabra and M. A. Ferreira (2006). *Principal component analysis as tool of characterization of quince (Cydonia oblonga Miller) jam*. Food Chemistry, 94, 504–512

M., Berth, F. M., Moser, M., Kolbe and J., Bernhardt (2007). *The state of the art in the analysis of two-dimensional gel electrophoresis images.* Appl Microbiol Biotechnol, 76, 1223–1243

M., Bill and P. S., Taoukis (1998). *Evaluation of shelf life of flavored dehydrated products using accelerated shelf life testing and the Weibull Hazard sensory analysis*. Developments in Food Science, 40, 627–637

M., Bunzel and J., Ralph (2006). *NMR Characterization of Lignins Isolated from Fruit and Vegetable Insoluble Dietary Fiber*. J. Agric. Food Chem, 54, 8352–8361

M., Carbonaro (2004). *Proteomics: present and future in food quality evaluation*. Trends in Food Science & Technology 15, 209–216

M., Carbonaro, G., Grant, M., Cappelloni and A. Pusztai (2000). *Perspectives into factors limiting in vivo digestion of legume proteins: antinutritional compounds or storage proteins?* Journal of Agricultural and Food Chemistry, 48, 742–749

M., Carbonaro, I., Sondergaard and K., Bukhave (2003). *Proteomics applied to bioavailability studies: traceability of in vivo gastrointestinal pattern of milk proteins.* In NFIF 2003 Conference, Copenhagen, 9–11 April, 2003

M., D'Imperio L., Mannina, D., Capitani, O., Bidet, E., Rossi, F. M., Bucarelli, G. B., Quaglia and A., Segre (2007). *NMR and statistical study of olive oils from Lazio: A geographical, ecological and agronomic characterization*. Food Chemistry 105, 1256–1267

M., Defernez, Y. M., Gunning, A. J., Parr, L. V. T., Shepherd, H. V., Davies and I. J., Colquhoun (2004). *NMR and HPLC-UV Profiling of Potatoes with Genetic Modifications to Metabolic Pathways*. J. Agric. Food Chem., 52, 6075–6085

M., Fiorini (1995). *Preparative high-performance liquid chromatography for the purification of natural anthocyanins.* Journal of Chromatography A, 692, 213–219

M., Galvani, M., Hamdan and P. G., Righetti (2001). *Two-dimensional gel electrophoresis/matrix-assisted laser desorption ionization mass spectrometry of commercial bovine milk.* Rapid Communication in Mass Spectrometry, 15, 258–264

M., Hajduch, A., Ganapathy, J. W., Stein and J. J., Thelen (2005). *A systematic proteomic study of seed filling in soybean. Establishment of high-resolution two-dimensional reference maps, expression profiles, and an interactive proteome database.* Plant Physiol, 137, 1397–1419

M., Hajduch, J. E., Casteel, K. E., Hurrelmeyer, Z., Song, G. K., Agrawal and J. J., Thelen (2006). *Proteomic analysis of seed filling in Brassica napus. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis.* Plant Physiol, 141, 32–46

M., Ringnér (2008). *What is principal component analysis?* Nature Biotechnology 26, 303–304

M., Servili, R., Selvaggini, S., Esposto, A., Taticchi, G., Montedoro and G., Morozzi (2004). *Health and sensory properties of virgin olive oil hydrophilic phenols: agronomic and technological aspects of production that affect their occurrence in the oil*. Journal of Chromatography A, 1054, 113-127

M-Á., Rodríguez-Delgado, G., González-Hernández, J-E., Conde-González and J-P., Pérez-Trujillo (2002). *Principal component analysis of the polyphenol content in young red wines*. Food Chemistry, 78, 4523–532

N. A., Abumrad (2005). *CD36 may determine our desire for dietary fats*. Journal of Clinical Investigations, 115, 2965–2967

N. J., Serkova, Y., Zhang, J. L.,Coatney, L., Hunter, M. E., Wachs, C. U., Niemann, et al. (2007). *Early detection of graft failure using the blood metabolic profile of a liver recipient*. Transplantation, 83, 517–521

N., Amiour, M., Merlino, P., Leroy and G., Branlard (2002). *Proteomic analysis of amphiphilic proteins of hexaploid wheat kernels*. Proteomics, 2, 632–641

N., Michał (2007). *Mineral Components* in *Chemical and Functional Properties of Food Components, Third Edition*, Z. E., Sikorski Eds., CRC Press Taylor & Francis Group, Boca Raton (FL), London, New York, chap. 4

N., Ogrinc, I. J,. Košir, J. E., Spangenberg and J., Kidric (2003). *The application of NMR and MS methods for detection of adulteration of wine, fruit juices and olive oil. A review.* Analytical and Bioanalytical Chemistry, 376, 424–430

O. E., De Noord, (1994). *Multivariate calibration standardization*. Chemometr. Intell. Lab. Syst, 25, 85–97

O. P., Kuipers (1999). *Genomics for food biotechnology: prospects of the use of high-throughput technologies for the improvement of food micro-organisms*. Current Opinion in Biotechnology, 10, 511–516

O., Moreira Sampaio, R., Vicente Reche and D., Wagner Franco (2008). *Chemical Profile of Rums as a Function of their Origin. The Use of Chemometric Techniques for their Identification*. J. Agric. Food Chem., 56, 1661–1668

P. A., Haynes, L., Hays, D., Schieltz, R., Ulaszek, J., Wei, D., Wolters and J. R., Yates (2002). *Proteomic survey of metabolic pathways in rice*. Proceedings of the National Academy of Science USA, 99, 11564–11566

P. G., Righetti, F., Nembri, A., Bossi and M., Mortarino (1997). *Continuous enzymatic hydrolysis of β-casein and isoelectric collection of some of the biologically active petides in an electric field*. Biotechnology Progress, 13, 258–264

P., Benatti, G., Peluso, R., Nicolai and M., Calvani (2004) *Polyunsaturated Fatty Acids: Biochemical, Nutritional and Epigenetic Properties*. Journal of the American College of Nutrition, 23, 281–302

P., Roncada, A., Gaviraghi, G. F., Greppi and S., Gigli (2002). *A proteomic approach to compare meat from different species.* 48th ICoMST, 2, 640–641

P., Tenbült, N. K., De Vries, G., van Breukelen, E., Dreezens and C., Martijn (2008). *Acceptance of genetically modified foods: The relation between technology and evaluation*. Appetite 51, 129–136

R. J., Abraham, J. J., Byrne, L., Griffiths and R., Koniotou (2005). *^{1}H chemical shifts in NMR: Part 22 – Prediction of the ^{1}H chemical shifts of alcohols, diols and inositols in solution, a conformational and salvation investigation*. Magnetic Resonance Chemistry, 43, 611–624

R. K. D., Tiebach and M., Schramm (1980). *Application of preparative HPLC in the chemical analysis of food ingredients*. Chromatographia, 13, 403–407

R. L., Last, A. D., Jones and Y., Shachar-Hill (2007). *Towards the plant metabolome and beyond*. Nature Reviews Molecular Cell Biology 8, 167-174

R. M., Adibhatla, J., Hatcher and R., Dempsey (2006). *Lipids and Lipidomics in Brain Injury and Diseases*. AAPS Journal, 8, E314–E321

R. M., Elliott and I. T., Johnson (2007). *Nutrigenomic approaches for obesity* research. Obesity Reviews, 8, 77–81

R. R., Ruan and P. L., Chen (1998). *Water in Foods and Biological Materials*, Technomic Publishing Co. Inc., Lancaster, PA

R., Aebersold and M., Mann (2003). *Mass spectrometry-based proteomics*. Nature 422, 198-207

R., Cozzolino, S., Di Giorgi, S., Fisichella, D., Garozzo, D., Lafiandra and A., Palermo (2001). *Proteomics of gluten: mapping of subunit 1AX2* in Cheyenne cultivar by matrix-assisted laser desorption/ionization.* Rapid Communication in Mass Spectrometry, 15, 1129–1135

R., Duda, P., Hart and D., Stork (2000). *Pattern classification*. Wiley

R., Gianferri, V., D'Aiuto, R., Curini, M., Delfini and E., Brosio (2007). *Proton NMR transverse relaxation measurements to study water dynamic states and age-related changes in Mozzarella di Bufala Campana cheese*. Food Chemistry, 105 , 720–726

R., Goodacre, S., Vaidyanathan, W. B., Dunn, G. G., Harrigan and D. B., Kell (2004). *Metabolomics by numbers: acquiring and understanding global metabolite data*. TRENDS in Biotechnology, 22, 245–252

R., Lametsch and E., Bendixen (2001). *Proteome analysis applied to meat science: characterizing post mortem changes in porcine muscle.* Journal of Agricultural and Food Chemistry, 49, 4531–4537

R., Sacchi , F., Addeo and L. Paolillo (1997). *^{1}H and ^{13}C NMR of virgin olive oil.* Magnetic Resonance in Chemistry, 35, S133–S145

R., Sacchi and L. Paolillo (2007). *NMR for Food Quality and Traceability* in *Advances in food diagnostics*, L. M. L., Nollet and F., Toldrá Eds, Blackwell publishing, cap V, 101–117.

S. G., Oliver (2002). *Functional genomics: lessons from yeast*. Phil. Trans. R. Soc. Lond. B, 357, 17–23

S. J., Dixon, Y., Xu, R. G., Brereton, H. A., Soini, M. V., Novotny, E., Oberzaucher, K., Grammer and D. J., Penn (2007). *Pattern recognition of gas chromatography mass spectrometry of human volatiles in sweat to distinguish the sex of subjects and determine potential discriminatory marker peaks*. Chemom. Intell. Lab. Syst, 87, 161–172

S. U., O'Hara and S., Stagl (2001). *Global Food Markets and Their Local Alternatives: A Socio-Ecological Economic Perspective*. Population and Environment: A Journal of Interdisciplinary Studies, 22, 533–554

S., Chevallier, D., Bertrand, A., Kohler and P., Courcoux (2006). *Application of PLS-DA in multivariate image analysis*. J. Chemometrics, 20, 221–229

S., Moco, R. J., Bino, O., Vorst, H. A., Verhoeven, J., de Groot, T. A., van Beek, et *al.* (2006). *A liquid chromatography-mass spectrometry-based metabolome database for tomato*. Plant Physiology, 141, 1205–1218

S., Quaranta, M. G., Giuffrida, M., Cavaletto, C., Giunta, J., Godovac-Zimmermann, B., Canas, C., Fabris, E., Bertino, M., Mombro and A. Conti (2001). *Human proteome enhancement: high-recovery me-*

thod and improved two-dimensional map of colostral fat globule membrane proteins. Electrophoresis, 22, 1810–1818

S., Wold (1994). *PLS for Multivariate Linear Modeling QSAR: Chemometric Methods in Molecular Design. Methods and Principles* in *Medicinal Chemistry*. Van de Waterbeemd H (Editor) Verlag-Chemie

S., Yoo-Soo, B., Kyong-Hwan, I., Dong-Soo, K., Ok-Tae, H., Dong-Yoon, A., In-Ok Ahn, K., Bon Cho, K., Suk-Weon, S., Nak-Sul, C., Seon-Woo, L., Dongho, and C., Hyung-Kyoon (2007). *Finger- printing Analysis of Fresh Ginseng Roots of Different Ages Using ^{1}H-NMR Spectroscopy and Principal Components Analysis*. Arch Pharm Res, 30, 1625-1628

T. M., Ehrman, D. J., Barlow, and P. J., Hylands, (2007). *Phytochemical databases of Chinese herbal constituents and bioactive plant compounds with known target specificities*. Journal of Chemical Informatics and Modeling, 47, 254–263

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

T., Majoul, E., Bancel, E., Triboi, J., Ben Hamida and G., Branlard (2003). *Proteomic analysis of the effect of heat stress on hexaploid wheat grain: characterization of heat-responsive proteins from total endosperm.* Proteomics, 3, 175–183

V. A. McKusick and Frank H. Ruddle (1987), *A new discipline, a new name, a new journal.* Genomics, 1, 1–2

V. M. F., Lai and C. Y., Lii (2004). *Role of saccharides in texturization and functional properties of foodstuffs* in *Chemical and Functional Properties of Food Saccharides*, P., Tomasik Ed., CRC Press, Boca Raton (FL), chap. 11

W., Wang and P. A., Kollman (2001). *Computational study of protein specificity: The molecular basis of HIV-1 protease drug resistance*. PNAS, 98, 14937–14942

X., Jia, K., Hollung, M., Therkildsen, K. I., Hildrum and E., Bendixen (2006). *Proteome analysis of early post-mortem changes in two bovine muscle types: M. longissimus dorsi and M. semitendinosis.* Proteomics, 6, 936–944

X., Sun and R. G. Moreira (1996). *Correlation between NMR proton relaxation time and free fatty acid and total polar materials of degraded soybean oils.* Journal of Food Processing and Preservation 20, 157–167

Y., Iwahashi and H., Hosoda (2000). *Effect of heat stress on tomato fruit protein expression*. Electrophoresis, 21, 1766–1771

Y., Wang, H., Tang, J. K., Nicholson, P. J. Hylands, J. Sampson, I., Whitcombe, C. G., Stewart, S., Caiger, I., Oru and E., Holmes (2004). *Metabolomic strategy for the classification and quality control of phytomedicine: a case study of chamomile flower (Matricaria recutita L.)*. Planta Med., 70(3), 250–255

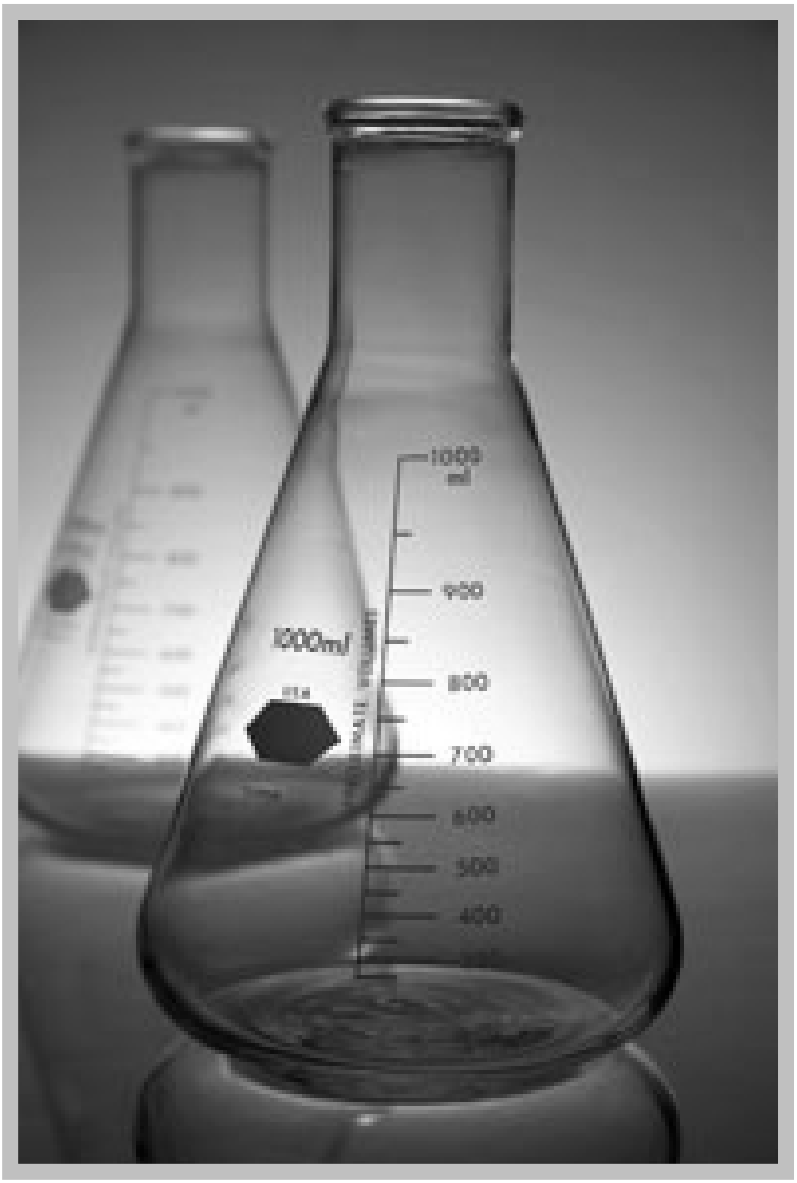
Z. E., Sikorski (2007). *Chemical and Functional Properties of Food Components*, *Third Edition*. Ed., CRC Press Taylor & Francis Group, Boca Raton (FL), London, New York, Introduction

Z., Fang, M., Zhang, G., Tao, Y., Sun and J., Sun (2006). *Chemical Composition of Clarified Bayberry (Myrica rubra Sieb. et Zucc.) Juice Sediment.* J. Agric. Food Chem., 54, 7710–7716

Z., Xiaoyu, W., Huifeng, L., Peiqiu, L., Xiaojing and N., Jiazuan and P., Fengkui (2006). *NMR-based metabonomic study on the sub-acute toxicity of aristolochic acid in rats*. Food and Chemical Toxicology, 44, 1006–1014

CHAPTER TWO

- MATERIAL -



2 MATERIAL

All the materials (from the equipments to the instruments) listed in this chapter and used for this research work are conformed to the most rigorous European and International standards (http://www.msds-europe.com).

2.1 EQUIPMENTS

2.1.1 Lab supplies

- Pipetting Standard - Gilson's Pipetman® P (P10, P20, P100, P200 and P1000. Range of volumes from 10 µL to 1000 µL) with suitable tips (Diamond® precision tip)
- SARSTEDT Polypropylene 50 and 15 mL Conical and Round Bottom Centrifuge Tubes, Falcon™ Type
- Spatula and spoons stainless steel, (VWR international)
- Square, polystyrene and anti-static weighing Dishes (volume 7, 100 e 250 mL)
- Electronic pipet filler Eppendorf Easypet® for graduated and one-mark pipettes from 0.1 mL to 100 mL
- Eppendorf® Safe-Lock® microcentrifuge tubes volume 0.5 and 1.5 mL
- Laboratory ceramic mortar grinder with pestle
- "High-speed" bottles, for High-Performance Centrifuge® by Beckman Coulter™, with screw stopper for reinforced tightness (capacity 500 and 50 mL)

- AMPOL NMR sample Tubes for use up to 700 MHz NMR, (203 mm, round bottom)
- Laboratory glassware: Beakers, low form, with spout (50, 250 and 600 mL) by Simax;
- Corning® Disposable Pasteur Pipettes, Bulk Pack, Non-Sterile, SIGMA-ALDRICH®
- Pasteur pipette rubber bulbs

2.1.2 Safety and protection supplies

- Safety Eyewear Glasses
- Natural Latex Powdered and Powder-Free Exam Gloves
- White lab coat

2. 2 REAGENTS

- Acetic acid (glacial) 100% anhydrous (CH_3COOH, 60.05 g/mol), MERCK
- Methanol (CH_3OH, 32.04 g/mol, 99,5% purity), MERCK
- Deuterium oxide (D_2O, 20.04 g/mol, 99.9% purity), MERCK
- Milliq demineralized water
- Potassium Hydroxide (KOH, 56.11 g/mol) in flave, MERK
- Perchloric acid ($HClO_4$, 100.46 g/mol, 70% solution), SIGMA-ALDRICH®
- Potassium dihydrogen phosphate (KH_2PO_4, 136.09 g/mol), Panreac
- Potassium Hydrogen phosphate anhydrous (K_2HPO_4, 174.18 g/mol), Panreac

- 21 L-Amino Acids plus glycine, 1 g of each, SIGMA-ALDRICH®
- L-Anserine Nitrate Salt ($C_{10}H_{16}N_4O_3$, 303.3 g/mol), SIGMA-ALDRICH
- Taurine in powder ($C_2H_7NO_3S$, 125.15 g/mol), SIGMA-ALDRICH®
- Creatine in powder ($C_4H_9N_3O_2$, 131.13 g/mol), ACROS ORGANICS
- Betaine in powder ($C_5H_{11}N_1O_2$, 117.14634 g/mol), FLUKA
- Trimehylamin Hydrochlorid 98%, TMA, in powder ($C_3H_9N.HCl$, 95.57 g/mol), SIGMA-ALDRICH
- Adenosine diphospahte ADP ($C_{10}H_{15}N_5O_{10}P_2$, 427.201 g/mol), SIGMA-ALDRICH
- Hypoxanthine, 99.5% ($C_5H_4N_4O$, 136.11 g/mol), ACROS ORGANICS

2. 3 BUFFER SOLUTIONS

In this work, two different solutions, due to the different biological matrices (fruits and vegetables), have been used.

- Phosphate Buffer, 100mM and pH 7.00
- Acetic acid/acetate buffer (CH_3COOH/CH_3COONa), 50mM and pH 5.00

2.3.1 Practical ways to make a buffer

Generally, three methods can be used to obtain a buffer solution:

1. The buffer pK_a method;

2. the two solutions method;

3. acid and Basic titration method.

For the phosphate buffer, the two solutions method were used. This method consists to make separate solutions of the acid form, potassium dihydrogen phosphate (KH_2PO_4) and base form of the buffer, potassium phosphate dibasic (K_2HPO_4), both solutions having the same buffer concentration (1M), and ionic strength if required, as the concentration of total buffer in the final solution. To obtain the desired pH, one solution is added to the other while monitoring the pH with a pH meter.

For the Acetic acid/acetate buffer the buffer pK_a method was used. In water, acetic acid establishes an equilibrium between the weak acid, acetic acid, and the conjugate base, acetate ion.

For acetic acid, the value of K_a equals 1.76×10^{-5} and pK_a equals 4.75. The magnitudes of the K_a and pK_a values of different weak acids give us a comparison of their relative strength. A weaker acid has less dissociation to the conjugate base and the equilibrium favors the undissociated weak acid form. This results in a smaller K_a value. A smaller K_a value corresponds to a larger pKa. In other words, the weaker the acid, the larger the pK_a value.

For experimental work in aqueous solutions, it is fundamentally important to be able to prepare a buffer solution at a desired pH. The pKa methods is based on the Henderson-Hasselbalch relationship (R. H. C. Strang, 1981) written in Equation 2.1.

$$K_A = \frac{[H^+][A^-]}{[HA]} \quad \text{(Eq. 2.1)}$$

the equation allows to calculate the correct ratio of basic form to acidic form which can be mixed to achieve the desired buffered pH. For the acetic acid/acetate buffer the Equation 2.1, considering the definition of Ka and pH, is rearranged as shown in Equation 2.2:

$$pH - pK_A = \log \frac{[CH_3COO^-]}{[CH_3COOH]} \quad \text{(Eq. 2.2)}$$

Note that the Henderson-Hasselbalch relationship indicates that the pH of a buffer solution does not depend on the total concentration of the buffering acid and conjugate base but only on the pKa and the ratio of the concentration of these two species.

On the other hand, the buffering capacity of a solution quantifies the amount of H_3O^+ or OH^- the solution is capable of neutralizing before the acid or conjugate base form is saturated and the pH begins to fall or rise precipitously.

This will depend on the total concentration of the acid and conjugate base buffer ions. Also, the buffering capacity may be different towards addition of acid than towards base. This will be true unless the pH of the buffer solution is identical to the pKa of the buffering acid-base equilibrium.

2. 4 STANDARD SOLUTIONS

All the standard solutions used for this research work were prepared since chemical reagents of analytical grade. Each standard was 0.5M in 1 mL of D_2O and the amount of reagents is obtained by applying the following equation (Eq. 2.3)

$$m = MW_{Std} \cdot [Std] \cdot Vol \quad \text{(Eq. 2.3)}$$

where *m* is the final amount of the standard, MW_{std} is the molecular weight, [*Std*] the final concentration (0.5M) and *Vol* the final volume (1 mL).

In Table 2.1 are listed all the standards used for the present work.

Table 2.1 – first part: list of standards used in this researcher work

Standard	Molecular Weight g/mol	Final Concentration [M]
Taurine	125.15	0.5
Trimetilammina	95.57	0.5
Anserine Nitrate Salt	303.3	0.5
Hypoxantine	136.11	0.5
Betaine	135.16	0.5
Taurine	125.15	0.5
L-Serine	105.09	0.5
L-Alanine	89.09	0.5
L-Cystine	121.16	0.5
L-Proline	115.13	0.5
L-Hystidine	209.63	0.5
L-Arginine	210.7	0.5
L-Tryptophan	204.23	0.5
L-Phenylalanine	165.19	0.5
L-Methionine	149.21	0.5
Malic Acid	134.09	0.5
Citic Acid	192.13	0.5
Tartaric Acid	150.09	0.5

Table 2.2 – second part: list of standards used in this researcher work

Standard	Molecular Weight g/mol	Final Concentration [M]
L-Lysine	182.65	0.5
L-Asparagine	132.12	0.5
L-Valine	117.15	0.5
L-Tyrosine	181.19	0.5
L-Leucine	131.18	0.5
L-Aspartic Acid	133.1	0.5
L-Glutamine	146.15	0.5
L-Glycine	75.07	0.5
L-Glutamic Acid	147.13	0.5
L-Cysteine hydrocloride	157.61	0.5

2. 5 INSTRUMENTS

- Eletronic digital technical Balance (max 2200 g, d= 0.01 g), SCALTEC (SBA 52)
- Eletronic digital analytical Balance (max 220g, d= 0.0001g), SCALTEC (SBA 31)
- IKA® ULTRA-TURRAX® homogenizer T 18, basic, AC input 115 V

- JENWAY Model 3310 pH Meter with glass bodied combination electrode, swing arm electrode holder & ATC probe.
- Heating magnetic stirrer mod. ARE, VELP Scientifica®
- Heating magnetic stirrer mod. ARED, by VELP Scientifica®
- Beckman Coulter™ Microfuge® 18 Microcentrifuge (max 14000 rpm adjustable in 500 increments)
- Avanti® J-25 High-Performance Centrifuge® by Beckman Coulter™ equibed by Beckman Coulter's aluminum rotors JA10 RPM_{max} 10000, JA14 RPM_{max} 14000, JA25.50, RPM_{max} 25000
- NMR Varian Mercury-plus AS400/54 (400MHz) spectometer equipped with a 5mm PFG gradient 4 nuclei ($^{1}H/^{19}F/^{13}C/^{31}P$) probehead, with a 400MHz (9.4 Tesla) superconducting magnatic system by Oxford Active Shielded and equipped with Sun BLADE 150 Host Workstation with Solaris 10 OS (80 GB Hard Disk, 512 MB Ram, CD ROM SCSI Drive and VnmrJ 1.1D Software)

2. 6 SOFTWARE

2.6.1 NMR data processing

MestRec (www.mestrec.com), Magnetic Resonance Companion, "is a software package that offers state-of-the-art facilities for data processing, visualization, and analysis of high resolution nuclear magnetic resonance (NMR) data, combined with a robust, user-friendly graphical interface that fully exploits the power and flexibility of the Windows platform. The program provides a variety of conversion facilities for most NMR spectrometer formats

and includes all the conventional processing, displaying, and plotting capabilities of an NMR program, as well as more advanced processing techniques" (J. C. Cobas, et *al.*, 2003).

A pdf format tutorial designed to help to become familiar with MestRe-C's features is available at http://nmr-aci.uni hd.de/Anleitungen/mestrec/mestrec.pdf

2.6.2 Chemometrics data processing

The multivariate statistical analysis were carried out using different statistical software due to the different kind of analysis: R programm, Matlab and Latentix were the software used for these porpoises.

R (http://www.r-project.org) is a language and environment for statistical computing and graphics. The language provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques, and is highly extensible. One of its strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.

Matlab (http://www.mathworks.com) is a high-performance language for technical computing. The name stands for matrix laboratory and was originally written to provide easy access to matrix software. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation; this allows to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of the time it would take to write a program in a scalar noninteractive language such as C or Fortran.

Latentix (http://www.latentix.com) is a new user-friendly stand-alone program for chemometric data analysis. It offers a raw data plot facility as well as comprehensive PCA and PLS modeling.

REFERENCES

R. H. C., Strang (1981). *Phosphate Buffer and the Henderson-Hasselbalch Equation: An Exercise in Theory and Practice.* Biochemical Education, 9, 66

CHAPTER THREE
-TRANSGENIC GRAPE-

3 A METABOLIC PROFILE OF TRANSGENIC GRAPE

The present work proposes two analytical approaches, based on the nuclear magnetic resonance spectroscopy:

1. **PART ONE**: comparing and measuring the global difference in the chemical composition of table grapes wild type with respect to their genetically modified derivative (**metabonomic approach**); Part one is submitted

2. **PART TWO**: attempting to identify and/or quantify as many compounds as possible in order to find out the most important molecular markers responsible of the differences between grape wild type and grape genetically modified (**metabolomic approach**). Manuscript is in preparation

These kinds of approach evaluate the metabolic profile of grapes, as determined on their hydroalcoholic extracts, and exploits the principles of metabon(l)omics (as described in Chapter 1, par. 1.2.3) to analyze the changes induced by genetic modifications.

3.1 INTRODUCTION

Food safety assessment of a transgenic plant or product is strictly defined by international rules, and European Union and United States regulators had formerly adopted as main concept the approach of substantial equivalence: if a new food or food component is found to be substantially equivalent to an existing food or food component, it can be treated in the same manner with respect to safety (L. Levidow et *al.*, 2007). Substantial equivalence might be used to demonstrate similarity, and therefore, safety, mainly through tests of physicochemical composition. However, many criticisms were addressed to the concept of substantial equivalence be-

tween transgenic and traditional foods: it emphasizes chemical composition at the expense of biological, toxicological, and immunological tests; it does not define the point at which a food is no longer substantially equivalent; and the concept actually impedes risk research (E. Millstone et *al.*, 1999). Presently, substantial equivalence is still considered a key step in the safety assessment process of transgenic foods, and it is not a safety assessment in itself. The concept of substantial equivalence should be improved by the development and application of new techniques, which can help to identify unintended and potentially harmful changes (Anon, 2000). Such techniques should provide a general profile of the biochemical composition in order to detect unknown changes. When comparing a transgenic food and the conventional counterpart, "if the differences exceed natural variations, a nutritional and toxicological assessment is required for the transgenic food" (World Health Organization, 2000). This became relevant when the differences from natural variations are not expected by knowing the modifications induced by the inserted gene.

In general, not only for GMO, it should be noted that a change of foodstuff composition is quantifiable even in the absence of a clear identification of each individual molecule undergoing the change of concentration. For example, a change of the area of an unidentified peak in a chromatogram corresponds to a proportional change of the concentration of its corresponding molecule. Moreover, a change may have a negative sign and therefore does not arouse toxicological alarm, but poses a nutritional problem. In this perspective, the entire molecular profiles should be compared in order to evaluate whether the variation observed for the transgenic derivative is higher than the one found in the conventional line. The next step identifies only those molecules undergoing the change of concentration. The alarm arouses if the change may have effects on health, such as an increased concentration of potentially toxic substances or a decreased concentration, under a critical threshold, of substances with high nutritional importance. Several strategies

have been developed to identify unintended alterations in the composition of genetically modified (GM) food crops that may occur as a result of the genetic modification process (H. A. Kuiper et *al.*, 2003). These include comparative chemical analysis of single compounds in GM food crops and their conventional non-GM counterparts, and profiling methods. Three main cell constituent groups are targeted by profiling technologies: RNA - microarray technologies (R. Batista et *al.*, 2008), proteins – proteomics (Chapter 1, par 1.2.2) and metabolites – metabonomic (Chapter 1, par 1.2.3). Different analytical platform have been employed to investigate on the metabolites' profile such as magnetic resonance (NMR) and mass spectrometry. In particular, the NMR technique has the advantage to provide at least one signal for each molecule, present in the mixture at a detectable concentration, and for this reason is often indicated as the universal detector. The spectroscopic data are then explored by chemometric techniques, such as the unsupervised Principal Component Analysis (PCA), in order to simplify and condensate in few parametric descriptors the global information given by the spectra that describe the whole chemical composition (K. V. Mardia et *al.*, 1979).

3.2 MATERIALS AND METHODS

3.2.1 Plant Material

In grape and other species, fruit set and growth is usually triggered by pollination, and correlates with elevated endogenous auxin (IAA) levels. In this regard, genes either increasing auxin synthesis or sensitivity and altering auxin signal transduction allow fruit set in the absence of pollination (parthenocarpy) (G. L. Rotino et *al.*, 1997; N. Ficcadenti et *al.*, 1999; N. Acciarri et *al.*, 2002; T. Pandolfini et *al.*, 2002; N. Carmi et *al.*, 2003; B. Mezzetti et *al.*, 2004; H. Wang et *al.*, 2005; Z. Yin et *al.*, 2006). Thus, auxin might

improve fruit set and consequently the fruit number per plant. In perennial species, auxin also affects inflorescence development. This is indicated by the increased number of inflorescences on raspberry and strawberry plants genetically engineered with an ovule-specific auxin-synthesising gene (B. Mezzetti et *al*., 2004). The *DefH9*-iaaM gene construct (Figure 3.1) consists of the ovule-specific regulatory regions from *DefH9* isolated by *Antirrhinum majus* and the iaaM coding region from *Pseudomonas savastonoi*.

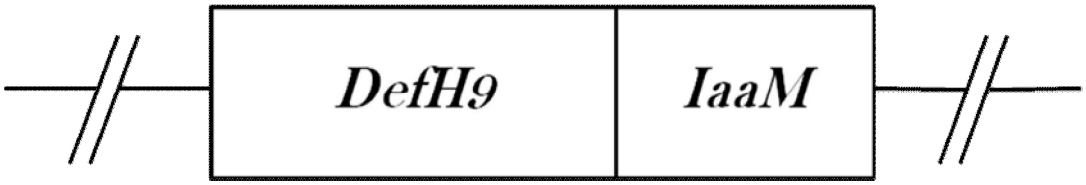


Figure 3.1: the chimeric *DefH9*-iaaM gene construct

The iaaM codes for a tryptophan-2-monooxygenase enzyme that converts tryptophan to indole-3-acetamide which is then hydrolyzed to the auxin indole acetic acid (IAA) (G. L. Rotino et *al*., 1997) (Figure 3.2).

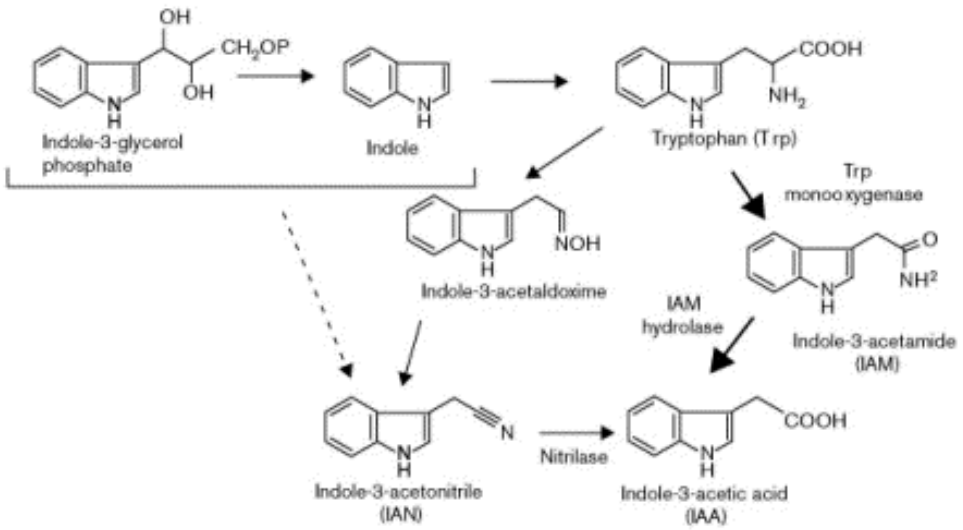


Figure 3.2: biosynthetic intermediates and pathways for IAA biosynthesis. The pathway from tryptophan (Trp) to IAA via indole-3-acetamide (thick arrows) has been demonstrated so far only in bacteria. Indole-3-acetonitrile, which is converted to IAA by nitrilase, is thought to be a biosynthetic intermediate for IAA in *Arabidopsis* and may be formed from Trp and by Trp-independent pathways (Figure is taken from P., Hedden and A. L., Phillips (2000). *Manipulation of hormone biosynthetic genes in transgenic plants.* Current Opinion in Biotechnology 11, 130-137**)**

DefH9-iaaM was introduced into the genome of two grape cultivars with different levels of fecundity mainly due to different shoot fruitfulness (B. Mezzetti et *al.*, 2004):

Thompson Seedless, a well known table grape cultivar has a low shoot fruitfulness;

Silcora, another table grape cultivar, has a shoot fruitfulness higher than Thompson Seedless.

The open field experimental trial with transgenic and control clones was established at the Experimental Farm of the Marche Polytechnic University in March 2001, by following the EC (CE 2001/18) rules for GM plants.

For Silcora cultivars two kind of modified lines were created: one with one *DefH9*-iaaM gene copy and a second with three copies of the same gene.

The Silcora three copies of gene modified lines didn't give significant results, whilst the (one gene copy) transgenic clones of both Thompson and Silcora were identified, propagated and compared in field trial with their corresponding controls not GM. Transgenic *DefH9*-iaaM lines of both cultivars have been cultivated under open field conditions to compare their fecundity to that one of control non-transgenic plants. Shoot fruitfulness and other relevant parameters were recorded over a 3 year-long production cycle lasting from 2004 to 2006 (E. Costantini et *al.*, 2007). In Thompson cultivar the *DefH9*-iaaM gene causes a two-fold increase of shoot fruitfulness, whilst shoot fruitfulness was unaffected in genetically modified Silcora. Both genetically modified cultivars showed an increased number of berries per cluster (Table 3.1), but with an higher entity for the Thompson (30%) in comparison with Silcora (15%). Berries of the GM cultivars maintained a substantial equivalent nutritional quality (Table 3.2) (E. Costantini et *al.*, 2007).

Table I. *Vegetative development[a], fecundity[b], and total plant fruit production of Thompson seedless and Silcora DefH9-iaaM lines in comparison with controls*

Data collected during the whole production cycle consisting of three years of cultivation: 2004, 2005, and 2006. Mean ± SE; means in the same row that are followed by the same letter are not different ($P \leq 0.05$) using the LSD test.

Clones	No. Shoots/Plant	Average Shoot Dry Weight	Fruitful Shoots	Shoot Fruitfulness	No. Bunches/Plant	No. Berries/Bunch	Berry Weight	Bunch Weight	Plant Fruit Production
		g	%				*g*	*g*	*kg*
[AQ28] TS CT	17.5 ± 0.36a	34 ± 2.7a	14.1 ± 2.5b	0.14 ± 0.01b	2.88 ± 0.33b	189.0 ± 7.7b	0.95 ± 0.02a	180.7 ± 8.8b	0.55 ± 0.05b
TS GM	16.1 ± 0.32b	31 ± 2.3a	27.2 ± 7.8a	0.27 ± 0.02a	5.97 ± 0.47a	247.6 ± 10.3a	0.87 ± 0.01b	217.2 ± 10.1a	1.17 ± 0.09a
S CT	16.4 ± 0.58a	13 ± 1.8a	70.0 ± 4.6a	0.86 ± 0.04a	17.38 ± 1.00a	98.4 ± 3.7b	2.88 ± 0.13b	289.9 ± 20.2b	4.88 ± 0.43b
S GM	17.0 ± 0.56a	15 ± 2.2a	66.6 ± 3.8a	0.80 ± 0.05a	15.38 ± 1.25a	113.5 ± 5.3a	3.24 ± 0.19a	376.1 ± 29.3a	5.31 ± 0.44a

[a]Expressed as number of shoots per plant and average shoot dry weight (grams). [b]Expressed as percentage of fruitful shoots, shoot fruitfulness (average number of inflorescences per shoot), number of bunches per plant, number of berries per bunch, berry weight (grams), and bunch weight (grams).

Table 3.1: comparison of shoot fruitfulness, bunches and berry number data between cultivars wild type (Thompson and Silcora) and *DefH9-iaaM* genetically modified ones. The Figure was adapted from E., Costantini, L., Landi, O., Silvestroni, T., Pandolfini, A., Spena and B. Mezzetti (2007*). Auxin synthesis-encoding Transgene Enhaches fecundity*. Plant Physiol, 143, 1689-94

Table II. *Berry quality[a] and nutritional value[b] of Thompson seedless and Silcora DefH9-iaaM lines in comparison with controls*

Data collected at harvest of the three years of cultivation: 2004, 2005, and 2006. Mean ± SE; means in the same row that are followed by the same letters are not different ($P \leq 0.05$) using the LSD test.

Clones	SS	TA	pH	TaA	MA	CA	TPH	TAC Trolox
				g/L	*g/L*	*g/L*	*mg gallic acid/g fruit*	*µmol/[AQ29] g*
TS CT	18.8 ± 0.35a	9.6 ± 0.35a	3.1 ± 0.02b	8.6 ± 0.27a	3.4 ± 0.23a	0.30 ± 0.02a	0.88 ± 0.03b	3.1 ± 0.08a
TS GM	18.7 ± 0.38a	8.0 ± 0.24b	3.2 ± 0.03a	8.3 ± 0.19a	2.4 ± 0.32b	0.22 ± 0.01b	0.96 ± 0.02a	2.9 ± 0.06a
S CT	15.2 ± 0.60a	5.5 ± 0.40a	3.3 ± 0.05a	7.0 ± 0.36a	1.3 ± 0.18a	0.11 ± 0.01a	1.41 ± 0.11b	4.8 ± 0.21b
S GM	15.6 ± 0.39a	5.4 ± 0.39a	3.3 ± 0.04a	6.7 ± 0.32a	1.5 ± 0.20a	0.12 ± 0.01a	1.76 ± 0.06a	6.6 ± 0.07a

[a]Expressed as soluble solids contents (SS), TA, pH, tartaric acid (TaA), MA, and CA. [b]Expressed as TPH and TAC.

Table 3.2: comparison of nutritionl values data between cultivars wild type (Thompson and Silcora) and *DefH9-iaaM* genetically modified ones. The Figure was adapted from E., Costantini, L., Landi, O., Silvestroni, T., Pandolfini, A., Spena and B. Mezzetti (2007*). Auxin synthesis-encoding Transgene Enhaches fecundity*. Plant Physiol, 143, 1689-94

3.2.2 Experimental design for NMR analysis

Fourty-five NMR samples were prepared from the extracts of the two cultivars and their modified lines. At the 2007 harvest, berries of Thompson Seedless control (Thompson-WT) and of the *DefH9-iaaM* Line (Thompson-GM1), were sampled. At the same time berries of Silcora from control plants (Silcora-WT) and from two *DefH9-iaaM* modified lines, one containing a single copy of the gene (Silcora-GM1), the other containing 3 copies of it (Silcora-GM2), were also sampled (E. Costantini et *al.*, 2007). At harvest, samples were freeze dried at -80° and then shipped in dry ice at the laboratory of analyses.

3.2.3 Sample Preparation

For each genotype (control and transgenic lines of both cultivars) was prepared a bulk of 100 grams of berries picked randomly from grape clusters of plants growing in different plots of the experimental vineyards. Both skin and pulps of each sample were homogenised under ice chilling using an ultra turrax T18 basic dispersing tool (IKA®). Three aliquots of about 10,0 g of homogenate, poured in 50 ml Falcon tubes, were separately vortexed with 10,0 ml of a mixture of methanol and 50mM CH_3COOH/ Na^+ buffer pH 5.0 (2:1) in order to perform the solvent extraction. The suspensions were centrifuged at 10,000 rpm for 20 minutes at 4° C. The resulting hydro-alcoholic solution was dispensed in different Eppendorf tubes, as 1 ml aliquots, and stored at -80° C.

Before the FID acquisition, 10% (v/v) D_2O was added to each 1 ml extract, thus centrifuged at 14,000 rpm for 5 minutes at room temperature. A volume of 800 μl was transferred to a 5-mm NMR tube in order to acquire a single NMR FID. Three extracts (**E1-3**) for each homogenate (**H1-3**) obtained from all genotypes (**G1-5**) were subjected to NMR analysis, by alternating samples so that the five extracts E1:H1:G1-5 were firstly analyzed, thus followed by the

series E2:H2:G1-5. When the last series E3:H3:G1-5 was analyzed, totally 45 spectra were acquired. With this sampling scheme, illustrated in Table 3.3, accuracy and precision of the instrumental analysis, as well as the storage effects were assessed.

3.2.4 NMR Spectroscopy

The ^{1}H-NMR spectra were recorded at T= 300K on a Varian Mercury-plus spectrometer, operating at ^{1}H frequency of 400 MHz; for each spectrum, 2048 scans were acquired, with data collected into 16K data points with a spectral width of 16 ppm, a pulse angle of 60°, a recycle delay of 1.0 s, and acquisition time of 2.561 s. The water and methanol singlets were suppressed using the WET presaturation sequence, with irradiation at the water and methanol frequencies. Methanol satellites were suppressed by irradiation on the ^{13}C frequency of the solvent. The data were acquired under an automatic procedure, requiring about 2 hours per sample.

Free induction decays (FID) were Fourier transformed, with the MestReC Software, by performing an exponential multiplication with a 1 Hz line broadening.

The glucose's β-anomeric signal at 4.4 ppm was taken as chemical shift reference for all spectra. Phase and multipoint manual baseline corrections were performed in duplicate for each FID in order to evaluate the errors due to the processing steps. Spectral data points were reduced from 16 K (16384) to 8000 points, by deleting the edge parts of the spectra, containing any signal above the noise, and by cutting off the solvents' and buffer signals (water, methanol and acetic acid).

The spectral information was further condensed by subdividing the spectra into 80 bins, each integrating 100 data points.

The resulting binned spectra were saved as ASCII file for the subsequent statistical analysis.

G1(Thompson-WT)	G2(Thompson-GM1)	G3(Silcora-WT)	G4(Silcora-GM1)	G5(Silcora-GM2)

Homogeneization (H)

100 grams of berries homogenized from which 3 aliquots of 10 grams were separately subjected to extraction

Extraction (E)

3 extractions (E1-3) from 10 g of each homogenized (H1-3) for each genotype (G1-5)

G1	G1	G1	G2	G2	G2	G3	G3	G3	G4	G4	G4	G5	G5	G5
H1	H2	H3	H1	H2	H3	H1	H2	H3	H1	H2	H3	H1	H2	H3
E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3	E1	E2	E3
1	7	13	4	10	16	19	28	37	22	31	40	25	34	43
2	8	14	5	11	17	20	29	38	23	32	41	26	35	44
3	9	15	6	12	18	21	30	39	24	33	42	27	36	45

Table 3.3: sampling scheme for grape fruits sample. For each genotypes, 3 extract from three 10 grams aliquots of 100 grams of homogenized berries were performed

3.2.5 Data Analysis

The ASCII files corresponding to the ^{1}H-NMR spectra were subjected to multivariate analysis by using home-made algorithms (Appendix A) written in the R program language (version 2.4.0). PCA, ANOVA, T-Student test and LDA were carried out on each spectrum by using built in commands in the R program environment, i.e. *prcomp* (centered and scaled data), *anova*, *t.test* and *ld* respectively.

3.3 RESULTS AND DISCUSSION

3.3.1 NMR Spectra

Forty-five Free induced decay (FID) acquired on the same number of grape fruit's extracts were Fourier transformed in duplicate, obtaining 90 spectra, in order to evaluate the effect of the processing on the reproducibility of the data. In Table 3.4 are listed all the spectra samples with their relative names and pH values, whilst in appendix B are illustrated all the spectra subdivided in two three groups according to the three main spectral region (downfield, midfield and upfield regions). Figure 3.3 shows a typical proton spectrum (^{1}H-NMR spectrum), recorded at 400 MHz, of a Thompson seedless wild type (T-WT) sample. For clarity the downfield region (< 5.5 ppm) and the upfield region (<2.9 ppm) were amplified in order to better point out the signals belonging to these regions. Strong resonances, assigned to residual water (4.66 ppm), labelled with number 2, methanol (3.16 ppm) labelled with number 4 and acetic acid (1.86 ppm), labelled with number 5, are easily identified and excluded from the subsequent chemometric analysis. Among several hundred signals, belonging to grape's metabolites, the most intense ones arise from glucose (α-Glu-H1 at 5.01 ppm (label #1), β-Glu-H1 at 4.4 ppm (label #3) and H2-H6 in the range 3.5-4.0 ppm).

Table 3.4: data set of grape fruits' spectra (part one)

Cultivar	Extract	Samples' names	Spectra's names	pH
Thompson Seed-less Wild Type (T-WT)	1	1TW	1TW-1TWbis	4.25
		2TW	2TW-2TWBIS	4.29
		3TW	3TW-3TWbis	4.27
	2	4TW	4TW-4TWbis	4.21
		5TW	5TW-5TWbis	4.21
		6TW	6TW-6TWbis	4.18
	3	7TW	7TW-7TWbis	4.30
		8TW	8TW-8TWbis	4.16
		9TW	9TW-9TWbis	4.18
Thompson *DefH9-iaaM* gene modified line (T-GM1)	1	10TG	10TG-10TGbis	4.28
		11TG	11TG-11TGbis	4.29
		12TG	12TG-12TGbis	4.27
	2	13TG	13TG-13TGbis	4.27
		14TG	14TG-14TGbis	4.26
		15TG	15TG-15TGbis	4.18
	3	16TG	16TG-16TGbis	4.18
		17TG	17TG-17TGbis	4.27
		18TG	18TG-18TGbis	4.27
Silcora Wild Type (SIL)	1	1SW	1SW-1SWbis	4.27
		2SW	2SW-2SWBIS	4.27
		3SW	3SW-3SWbis	4.27
	2	4SW	4SW-4SWbis	4.29
		5SW	5SW-5SWbis	4.29
		6SW	6SW-6SWbis	4.29
	3	7SW	7SW-7SWbis	4.30
		8SW	8SW-8SWbis	4.27
		9SW	9SW-9SWbis	4.25
Silcora *DefH9-iaaM* gene modified line (GM1-SIL) one copy gene	1	10SG1	10SG1-10SG1bis	4.27
		11SG1	11SG1-11SG1bis	4.21
		12SG1	12SG1-12SG1bis	4.24
	2	13SG1	13SG1-13SG1bis	4.32
		14SG1	14SG1-14SG1bis	4.29
		15SG1	15SG1-15SG1bis	4.29
	3	16SG1	16SG1-16SG1bis	4.43
		17SG1	17SG1-17SG1bis	4.46
		18SG1	18SG1-18SG1bis	4.45

Table 3.4: data set of grape fruits' spectra (part two)

Cultivar	Extract	Samples' names	Spectra's names	pH
Silcora *DefH9-iaaM* gene modified line (GM2-SIL) three copies gene	1	19SG2	19SG2-19SG2bis	4.30
		20SG2	20SG2-20SG2bis	4.31
		21SG2	21SG2-21SG2bis	4.31
	2	22SG2	22SG2-22SG2bis	4.39
		23SG2	23SG2-23SG2bis	4.38
		24SG2	24SG2-24SG2bis	4.42
	3	25SG2	25SG2-25SG2bis	4.44
		26SG2	26SG2-26SG2bis	4.45
		27SG2	27SG2-27SG2bis	4.44

3.3.2 Spectral Data Pre-treatment

As it has been told in par 1.3.3 NMR spectral data often need to be preprocessed in several ways in order to conform to the prerequisites for chemometric data analysis. Thus, prior to multivariate analysis, data underwent pre-statistical improvement, aiming at minimizing unwanted sources of variation due to slightly different instrumental conditions, imperfect baseline and phase corrections, and sample preparation artefacts. Such operations include first the choice of a pH independent signal (peak) to which refer all spectra (**alignment**), than the **normalization**, which mostly minimizes the differences due to dilution errors during samples preparation, as well as to small differences in the tuning conditions of the spectrometer. Moreover, the signals belonging to some titrable organic acids still show some variations of chemical shift among different spectra, due to small differences (<0.05 units) in the pH of the extract being analyzed.

To avoid such a detrimental effect, the spectral data underwent to a **binning**, thus converting each spectrum in a collection of 80 bins, each consisting of the integral area over 100 consecutive spectral data points.

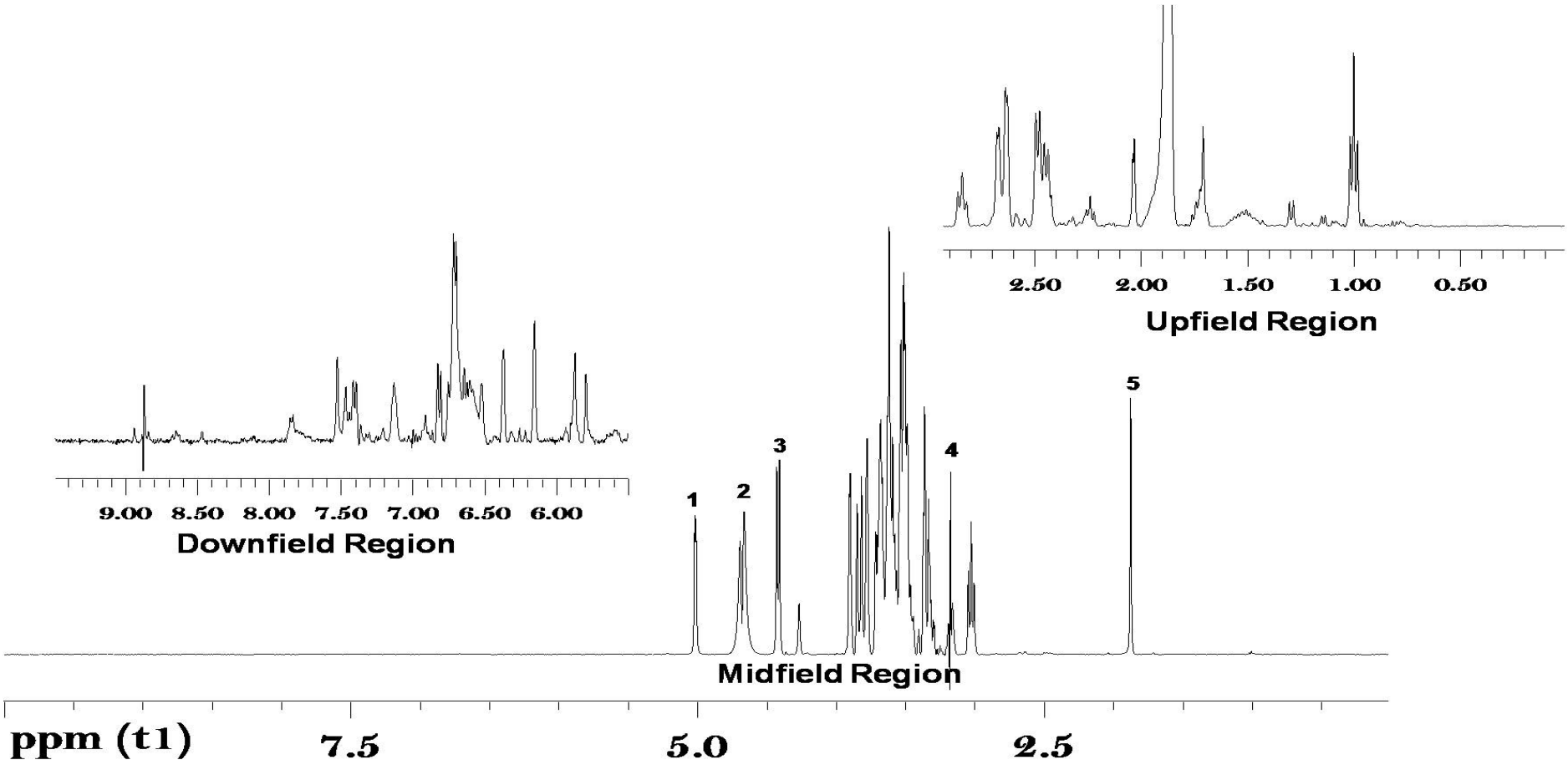


Figure 3.3: ^{1}H-NMR spectrum of hydro-alcoholic extract of Thompson seedless Wild Type (T-WT) grape (*Vitis vinifera*). Downfield and upfield regions were expanded on the vertical scale in order to appreciate the presence of small signals. Some signals, easily assigned, are labelled: **1**, α-D glucose; **2,** residual water; **3**, β-D glucose; **4**, residual methanol and **5**, Acetic Acid (the Figure was prepared by using MestReC software)

3.3.3 Multivariate Analysis

The multivariate analysis was carried out onto binned spectra through the principal component analysis (PCA). The latter has been chosen as the standard gold for comparison, since it is an unsupervised method able to describe the total sample variance by projecting it in a condensed space (Chapter 1, par 1.2.4). According to Chapter 1, par 1.3.3 and par 3.3.2, the spectral data set has been subjected to a prior pre processing in order to improve the PC analysis. Moreover, the PCA screening was applied using a correlation method in which each single data bin is standardized by centering its area integrals with respect to the mean value among all spectra (mean centering), and by scaling the variance of each bin integral by expressing it in terms of standard deviation units. Commonly this method is called **autoscaled analysis**. In this way, the resulting PC scores take into accounts each bin in the same manner independently of the fact that it represents major or minor constituents of the mixture. This choice is in agreement with the principle that the nutritional value or the toxicological effect of a substance or referred to food chemical composition may not be related only on its absolute amount and minority species should be considered in the same manner as the majority one. Since the present study is aimed at evaluating the effect of the genetic mutation on the metabolite content of grapes' extracts (**metabonomic approach**), it has been firstly performed separately the PC analysis on two subsets of the NMR spectra: the ones recorded on the Silcora seedless cultivar, including the wild type (S-WT) as well as the first (S-GM1) and the second genetic variant (S-GM2), and the ones relative to the Thompson seedless cultivar (both T-WT and T-GM).

3.3.4 PCA of the Silcora cultivar subset

The application of the principal component analysis, on the Silcora subset, originates the PC plot shown in Figure 3.4, where the first

principal component (PC1) describes 25% and the second one (PC2) 13% of the total subset variance.

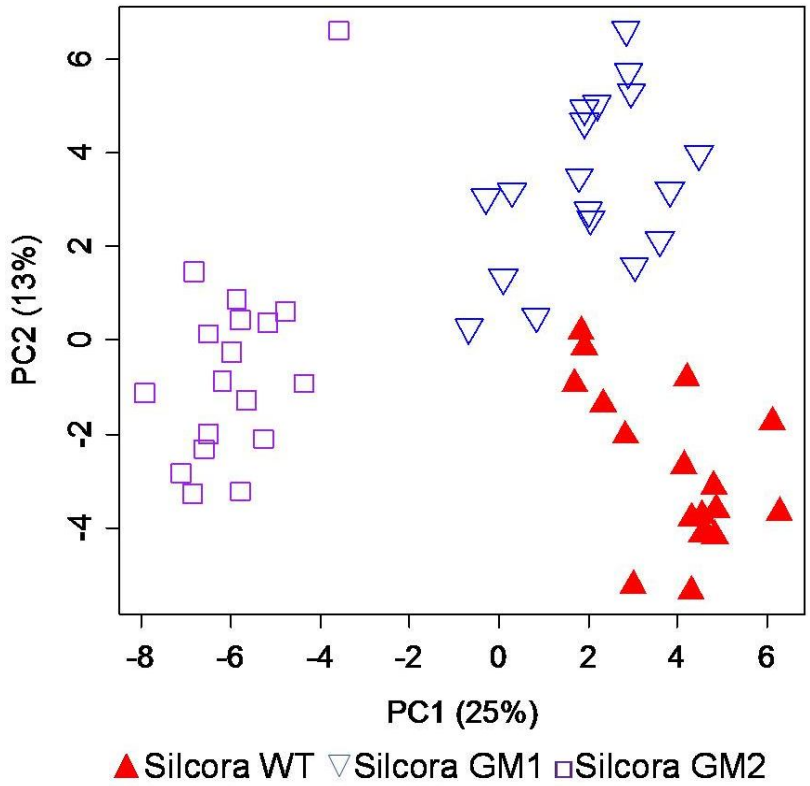


Figure 3.4: autoscaled (meancentered and scaled data) PCA of binned NMR spectral data set, including all the Silcora samples. According to the total variance (**Tv**), PC1 and PC2 explain 25% and 13% respectively. As it can be seen, all the genotypes are differentiated (the Figure was prepared by using R program; see Appendix A)

As emerging from the visual inspection of the PC plot, the variance among the genetically modified grapes and the wild type samples is higher than the variance internal to each group. Moreover, the direction of separation between S-GM1 and S-WT is different (mainly along PC1) from the one along which S-GM2 is separated from S-WT (mainly PC2). This result is mainly interpreted by assuming that the spectral features responsible for the differentiation of GM1 group from the natural species are different from the ones differentiating GM2. The further details arise from the inspection of the PC loadings of the most meaningful components. In fact, the contribution of each variable (a single NMR signal) to a PC can also be calculated, giving a **loading**. A high loading indicates a strong contribution of the original NMR signal to the investigated

PC. Loadings can be displayed in several manners, like the follow pictures. In Figure 3.5, the PC1 loadings are shown as a barplot in which are described as lines then, with a position equal to the position of the group of variables (bins) in the original spectra. The height of the lines-bars indicates the contribution of the variables-bins to the investigated direction. In summary, the loadings report the weight with which each bin influences the position of the sample within the PC plot. For instance, in the same picture are reported the main bins, clusterized into three groups.

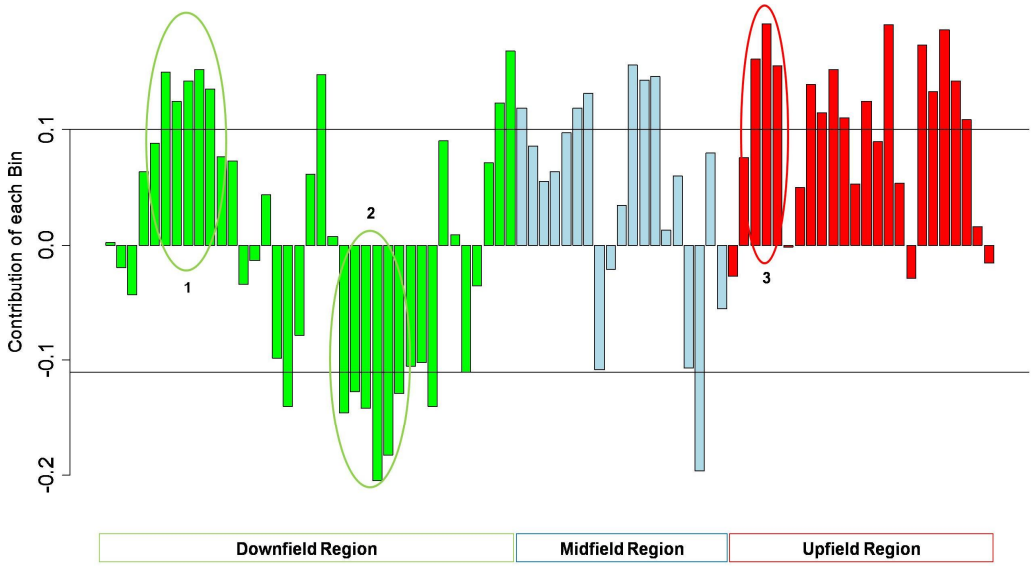


Figure 3.5: loading plots along PC1. Each bar corresponds to a spectral bin of 100 spectral points of length. Groups labeled with #1 and #2, located in the downfield region, and #3, in the upfield region, underlined with a colored circles, are responsible of samples' separation along PC1. Both y=a and y=-b lines indicate for convenience the maximum and minimum values to choose the most representative bins (the Figure was prepared by using R program; see Appendix A)

Increasing areas of the bins in the group labelled with #1 and 3, push the sample scores (S-WT and S-GM1) towards higher values of PC1. The opposite occurs with group labelled with #2, for which lower values of PC1 corresponding to an increasing areas of samples S-GM2 relative bins. At the same manner, according to PC2 loadings plot shown in Figure 3.6, increasing area of the bin labelled with #4 raises the PC2 score.

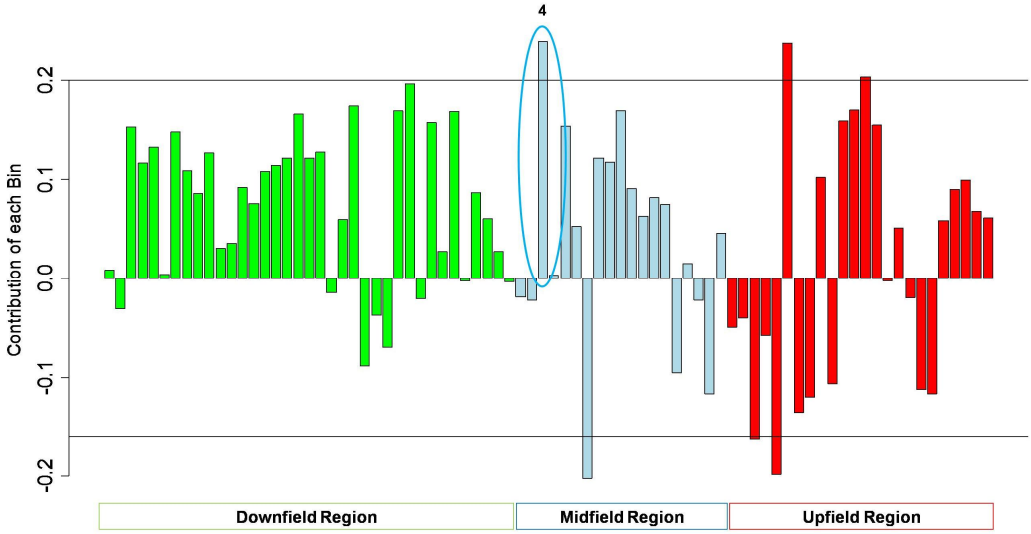


Figure 3.6: loading plots along PC2. Groups labeled with #4, located in the midfield region, underlined with a blue circle, is responsible of samples separation along PC2 (the Figure was prepared by using R program; see Appendix A)

It is worth noting here that the source of variation is not confined to few bins, thus few signals and then molecules. Rather, the whole metabolites' profile is subjected to changes. The absolute extent of such changes is, however, not directly interpretable from the analysis of PC loadings. In order to understand the extent of changes in the amount of metabolites, the absolute area of the bins' groups labelled in Figures 3.5 and 3.6 is reported in Figure 3.7 (A, B, C and D).

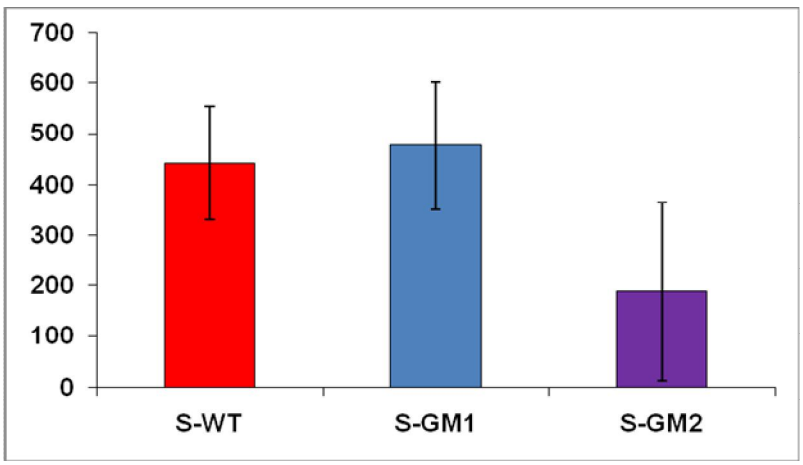


Figure 3.7 A: absolute area of the bins' groups labelled with # 1 in PC1 loadings plot. It corresponds to a bins' groups belonging to the downfield region characterized by the presence of aromatic compounds' signals. Standard errors are also shown as black lines on the top of each bar

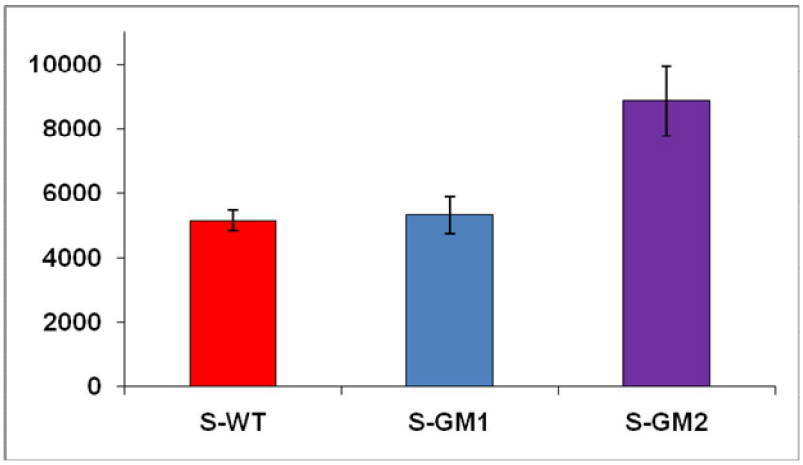


Figure 3.7 B: absolute area of the bins' groups labelled with # 2 in PC1 loadings plot. Like the previous bins' group, even these bins belong to the downfield region and thus to aromatic compounds. Standard errors are also shown as black lines on the top of each bar

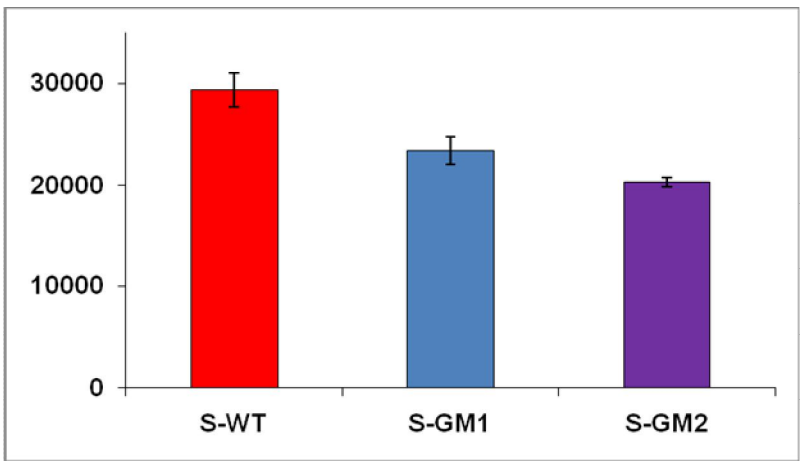


Figure 3.7 C: absolute area of the bins' groups labelled with # 3 in PC1 loadings plot. These bins belong to the upfield region and thus to organic compounds such as organic acids like citric and malic acid. Standard errors are also shown as black lines on the top of each bar

Bins' group #1 (including signals from aromatic substances) is clearly smaller in the second variant of the Silcora grapes than in the other two genotypes of the same cultivar. The same genotype, however, has higher amount of other aromatic compounds, grouped in bins labelled with #2, compared with S-WT and S-GM1. A different trend is observed for bins' group #3, including signals from organic acids, since its area decreases in the first variant (S-GM1) and even more in the second variant (S-GM2).

The bin labelled with #4 (Figure 3.9 D), including signals from the midfield spectra region such as those belonging to sugars, is involved in the separation along PC2. This bin, indeed, has area slightly higher in the first variant GM1 than in the other genotypes, and collects signals belonging to sugars.

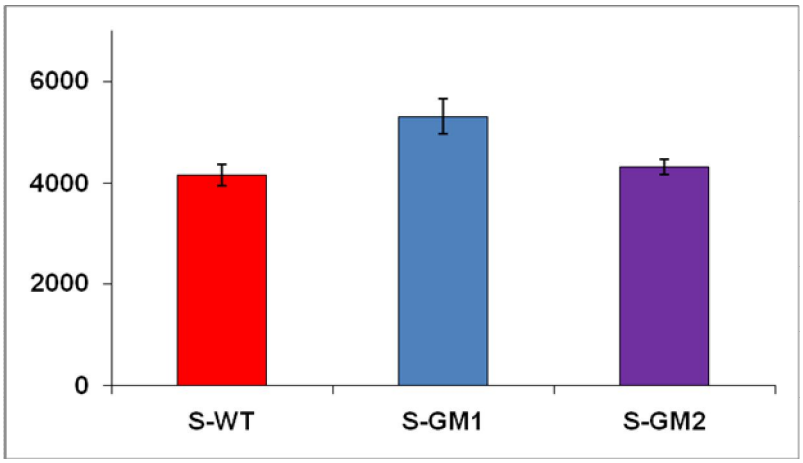


Figure 3.7 D: figure shows the absolute area of the bins' groups labelled with #4 in PC2 loadings plot. It corresponds to bin 40th. Standard errors are also shown as black lines on the top of each bar

A further step in the chemometric analysis of the effect induced by genetic modification of grape berries is represented by the statistical description of the discrimination ability covered by each PC dimension. The result of the t-student test, applied on the PC1 and PC2 scores of all variant, with respect to the WT genotype, is reported in Table 3.5.

The p-value represents the probability of being correct to assume that the two compared populations are equivalent on the basis of their PC score.

For the Silcora cultivar, it emerges that the two genetically modified variants are both statistically different form the wild type genotype, at least along one dimension of the PC space, the latter representing a large portion of the total variance attached to the metabolic profile described by the NMR spectrum.

Table 3.5: summary of t-student test for PC1 and PC 2 applied to Silcora and Thompson subsets (T-test was performed using R program; see appendix A)

PCs	Cultivars	| t |	df	p-value
1	Silcora (WT vs GM1)	4.19	34	1.8e-4
	Silcora (WT vs GM2)	24.16	34	2.2e-16
2	Silcora (WT vs GM1)	10.59	34	2.6e-12
	Silcora (WT vs GM2)	3.35	34	1.9e-3
1	Thompson (WT vs GM)	4.6804	34	4.4e-5
2	Thompson (WT vs GM)	0.6228	34	5.4e-1

| t | is the absolute t value observed in the statistical analysis; df means degrees of freedom

3.3.5 PCA of the Thompson cultivar subset

The PCA approach was also applied to the Thompson cultivar subset, and the results can be summarized through the corresponding PC plot shown in Figure 3.8.

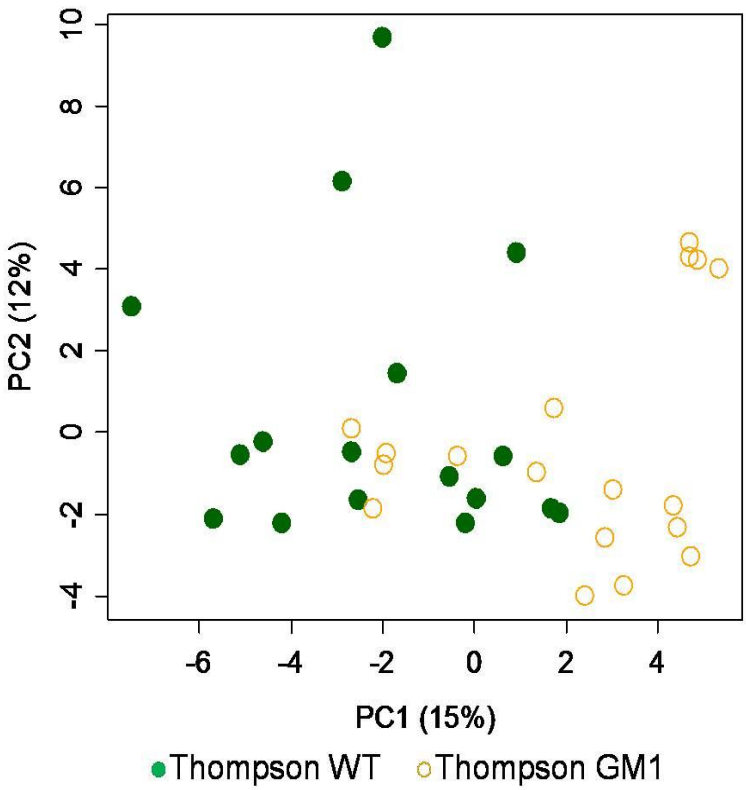


Figure 3.8: PCA autoscaled (meancentered and scaled data) of binned NMR spectral data set, including all the Thompson samples. According to the total variance (**Tv**), PC1 and PC2 explain 15% and 12% respectively. As it can be seen, only along PC1 is possible to observe a tendency to the separation of WT from GM (the Figure was prepared by using R program; see Appendix A)

The first consideration arise from the fact that only 27% of the total variance is described by the first two principal components, while up to 38% was obtained from the same number of PCs in the Silcora case. Such a result implies a lower descriptive power associated to the chosen PC plot which, however, can be raised by inspecting other dimensions of the PC space.

PC plot of Figure 3.8 allows us to assume that only PC1 has some tendency to discriminate between berries belonging to T-WT and T-GM1 genotypes, although such a tendency is not as clear as in the other cultivar.

The t-student test, indeed, gives results, reported in Table 3.5, with higher p-values (4.4e-5 vs. 2.2e-16 or 2.6e-12).

Also for the PC space calculated on the Thompson subset, the PC loadings give the weight of each bin in determining the extent of the separation along PC1 (Figure 3.9) and PC2 dimension (Figure 3.10).

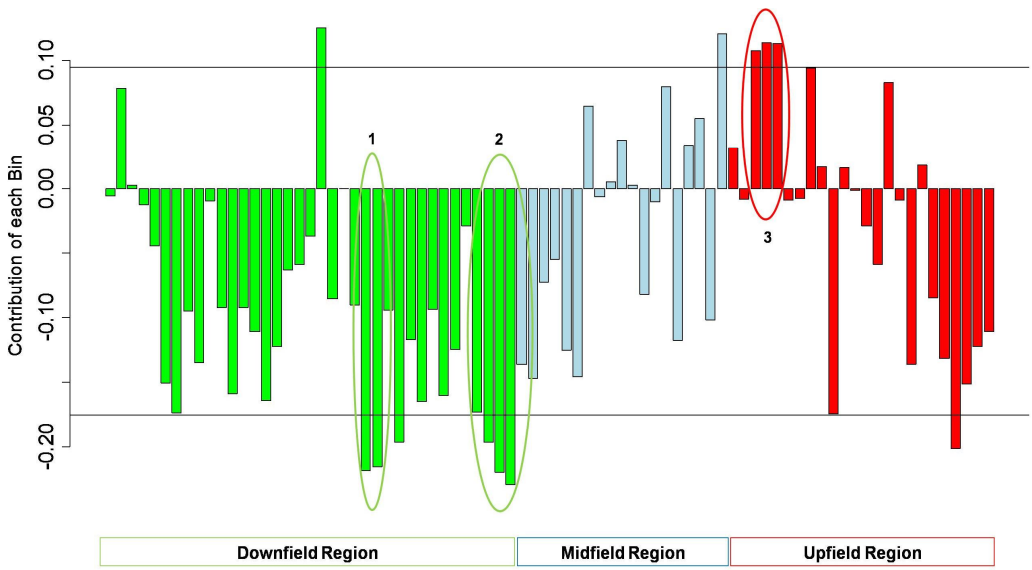


Figure 3.9: loading plots along PC1. Each bar corresponds to a spectral bin of 100 spectral points of length. Groups #1 and #2, located in the downfield region, and #3, in the upfield region, underlined with a colored circles, are responsible of samples separation along PC1.
Both y=a and y=-b lines indicate for convenience the maximum and minimum values to choose the most representative bins (the Figure was prepared by using R program; see Appendix A)

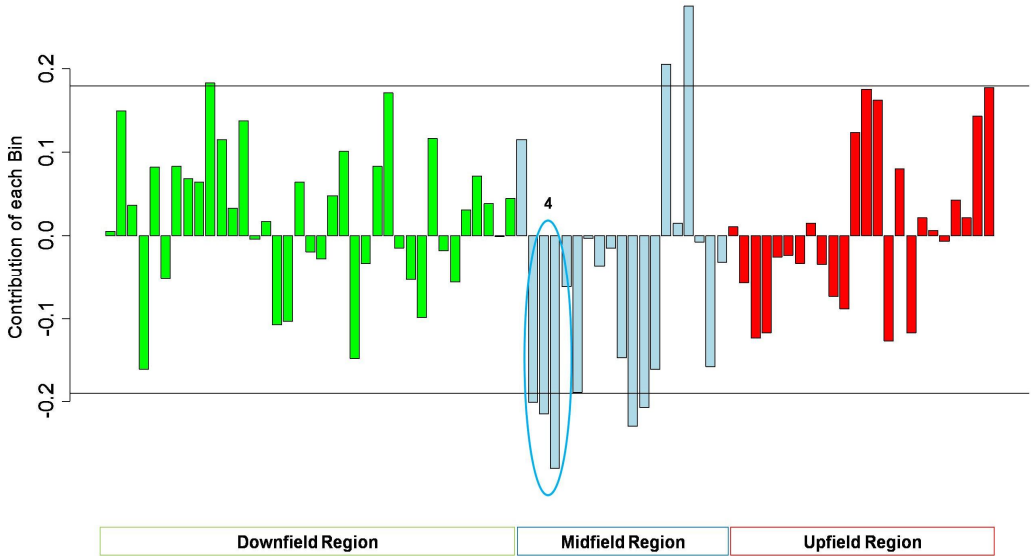


Figure 3.10: loading plots along PC2. Groups # 4, located in the midfield region, underlined with a blue circle, is responsible of samples' separation along PC2 (the Figure was prepared by using R program; see Appendix A)

In this way, bins' groups labelled with #1 and #2 (both in the aromatic region) are responsible of the even poor separation along PC1, so that a higher area is associated to the transgenic variant. This finding is also confirmed by integrating areas of such bins (Figures 3.11 A and B).

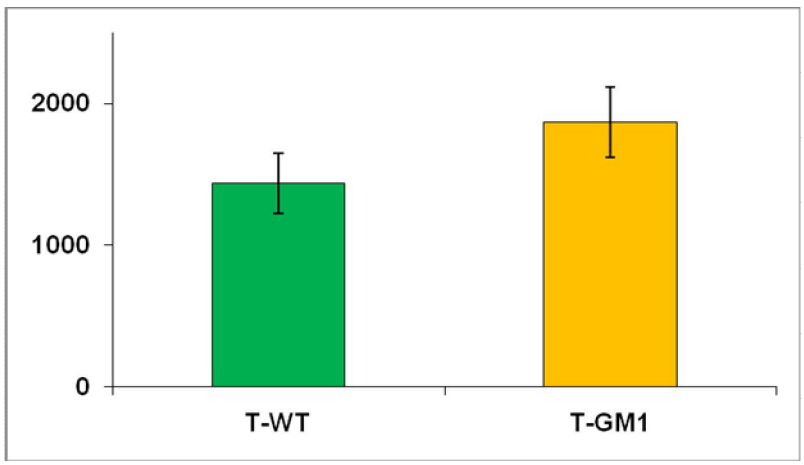


Figure 3.11 A: absolute area of the bins' groups labelled with # 1 in PC1 loadings plot. Bins are characterized by the presence of aromatic compounds' signals. Standard errors are also shown as black lines on the top of each bar

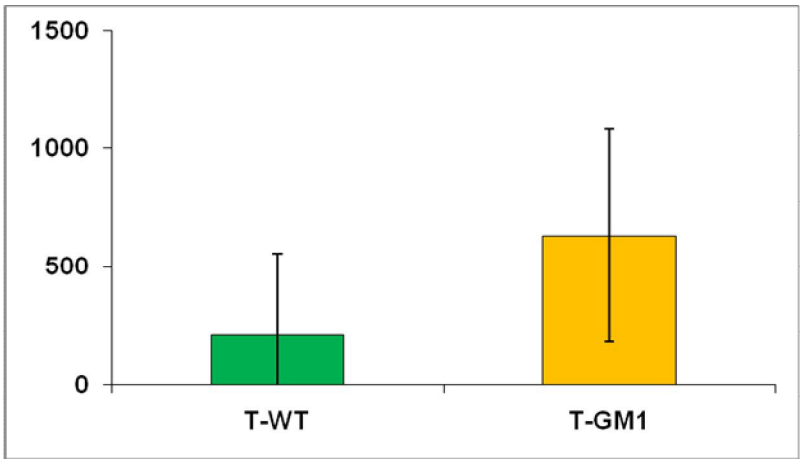


Figure 3.11 B: absolute area of the bins' groups labelled with # 2 in PC1 loadings plot. Like the previous bins' group, even these bins belong to the downfield region and thus to aromatic compounds. Standard errors are also shown as black lines on the top of each bar

Bins' group with label #3 exerts a slight decrease of its area, thus suggesting that the amount of organic acid is lower in the transgenic line than in its control group (Figure 3.11 C). On the contrary, although bins' group labelled with #4, containing sugar signals, has high weight on PC2, its cumulative area shows no meaningful difference when comparing transgenic and control lines (Figure 3.11 D). This is also expected on the basis of the fact that PC2 is, indeed, not able to discriminate between the two lines of the Thompson cultivar.

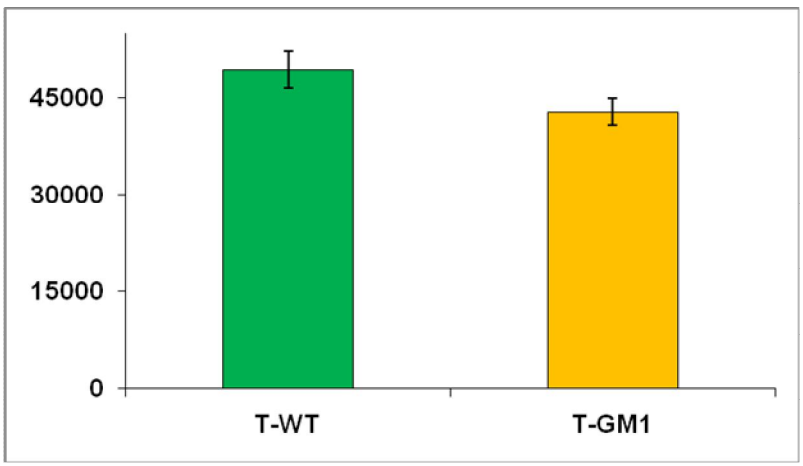


Figure 3.11 C: absolute area of the bins' groups labelled with # 3 in PC1 loadings plot. It corresponds to a group of bins in the downfield region, where organic acids' signals are abundant. Standard errors are also shown as black lines on the top of each bar

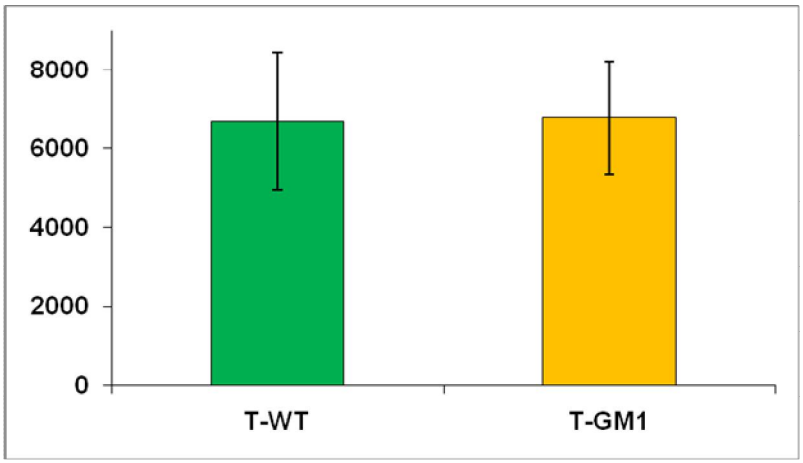


Figure 3.11 D: absolute area of the bins' groups labelled with #4 in PC2 loadings plot, falling into the midfield region. Standard errors are also shown as black lines on the top of each bar

3.3.6 PCA applied to all cultivars and genotypes

A global PCA, considering all samples belonging to both cultivars, is shown in Figure 3.12, together with the relative loadings shown in Figures 3.13 (PC1 loadings) and 3.14 (PC2 loadings).

This further analysis has not the purpose to explore the differences between cultivars but, rather, is aimed at determining whether the genetic modification shifts the metabolites profile towards the same direction of the PC plot, independently of the cultivar.

By looking at the PC plot, it appears that the first genetic modification occurring in the Silcora cultivar (S-GM1) shifts the metabolic profile along the PC1 direction in the same manner as it happens for the Thompson cultivar, although the latter at a less extent.

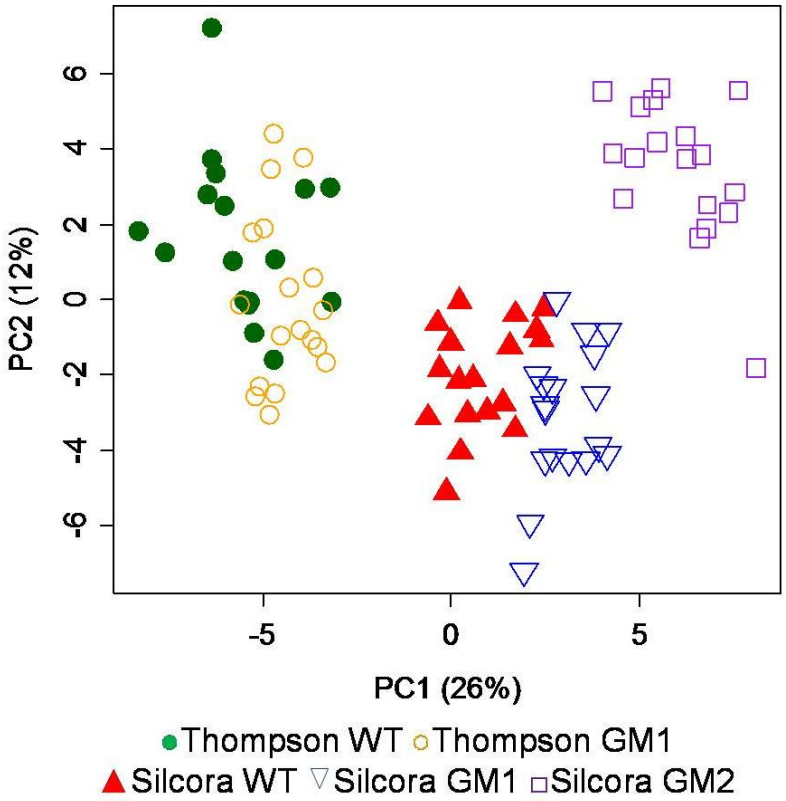


Figure 3.12: autoscaled PCA (meancentered and scaled data) of binned NMR spectral data set, including all samples. According to the total variance (**Tv**), PC1 and PC2 explain 26% and 12% respectively (the Figure was prepared by using R program; see Appendix A)

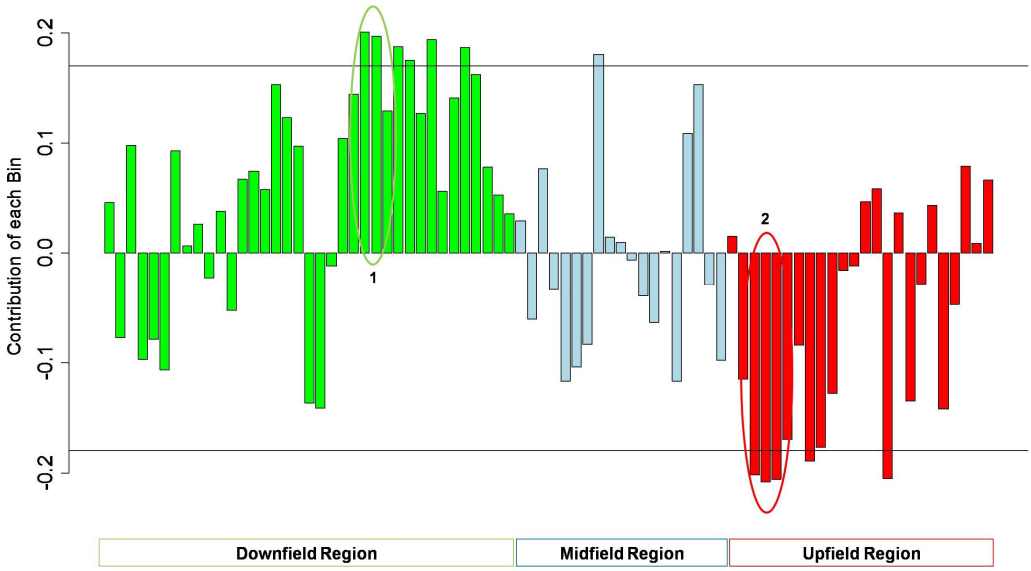


Figure 3.13: the loading plots for spectral bins along PC1 (the Figure was prepared by using R program; see Appendix A)

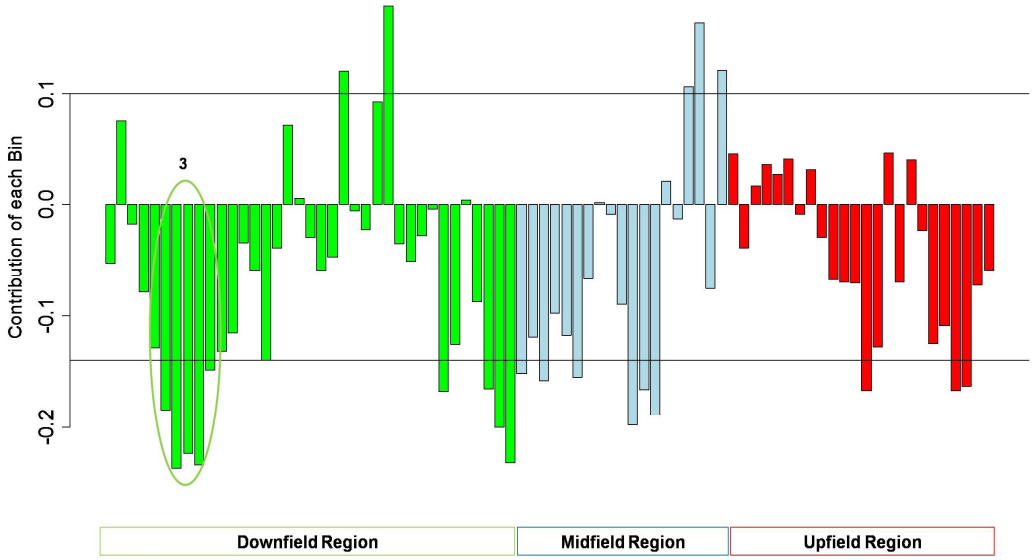


Figure 3.14: loading plots for spectral bins along PC2 (the Figure was prepared by using R program; see Appendix A)

Some further consideration can be extracted from the inspection of the integral areas of selected bins' groups.

Aromatic compounds of bins' group with #1 have higher concentrations in Silcora than in Thompson cultivars (Figure 3.15 A), and the GM variants contain higher amounts than the corresponding WT genotype, except for GM1 of the Thompson variety, for which integrals show comparable amounts.

The opposite trend appears for the organic acids collected in bins' group with label #2 (Figure 3.15 B), since Silcora berries contain lower concentrations and WT higher ones.

Again, the same trend was found in the chemical determination of total acids, although mostly due to tartaric acid variations rather than to changes in malic and citric acids concentration (E. Costantini et *al*., 2007).

It is interesting to note that Silcora GM2 have decreased amounts of the aromatic compounds grouped in bins labelled with #3 (Figure 3.15 C), whilst different aromatic compounds, grouped in bins #1, were subjected to an increase consequent to the same type of genetic mutation.

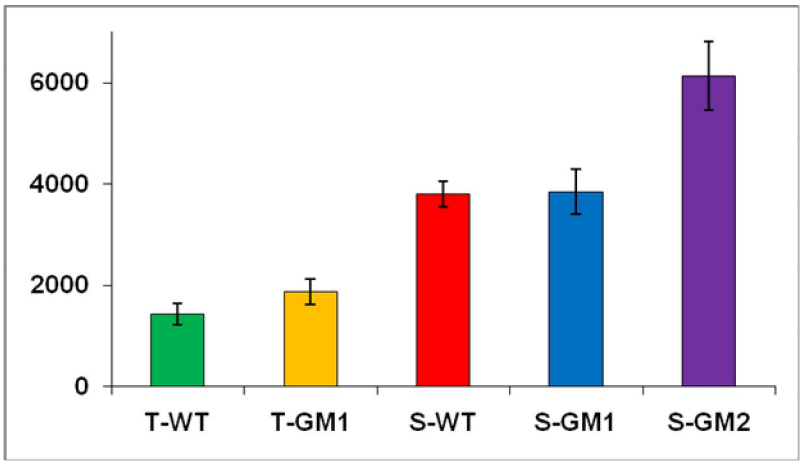


Figure 3.15 A: absolute area of the bins' groups labelled with #1 in PC1 loadings plot. It corresponds to bins from 24th to 25th. Standard errors are also shown as black lines on the top of each bar

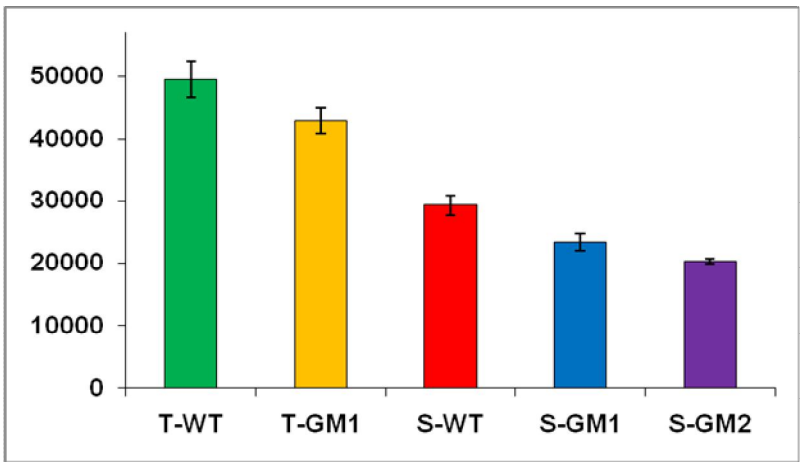


Figure 3.15 B: absolute area of the bins' groups labelled with #2 in PC1 loadings plot. It corresponds to bins from 60th to 62th. Standard errors are also shown as black lines on the top of each bar

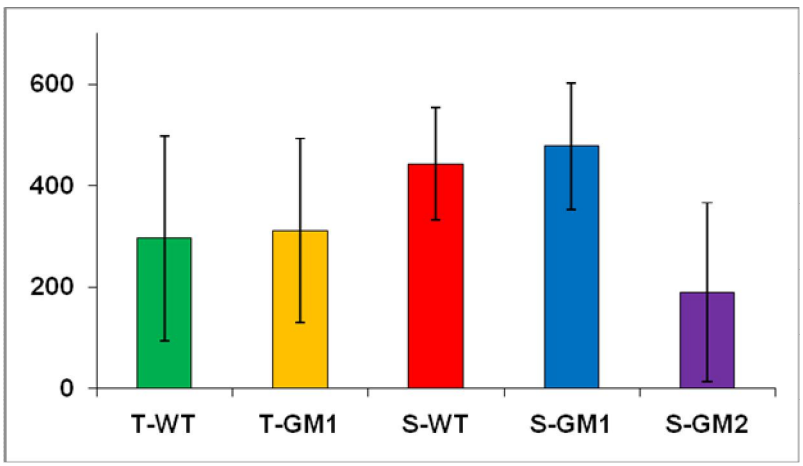


Figure 3.15 C: absolute area of the bins' groups labelled with #3 in PC2 loadings plot. It corresponds to bins from 6th to 10th. Standard errors are also shown as black lines on the top of each bar

According to the present results it is possible to identify some NMR spectral regions collecting signals from metabolites that show some statistically significant differences when comparing the genetically modified lines and their respective control samples. Such differences are mainly in the aromatic region, where tryptophan and its metabolites together with polyphenols have their signals. Since the inserted gene acts on the triptophane pathways, it would be plausible to assign some of the varying aromatic signals to such a molecule or its derivatives. However, meaningful variations are also observed in the organic acid regions. The latter have already been shown to change their concentration in a previous study based on chemical analysis. The critical role of auxin in plant growth and fruit development is a well known issue (A. W. Woodward et *al.*, 2005; B. Molesini et *al.*, 2008), and these results are evidencing the perspective offered by the NMR technology in identifying the entity of metabolic variation in specific tissue with modified auxin metabolism.

Further quantitative considerations need a more detailed study that must be based also on the assignment of those NMR signals which are mainly responsible for the global profile changes here observed.

3.3.7 LDA applied to all cultivars and genotypes

LD analysis (Chapter 1, par. 1.2.4) has been performed on PC Loadings obtained by PCA applied on all samples binned spectral data set. Before starting with the LD analysis, the one way ANOVA (J. Jionghua et *al.*, 2004) was applied on the first 14 PCs, together explaining 80% (Figure 3.16) of the total variance in order to find out those PCs having the highest discriminative power.

In this way, it has been possible to select the PCs to be used for the subsequent LD analysis in order to gain a better classification compared to the one already caught by Principle Components Analysis.

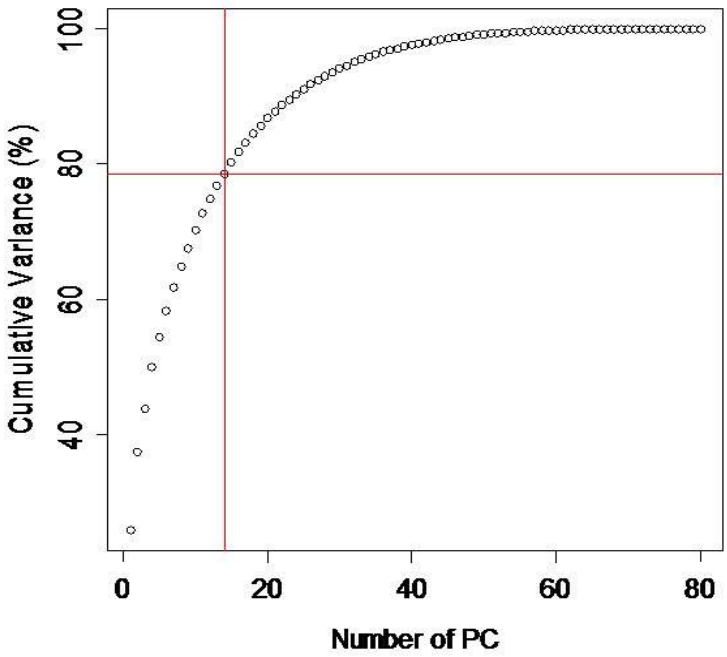


Figure 3.16: representation of the variation of TV according to the number of PC used. The first 14 PCs explain the 80% of the total variance (Figure was prepared by using R program; see Appendix A)

In Table 3.6 is listed the main results obtained from the ANOVA analysis performed used R.

Table 3.6: results from ANOVA analysis on the first 14 PCs (ANOVA was performed using R program; see Appendix A)

Number of PCs	p Value	Number of PCs	p Value
1	< 2.2e-16 ***	8	0.0008309 ***
2	< 2.2e-16 ***	9	1.704e-05 ***
3	0.1363	10	0.004832 **
4	4.05e-05 ***	11	0.1170
5	1.016e-06 ***	12	0.9194
6	0.0615 ^	13	0.9644
7	0.7329	14	0.3117

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '^' 0.1 ' ' 1

Figures 3.17 from A to D explain and illustrate the results of LDA applied on the 7 selected PCs (1, 2, 4, 5, 8, 9 and 10) resulting from the PCA of 90 spectra acquired on 45 grape samples (Thompson

and Silcora). The representation is by a box plot (J., Devore et *al.*, 1990; F. Daly et *al.*, 1995).

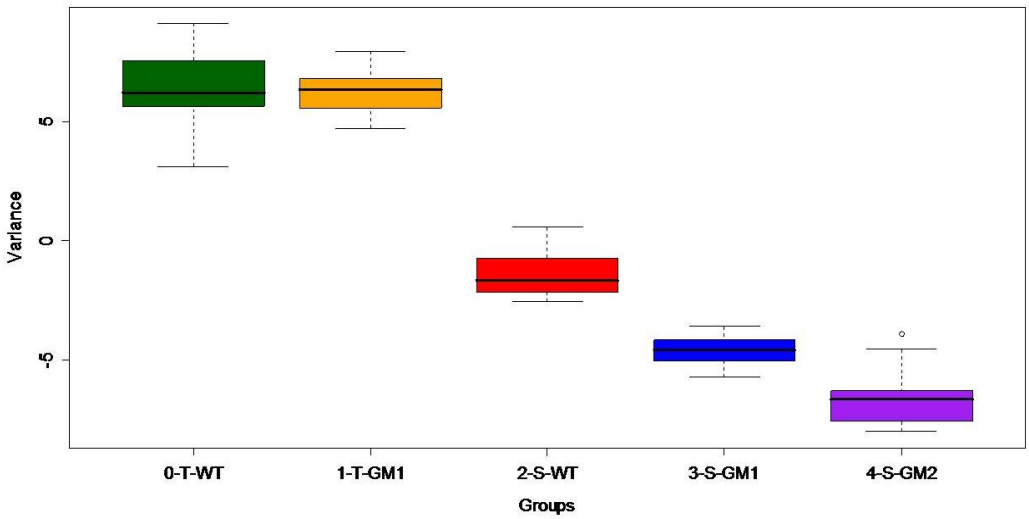


Figure 3.17 A: Figure shows the separation among groups according to LD1 values based on the seven PCs chosen from the analysis of variance (ANOVA). A boxplot, or box and whisker diagram, provides a simple graphical summary of a set of data. It shows a measure of central location (the median), two measures of dispersion (the range and inter-quartile range), the skewness (from the orientation of the median relative to the quartiles) and potential outliers (marked individually), in this case with a white dot. Samples are well grouped according to cultivars and related to Silcora, according to genotypes (the Figure was prepared by using R program; see appendix A)

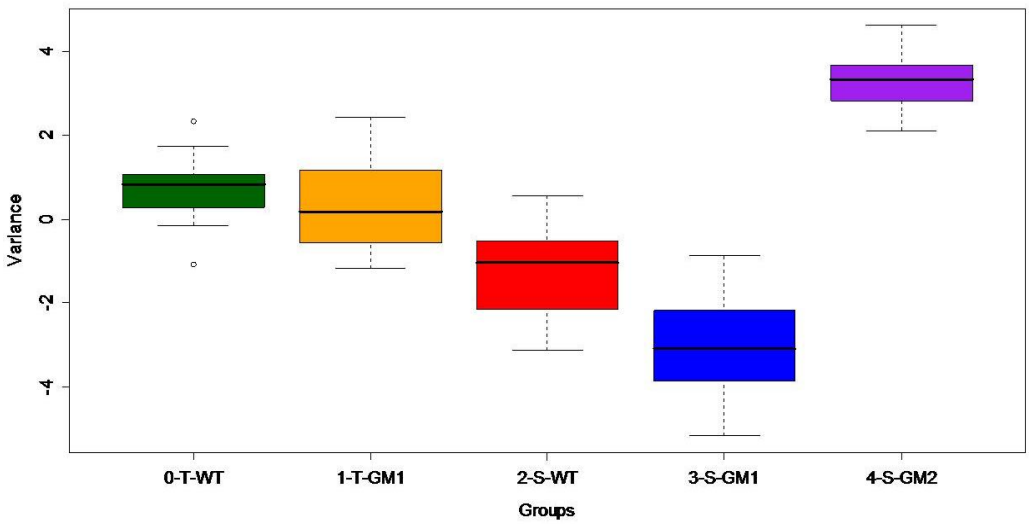


Figure 3.17 B: Figure shows the separation among groups according to LD2 values based on the seven PCs chosen from the analysis of variance (ANOVA). The Silcora GM2 shows a higher variance compared to the other groups (the Figure was prepared by using R program; see appendix A)

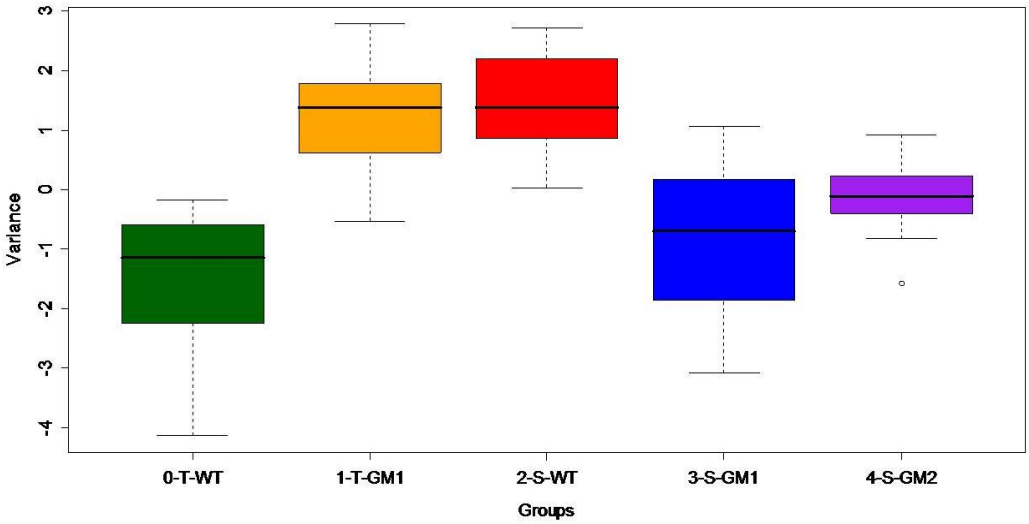


Figure 3.17 C: Figure shows the separation among groups according to LD3 values based on the seven PCs chosen from the analysis of variance (ANOVA) (the Figure was prepared by using R program; see appendix A)

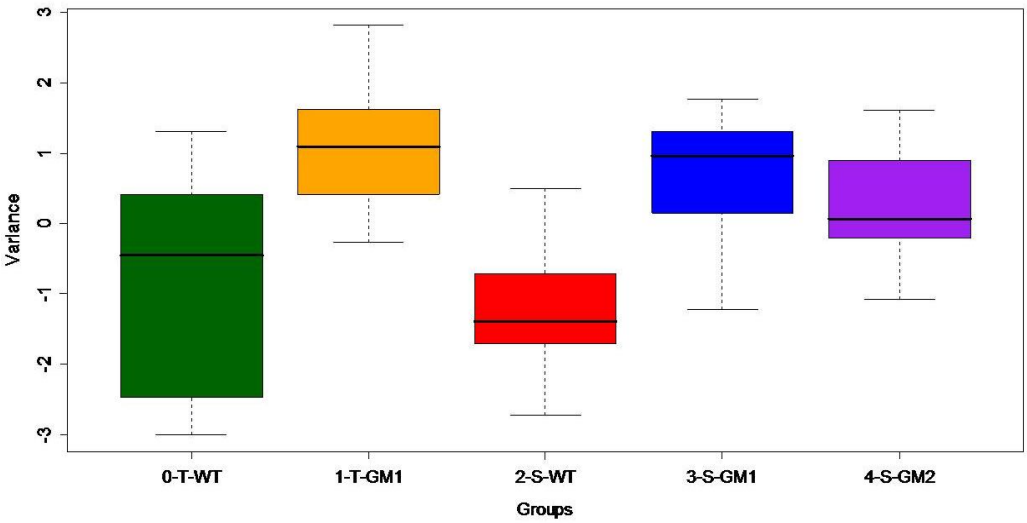


Figure 3.17 D: Figure shows the separation among groups according to LD4 values based on the seven PCs chosen from the analysis of variance (ANOVA) (the Figure was prepared by using R program; see Appendix A)

The examination of each LD and eigenvector loadings is useful to understand the basis of clustering behavior.

The combination of LD1 with LD2 better differentiates the two cultivars (Thompson from Silcora) and separates the wild type cultivars of Silcora from the other two genetically modified lines of the same cultivar. On the other variable of the plot, LD2 clearly allocates S-GM2 differently from the other two lines of the Silcora cultivar (Figure 3.18).

On the other side, the combination of LD3 with LD4 separates, in a slightly manner, T-WT from T-GM1 (Figure 3.19), especially along LD3.

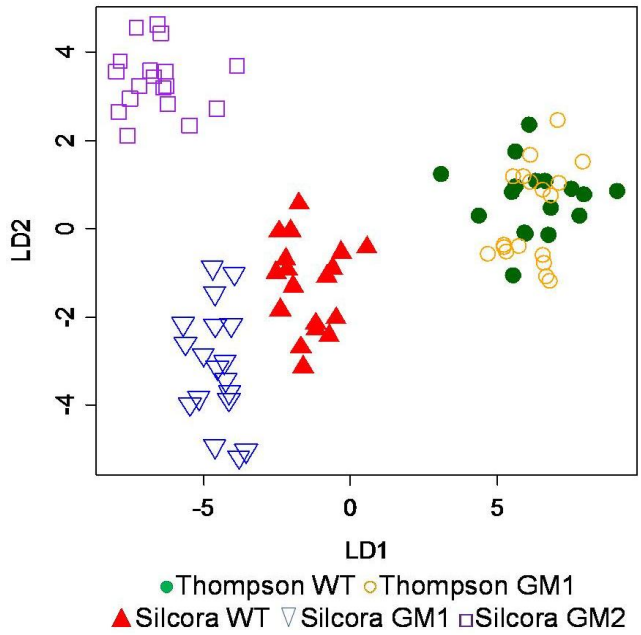


Figure 3.18: LD1 vs LD2 plot performed on PC Loadings obtained by PCA applied on all samples binned spectral data set (the Figure was prepared by using R program; see Appendix A)

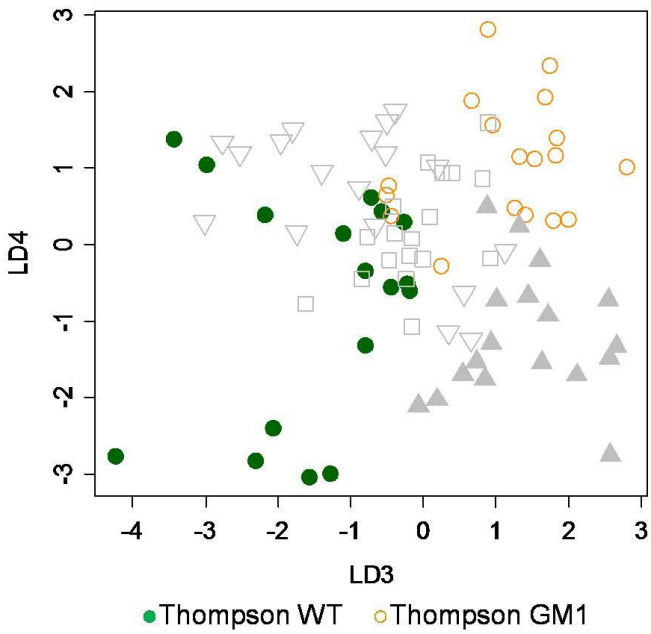


Figure 3.19: LD3 vs LD4 plot performed on PC Loadings obtained by PCA applied on all samples binned spectral data set. For a better visualization, Silcora samples were co-lored in grey (the Figure was prepared by using R program; see Appendix A)

Figure 3.20 shows the contribution of each spectral bin on LD, already weighted by their respective loading coefficient found in the seven selected PCs (Table 3.7) according to the Equation 3.1 below

$$LD1(i_1..i_{80}) = \sum_{n=1}^{7} vc_1 \cdot PC_n, ..., \sum_{n=1}^{7} vc_{80} \cdot PC_n \qquad \text{(Eq 3.1)}$$

in which *vc1* represents the canonical variable's value taken in consideration for bin #1 of LD1; *vc80* is, on the other side, the canonical variable's value of bin #80 always belonging to LD1; n is the number of PCs taken in consideration according to Table 3.6 (see Appendix A).

Table 3.7: coefficients of linear discriminant for each main PC selected on the base of ANOVA

Number of PCs	LD1	LD2	LD3	LD4
1	-1.19465559	0.05578327	0.02425739	0.02478813
2	0.09604888	0.65596033	-0.11150008	-0.01066208
4	0.17389633	0.11496348	-0.37502033	0.25615319
5	-0.16385861	-0.54232059	-0.30811770	0.28685386
8	-0.50760487	-0.38127295	-0.48677809	-0.49535749
9	-0.23438227	0.16122044	-0.42430681	0.09231256
10	-0.01937333	0.04063683	-0.27781272	-0.46917014

Such weighted contributions, here represented as a bar-plot, are useful to find out those bins that are the most responsible for the subdivision among categories, both cultivars and genotypes (as it has been also for PC analysis).

This representation allows to explain better the reasons of the separation between groups and between different genotypes on the basis of bins' content.

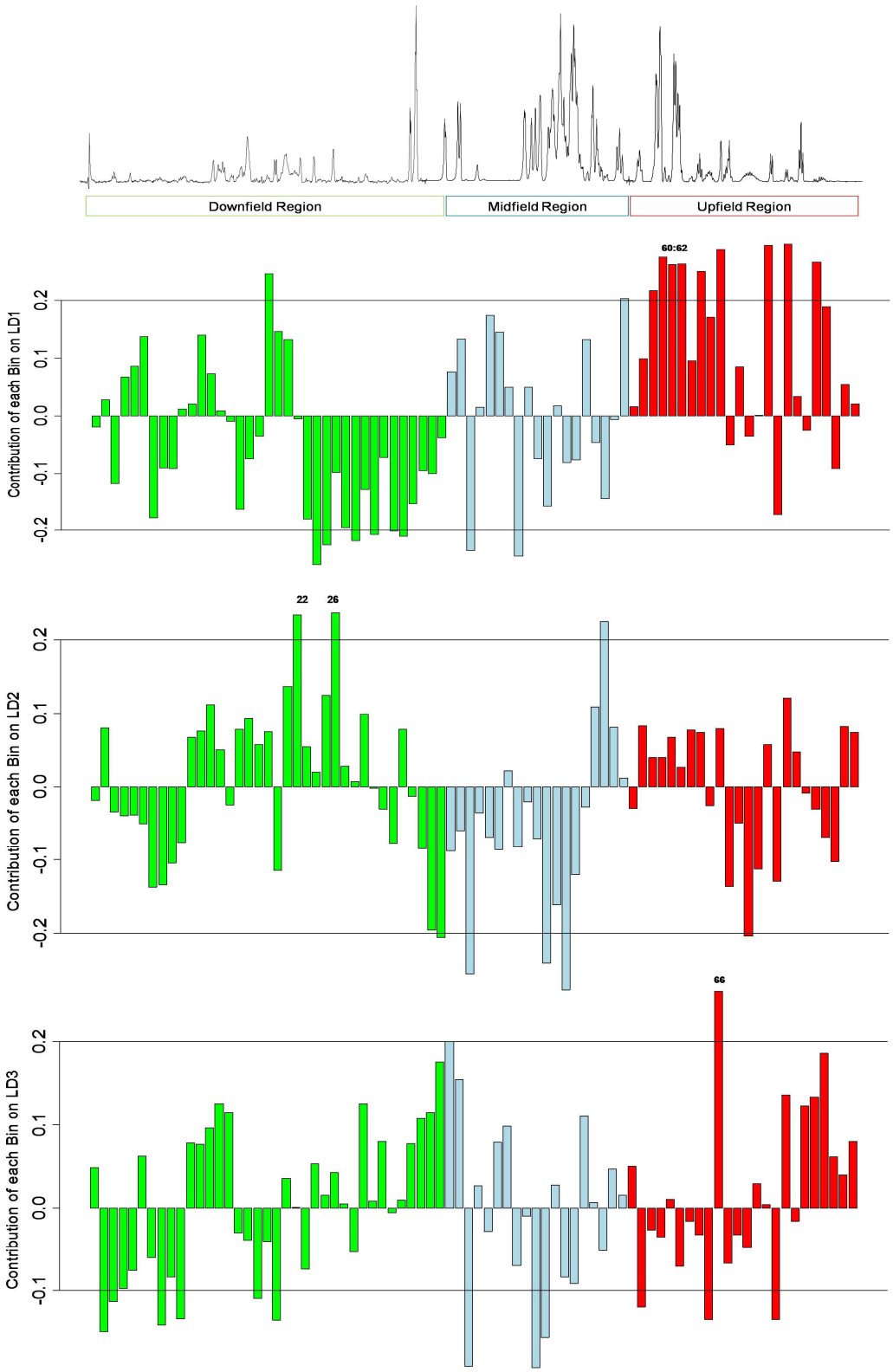


Figure 3.20 – first part: bar-plot representing the bins' weight on the cultivar/genotype separation along the four LD dimension. Each bar corresponds to a bin of 100 points of length

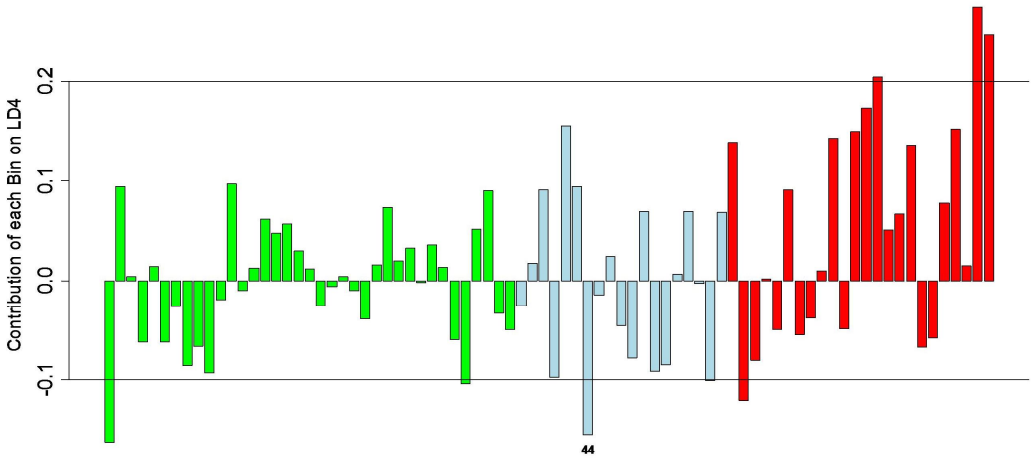


Figure 3.20 - second part: bar-plot representing the bins' weight on the cultivar/genotype separation along the four LD dimension. Each bar corresponds to a bin of 100 points of length

Combining the analysis of the loadings plot of LD1 and LD2 together with the plot of the loadings of LD3 and LD4 (Figure 3.20), it comes out that the most responsible bins involved above all in the separation between the two cultivars, and among genotypes, are bin 60th and bin 62nd together with bin 22nd and bin 26th, 66th and 44th.

It is found that the discriminant bins are spread over the whole spectral window. Negative contributions associated to a given bin means that a higher area for it push the LD towards negative values and vice versa. For example, the analysis of the portion of proton spectrum corresponding to region between positive bins 22nd and 26th (from 6.86 ppm to 6.27 ppm) in LD2 bin plot reveals an higher presence of the relative compounds (in this case, the range is characterized by the presence of aromatic compounds' signal such as polyphenols) for samples located along positive values of LD2.

So, the positive values, along LD2 in Figure 3.18, of S-GM2 means a more presence of these compounds, on the contrary it corresponds to a minor concentration of aromatic compounds for S-GM1 and S-WT. These bins points out the difference between cultivars, underlining also the difference among genotypes.

The same considerations can be done for the others bins; in particularly bin 60th and bin 62nd are more discriminant between cultivars: both of them show a larger integrals' intensity for those signals belonging to Thompson Seedless' samples.

Particularly interesting is the observation of the correspondent spectrum area. In fact, this area is characterized by the presence of signals belonging to organic acids such as malic and citric acid. Related to berry quality and nutritional values, malic and citric acids are two of the most important organic acids presented together with succinic and tartaric acids. The first two, on the basis of our observations, are more discriminant between cultivars then the others. Then, Thompson Seedless results to be the cultivar with the most percentage of total organic acids (malate + citrate) as it can be seen in Figure 3.21. Moreover, the two acids are presented in berry in different concentration. From the spectrum analysis, it emerges that the rate between malate and citrate is 1:10 that means that the concentration of malate is exactly about 10x the concentration of citrate (Figure 3.22).

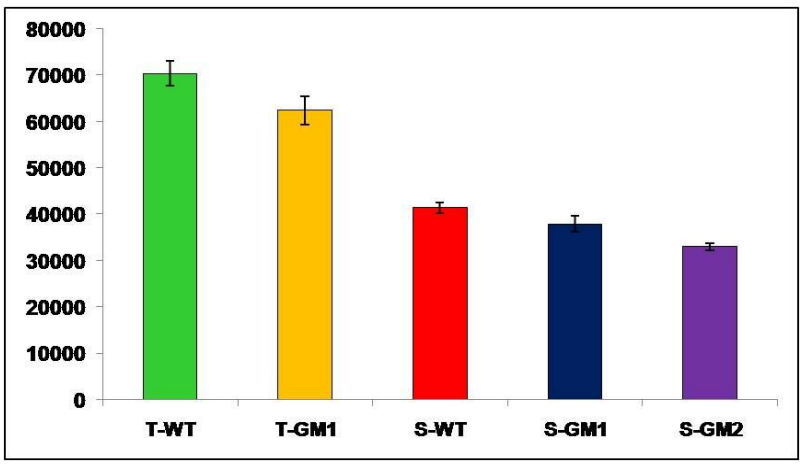


Figure 3.21: total organic acids (malic and citric acids) distribution among genotypes and between cultivars. Thompson seedless shows an higher concentration of these acids resulting most acid then Silcora cultivar. Standard errors are also shown as black lines on the top of each bar

In the end, considering the total organic acids S-GM1 is lightly less acid than S-WT, but more than S-GM2. At the end, the analysis of bins plot along LD3 and LD4 points out bins 66th for the

first one and 44th for the second. The importance of these bins are exclusively for the distinction from T-GM1 to T-WT.

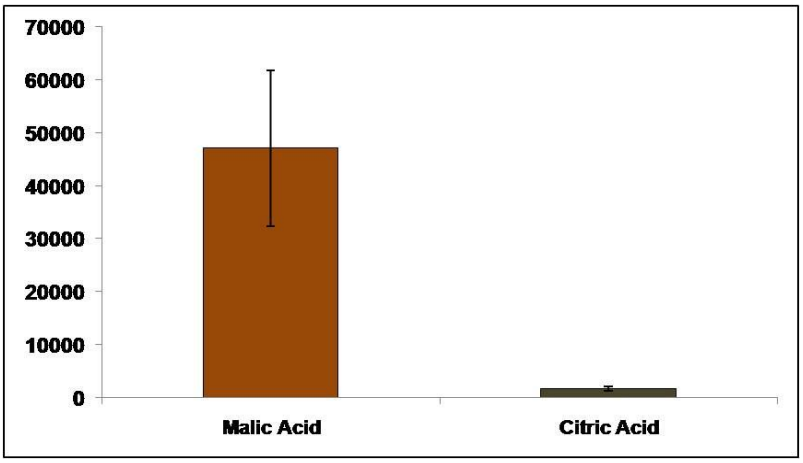


Figure 3.22: comparison between Malic and Citric acids. Malic's ratio is about 10 times Citric's one. Standard errors are also shown as black lines on the top of each bar

3.4 CONCLUSION

The chemometric analysis, based on the metabonomic approach, shows that a grape cultivar, more than one other, exert some statistically significant differences of their metabolites content, above the natural variability, induced by modifications introduced by genome's artificial alteration.

According to the genetic modification, the results were supposed to be relative mainly to metabolites involved in the ripening process. On the contrary, a global screening of the metabolites' profile shows changes not immediately expected as the gene modification consequences.

The combination of NMR spectroscopy together with the chemometric technique for data analysis becomes an important tool able to identify unexpected differences in the metabolic profile. Of course, NMR spectroscopy together with chemometrics, should not be taken alone as analytical tools for the description of substantial equivalence between GMO and not-GMO foodstuff. However, applications such as the one explored in the present research work,

may contribute to the definition of a more complete description of the food content. In the light of such considerations, the metabonomic approach applied to food science may result as a further tool in the hands of the scientists that need more and more food descriptors to be taken under control during the risk assessment strategy. Moreover, the presented approach may be used to define, together with other statistical and analytical tools, the point at which a GM-food is no longer substantially equivalent to the one that has not been modified.

Further developments of the present study should be addressed towards the characterization of the chemical nature of metabolites undergoing major concentration changes, in order to rationalize and, possibly, to understand the main metabolic pathways being influenced by the genetic modifications.

REFERENCES

A. W., Woodward and B., Bartel (2005). *Auxin: regulation, action, and interaction*. Ann Bot, 95, 707-35

A., Craig, O., Cloarec, E., Holme, J. K., Nicholson and J. C., Lindon (2006). *Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets*. Anal Chem, 78, 2262-2267

Anon (2000). *The EU-US Biotechnology Consultative Forum: Final Report*, 10, recommendation 5

B., Mezzetti, L., Landi, T., Pandolfini and A., Spena (2004). *The DefH9-iaaM auxin-synthesizing gene increases plant fecundity and fruit production in strawberry and raspberry*. BMC Biotechnology, 4, 4

B., Mezzetti, T., Pandolfini, O., Navacchi and L., Landi (2002). *Genetic transformation of Vitis vinifera via organogenesis.* BMC Biotechnology, 2, 18

B., Molesini, T., Pandolfini, G. L., Rotino, V., Dani and A., Spena (2008). *Aucsia Genes Silencing Causes Parthenocarpic Fruit Development in Tomato*. Plant Physiol Nov 5 [Epub ahead of print]

C. Castro, C., Manetti (2007). *A multiway approach to analyze metabonomic data: a study of maize seeds development*. Analytical Biochemistry, 371, 194-200

C. l., Gavaghan, I. D., Wilson and J. K., Nicholson (2002). *Physiological variation in metabolic phenotyping and functional genomic studies: use of orthogonal signal correction and PLS-DA*. FEBS Letters, 530, 191-196

C., Manetti, C., Bianchetti, M., Bizzarri, L., Casciani, C., Castro, G., D'Ascenzo, M., Delfini, M. E., Di Cocco, A., Laganà, A., Miccheli, M., Motto and F., Conti (2004). *NMR-based metabonomic study of transgenic maize.* Phytochemistry, 65, 23187-3198

E., Costantini, L. Landi, O., Silvestroni, T., Pandolfini, A., Spena and B., Mezzetti (2007). *Auxin synthesis-encoding Transgene Enhaches fecundity*. Plant Physiol, 143, 1689-94

E., Millstone, E., Brunner and S., Mayer (1999). *Beyond 'Substantial Equivalence".* Nature 158, 401:525-526

F., Daly, D. J., Hand, M. C., Jones, A. D., Lunn and K. J., McConway (1995). Elements of Statistics. Addison Wesley / The Open University

G. L., Rotino, E., Perri, M., Zottini, H., Sommer and A., Spena (1997). *Genetic engineering of parthenocarpic plants*. Nature Biotechnology, 15, 1398-1401

H. A., Kuiper, E. J., Kok and K. H., Engel (2003). *Exploitation of molecular profiling techniques for GM food safety assessment*. Curr. Opin. Biotechnol, 14, 238-243

H. C., Keun (2006). *Metabonomic modeling of drug toxicity*. Pharmacology & Therapeutics, 109, 92-106

H., Wang, B., Jones, Z., Li, P., Frasse, C., Delalande, F., Regad, S., Chaabouni, A., Latche, J. C., Pech JC and M., Bouzayen (2005). *The tomato Aux/IAA trascrition factor IAA9 is involved in fruit development and leaf morphogenesis*. Plant Cell, 17, 2676-2692

I. B., Abdel-Farid, K. H., Kyong, Y. H., Choi, R., Verpoorte (2007). *Metabolic Characterization of Brassica rapa Leaves by NMR Spectroscopy*. J Agric Food Chem, 55, 7936-7943

J., Devore and R., Peck (1990). Introductory Statistics. West Publishing Co

J., Jionghua and G., Huairui (2004). *ANOVA Method for Variance Component Decomposition and Diagnosis in Batch Manufacturing Processes.* International Journal of Flexible Manufacturing Systems, 15, 167-186

J. K., Nicholson, J. C., Lindon and E., Holmes (1999). *"Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data*. Xenobiotica, 11, 1181-1189

K. S., Solanky, N. J., Bailey, B. M., Beckwith-Hall, S., Bingham, A., Davis, E., Holmes, J. K., Nicholson and A., Cassidy (2005). *Biofluid ^{1}H NMR-based metabonomic techniques in nutrition research - metabolic effects of dietary isoflavones in humans.* The Journal of Nutritional Biochemistry, 16, 236-244

K. V., Mardia, J. T., Kent and J. M., Bibby (1979). *Multivariate analysis*. Academic Press. London.

L., Levidow, J., Murphy and S., Carr S (2007). *Recasting "Substantial Equivalence": Transatlantic Governance of GM Food*. Science, Technology & Human, 32, 26-64

M. C., Ruebelt, N. K., Leimgruber, M., Lipp, T. L., Reynolds, M. A., Nemeth, J. D., Astwood, K. H., Engel and K. D., Jany (2006). *Application of two-dimensional gel electrophoresis to interrogate alterations in the proteome of genetically modified crops. 1. Assessing analytical validation.* J Agric Food Chem, 54, 2154-2161

N., Acciarri, F., Restaino, G., Vitelli, D., Perrone, M., Zottini, T., Pandolfini, A., Spena, G. L., Rotino (2002). *Genetically modified parthenocarpic eggplants: improved fruit productivity under both greenhouse and open field cultivation*. BMC Biotechnology, 2, 4

N., Carmi, Y., Salts, B., Dedicova, S., Shabtai, R., Barg (2003). *Induction of parthenocarpy in tomato via specific expression of the rolB gene in the ovary*. Planta, 217, 726-735

N., Ficcadenti, S., Sestili, T., Pandolfini, C., Cirillo, G. L., Rotino and A., Spena (1999). *Genetic engineering of parthenocarpic fruit development in tomato*. Molecular Breeding, 5, 463-470

P., Hedden and A. L., Phillips (2000). *Manipulation of hormone biosynthetic genes in transgenic plants.* Current Opinion in Biotechnology 11, 130-137

R., Batista, N., Saibo, T., Lourenço, M. M., Oliveira (2008). *Microarray analyses reveal that plant mutagenesis may induce more transcriptomic changes than transgene insertion.* Proc Natl Acad Sci U S A, 105, 3640-3645

T., Pandolfini, G. L., Rotino, C., Camerini,R., Defez and A., Spena (2002). *Optimization of transgene action at the post-transcriptional level: high quality parthenocarpic fruits in industrial tomatoes*. BMC Biotechnology, 2, 1

World Health Organization (2000). *Safety aspects of genetically modified foods of plant origin.* Report of a Joint FAO/WHO Expert Consultation on Foods Derived from Biotechnology

Z., Xiaoyu, W., Huifeng, L., Peiqiu, L., Xiaojing, N., Jiazuan and P., Fengkui (2006). *NMR-based metabonomic study on the subacute toxicity of aristolochic acid in rats*. Food and Chemical Toxicology, 44, 1006-1014

Z., Yin, R., Malinowski, A., Ziolkowska, H., Sommer, W., Plader and S., Malepszi (2006). *The DefH9-iaaM-containing construct efficiently induces parthenocarpy in cucumber*. Cell Mol Biol, 11, 279-290

APPENDIX A AND B

Appendix A

```
############################################################

rm(list=ls(all=TRUE))

#####################################################################

# Thompson Wild Type Samples

G19<-scan("1TW.txt", dec=".")
G20<-scan("1TWbis.txt", dec=".")
G21<-scan("2TW.txt", dec=".")
G22<-scan("2TWbis.txt", dec=".")
G23<-scan("3TW.txt", dec=".")
G24<-scan("3TWbis.txt", dec=".")
G25<-scan("4TW.txt", dec=".")
G26<-scan("4TWbis.txt", dec=".")
G27<-scan("5TW.txt", dec=".")
G28<-scan("5TWbis.txt", dec=".")
G29<-scan("6TW.txt", dec=".")
G30<-scan("6TWbis.txt", dec=".")
G31<-scan("7TW.txt", dec=".")
G32<-scan("7TWbis.txt", dec=".")
G33<-scan("8TW.txt", dec=".")
G34<-scan("8TWbis.txt", dec=".")
G35<-scan("9TW.txt", dec=".")
G36<-scan("9TWbis.txt", dec=".")

# Thompson Genetically Modified 1 Samples

G37<-scan("10TG.txt", dec=".")
G38<-scan("10TGbis.txt", dec=".")
G39<-scan("11TG.txt", dec=".")
G40<-scan("11TGbis.txt", dec=".")
G41<-scan("12TG.txt", dec=".")
G42<-scan("12TGbis.txt", dec=".")
G43<-scan("13TG.txt", dec=".")
G44<-scan("13TGbis.txt", dec=".")
G45<-scan("14TG.txt", dec=".")
G46<-scan("14TGbis.txt", dec=".")
G47<-scan("15TG.txt", dec=".")
G48<-scan("15TGbis.txt", dec=".")
G49<-scan("16TG.txt", dec=".")
G50<-scan("16TGbis.txt", dec=".")
G51<-scan("17TG.txt", dec=".")
G52<-scan("17TGbis.txt", dec=".")
G53<-scan("18TG.txt", dec=".")
G54<-scan("18TGbis.txt", dec=".")

# Silcora wild Type Samples

G55<-scan("1SW.txt", dec=".")
G56<-scan("1SWbis.txt", dec=".")
G57<-scan("2SW.txt", dec=".")
G58<-scan("2SWbis.txt", dec=".")
```

```
G59<-scan("3SW.txt", dec=".")
G60<-scan("3SWbis.txt", dec=".")
G61<-scan("4SW.txt", dec=".")
G62<-scan("4SWbis.txt", dec=".")
G63<-scan("5SW.txt", dec=".")
G64<-scan("5SWbis.txt", dec=".")
G65<-scan("6SW.txt", dec=".")
G66<-scan("6SWbis.txt", dec=".")
G67<-scan("7SW.txt", dec=".")
G68<-scan("7SWbis.txt", dec=".")
G69<-scan("8SW.txt", dec=".")
G70<-scan("8SWbis.txt", dec=".")
G71<-scan("9SW.txt", dec=".")
G72<-scan("9SWbis.txt", dec=".")

# Silcora Genetically Modified 1 Samples

G73<-scan("10SG1.txt", dec=".")
G74<-scan("10SG1bis.txt", dec=".")
G75<-scan("11SG1.txt", dec=".")
G76<-scan("11SG1bis.txt", dec=".")
G77<-scan("12SG1.txt", dec=".")
G78<-scan("12SG1bis.txt", dec=".")
G79<-scan("13SG1.txt", dec=".")
G80<-scan("13SG1bis.txt", dec=".")
G81<-scan("14SG1.txt", dec=".")
G82<-scan("14SG1bis.txt", dec=".")
G83<-scan("15SG1.txt", dec=".")
G84<-scan("15SG1bis.txt", dec=".")
G85<-scan("16SG1.txt", dec=".")
G86<-scan("16SG1bis.txt", dec=".")
G87<-scan("17SG1.txt", dec=".")
G88<-scan("17SG1bis.txt", dec=".")
G89<-scan("18SG1.txt", dec=".")
G90<-scan("18SG1bis.txt", dec=".")

# Silcora Genetically Modified 2 Samples

G91<-scan("19SG2.txt", dec=".")
G92<-scan("19SG2bis.txt", dec=".")
G93<-scan("20SG2.txt", dec=".")
G94<-scan("20SG2bis.txt", dec=".")
G95<-scan("21SG2.txt", dec=".")
G96<-scan("21SG2bis.txt", dec=".")
G97<-scan("22SG2.txt", dec=".")
G98<-scan("22SG2bis.txt", dec=".")
G99<-scan("23SG2.txt", dec=".")
G100<-scan("23SG2bis.txt", dec=".")
G101<-scan("24SG2.txt", dec=".")
G102<-scan("24SG2bis.txt", dec=".")
G103<-scan("25SG2.txt", dec=".")
G104<-scan("25SG2bis.txt", dec=".")
G105<-scan("26SG2.txt", dec=".")
G106<-scan("26SG2bis.txt", dec=".")
G107<-scan("27SG2.txt", dec=".")
G108<-scan("27SG2bis.txt", dec=".")

##################################################
# Delete Water, Methanol and Acetic Acid Signals #
##################################################

G19<-G19[c(1:6200,6700:7970,8100:9200,9450:12295)]
G20<-G20[c(1:6200,6700:7970,8100:9200,9450:12295)]
G21<-G21[c(1:6200,6700:7970,8100:9200,9450:12295)]
G22<-G22[c(1:6200,6700:7970,8100:9200,9450:12295)]
G23<-G23[c(1:6200,6700:7970,8100:9200,9450:12295)]
G24<-G24[c(1:6200,6700:7970,8100:9200,9450:12295)]
G25<-G25[c(1:6200,6700:7970,8100:9200,9450:12295)]
```

```
G26<-G26[c(1:6200,6700:7970,8100:9200,9450:12295)]
G27<-G27[c(1:6200,6700:7970,8100:9200,9450:12295)]
G28<-G28[c(1:6200,6700:7970,8100:9200,9450:12295)]
G29<-G29[c(1:6200,6700:7970,8100:9200,9450:12295)]
G30<-G30[c(1:6200,6700:7970,8100:9200,9450:12295)]
G31<-G32[c(1:6200,6700:7970,8100:9200,9450:12295)]
G32<-G32[c(1:6200,6700:7970,8100:9200,9450:12295)]
G33<-G33[c(1:6200,6700:7970,8100:9200,9450:12295)]
G34<-G34[c(1:6200,6700:7970,8100:9200,9450:12295)]
G35<-G35[c(1:6200,6700:7970,8100:9200,9450:12295)]
G36<-G36[c(1:6200,6700:7970,8100:9200,9450:12295)]
G37<-G37[c(1:6200,6700:7970,8100:9200,9450:12295)]
G38<-G38[c(1:6200,6700:7970,8100:9200,9450:12295)]
G39<-G39[c(1:6200,6700:7970,8100:9200,9450:12295)]
G40<-G40[c(1:6200,6700:7970,8100:9200,9450:12295)]
G41<-G41[c(1:6200,6700:7970,8100:9200,9450:12295)]
G42<-G42[c(1:6200,6700:7970,8100:9200,9450:12295)]
G43<-G43[c(1:6200,6700:7970,8100:9200,9450:12295)]
G44<-G44[c(1:6200,6700:7970,8100:9200,9450:12295)]
G45<-G45[c(1:6200,6700:7970,8100:9200,9450:12295)]
G46<-G46[c(1:6200,6700:7970,8100:9200,9450:12295)]
G47<-G47[c(1:6200,6700:7970,8100:9200,9450:12295)]
G48<-G48[c(1:6200,6700:7970,8100:9200,9450:12295)]
G49<-G49[c(1:6200,6700:7970,8100:9200,9450:12295)]
G50<-G50[c(1:6200,6700:7970,8100:9200,9450:12295)]
G51<-G51[c(1:6200,6700:7970,8100:9200,9450:12295)]
G52<-G52[c(1:6200,6700:7970,8100:9200,9450:12295)]
G53<-G53[c(1:6200,6700:7970,8100:9200,9450:12295)]
G54<-G54[c(1:6200,6700:7970,8100:9200,9450:12295)]
G55<-G55[c(1:6200,6700:7970,8100:9200,9450:12295)]
G56<-G56[c(1:6200,6700:7970,8100:9200,9450:12295)]
G57<-G57[c(1:6200,6700:7970,8100:9200,9450:12295)]
G58<-G58[c(1:6200,6700:7970,8100:9200,9450:12295)]
G59<-G59[c(1:6200,6700:7970,8100:9200,9450:12295)]
G60<-G60[c(1:6200,6700:7970,8100:9200,9450:12295)]
G61<-G61[c(1:6200,6700:7970,8100:9200,9450:12295)]
G62<-G62[c(1:6200,6700:7970,8100:9200,9450:12295)]
G63<-G63[c(1:6200,6700:7970,8100:9200,9450:12295)]
G64<-G64[c(1:6200,6700:7970,8100:9200,9450:12295)]
G65<-G65[c(1:6200,6700:7970,8100:9200,9450:12295)]
G66<-G66[c(1:6200,6700:7970,8100:9200,9450:12295)]
G67<-G67[c(1:6200,6700:7970,8100:9200,9450:12295)]
G68<-G68[c(1:6200,6700:7970,8100:9200,9450:12295)]
G69<-G69[c(1:6200,6700:7970,8100:9200,9450:12295)]
G70<-G70[c(1:6200,6700:7970,8100:9200,9450:12295)]
G71<-G71[c(1:6200,6700:7970,8100:9200,9450:12295)]
G72<-G72[c(1:6200,6700:7970,8100:9200,9450:12295)]
G73<-G73[c(1:6200,6700:7970,8100:9200,9450:12295)]
G74<-G74[c(1:6200,6700:7970,8100:9200,9450:12295)]
G75<-G75[c(1:6200,6700:7970,8100:9200,9450:12295)]
G76<-G76[c(1:6200,6700:7970,8100:9200,9450:12295)]
G77<-G77[c(1:6200,6700:7970,8100:9200,9450:12295)]
G78<-G78[c(1:6200,6700:7970,8100:9200,9450:12295)]
G79<-G79[c(1:6200,6700:7970,8100:9200,9450:12295)]
G80<-G80[c(1:6200,6700:7970,8100:9200,9450:12295)]
G81<-G81[c(1:6200,6700:7970,8100:9200,9450:12295)]
G82<-G82[c(1:6200,6700:7970,8100:9200,9450:12295)]
G83<-G83[c(1:6200,6700:7970,8100:9200,9450:12295)]
G84<-G84[c(1:6200,6700:7970,8100:9200,9450:12295)]
G85<-G85[c(1:6200,6700:7970,8100:9200,9450:12295)]
G86<-G86[c(1:6200,6700:7970,8100:9200,9450:12295)]
G87<-G87[c(1:6200,6700:7970,8100:9200,9450:12295)]
G88<-G88[c(1:6200,6700:7970,8100:9200,9450:12295)]
G89<-G89[c(1:6200,6700:7970,8100:9200,9450:12295)]
G90<-G90[c(1:6200,6700:7970,8100:9200,9450:12295)]
G91<-G91[c(1:6200,6700:7970,8100:9200,9450:12295)]
G92<-G92[c(1:6200,6700:7970,8100:9200,9450:12295)]
G93<-G93[c(1:6200,6700:7970,8100:9200,9450:12295)]
```

```
G94<-G94[c(1:6200,6700:7970,8100:9200,9450:12295)]
G95<-G95[c(1:6200,6700:7970,8100:9200,9450:12295)]
G96<-G96[c(1:6200,6700:7970,8100:9200,9450:12295)]
G97<-G97[c(1:6200,6700:7970,8100:9200,9450:12295)]
G98<-G98[c(1:6200,6700:7970,8100:9200,9450:12295)]
G99<-G99[c(1:6200,6700:7970,8100:9200,9450:12295)]
G100<-G100[c(1:6200,6700:7970,8100:9200,9450:12295)]
G101<-G101[c(1:6200,6700:7970,8100:9200,9450:12295)]
G102<-G102[c(1:6200,6700:7970,8100:9200,9450:12295)]
G103<-G103[c(1:6200,6700:7970,8100:9200,9450:12295)]
G104<-G104[c(1:6200,6700:7970,8100:9200,9450:12295)]
G105<-G105[c(1:6200,6700:7970,8100:9200,9450:12295)]
G106<-G106[c(1:6200,6700:7970,8100:9200,9450:12295)]
G107<-G107[c(1:6200,6700:7970,8100:9200,9450:12295)]
G108<-G108[c(1:6200,6700:7970,8100:9200,9450:12295)]

########################################
# Delete Points from the Spectrum Tails#
########################################

g19<-G19[c(2001:10000)]
g20<-G20[c(2001:10000)]
g21<-G21[c(2001:10000)]
g22<-G22[c(2001:10000)]
g23<-G23[c(2001:10000)]
g24<-G24[c(2001:10000)]
g25<-G25[c(2001:10000)]
g26<-G26[c(2001:10000)]
g27<-G27[c(2001:10000)]
g28<-G28[c(2001:10000)]
g29<-G29[c(2001:10000)]
g30<-G30[c(2001:10000)]
g31<-G31[c(2001:10000)]
g32<-G32[c(2001:10000)]
g33<-G33[c(2001:10000)]
g34<-G34[c(2001:10000)]
g35<-G35[c(2001:10000)]
g36<-G36[c(2001:10000)]
g37<-G37[c(2001:10000)]
g38<-G38[c(2001:10000)]
g39<-G39[c(2001:10000)]
g40<-G40[c(2001:10000)]
g41<-G41[c(2001:10000)]
g42<-G42[c(2001:10000)]
g43<-G43[c(2001:10000)]
g44<-G44[c(2001:10000)]
g45<-G45[c(2001:10000)]
g46<-G46[c(2001:10000)]
g47<-G47[c(2001:10000)]
g48<-G48[c(2001:10000)]
g49<-G49[c(2001:10000)]
g50<-G50[c(2001:10000)]
g51<-G51[c(2001:10000)]
g52<-G52[c(2001:10000)]
g53<-G53[c(2001:10000)]
g54<-G54[c(2001:10000)]
g55<-G55[c(2001:10000)]
g56<-G56[c(2001:10000)]
g57<-G57[c(2001:10000)]
g58<-G58[c(2001:10000)]
g59<-G59[c(2001:10000)]
g60<-G60[c(2001:10000)]
g61<-G61[c(2001:10000)]
g62<-G62[c(2001:10000)]
g63<-G63[c(2001:10000)]
g64<-G64[c(2001:10000)]
g65<-G65[c(2001:10000)]
g66<-G66[c(2001:10000)]
```

```
g67<-G67[c(2001:10000)]
g68<-G68[c(2001:10000)]
g69<-G69[c(2001:10000)]
g70<-G70[c(2001:10000)]
g71<-G71[c(2001:10000)]
g72<-G72[c(2001:10000)]
g73<-G73[c(2001:10000)]
g74<-G74[c(2001:10000)]
g75<-G75[c(2001:10000)]
g76<-G76[c(2001:10000)]
g77<-G77[c(2001:10000)]
g78<-G78[c(2001:10000)]
g79<-G79[c(2001:10000)]
g80<-G80[c(2001:10000)]
g81<-G81[c(2001:10000)]
g82<-G82[c(2001:10000)]
g83<-G83[c(2001:10000)]
g84<-G84[c(2001:10000)]
g85<-G85[c(2001:10000)]
g86<-G86[c(2001:10000)]
g87<-G87[c(2001:10000)]
g88<-G88[c(2001:10000)]
g89<-G89[c(2001:10000)]
g90<-G90[c(2001:10000)]
g91<-G91[c(2001:10000)]
g92<-G92[c(2001:10000)]
g93<-G93[c(2001:10000)]
g94<-G94[c(2001:10000)]
g95<-G95[c(2001:10000)]
g96<-G96[c(2001:10000)]
g97<-G97[c(2001:10000)]
g98<-G98[c(2001:10000)]
g99<-G99[c(2001:10000)]
g100<-G100[c(2001:10000)]
g101<-G101[c(2001:10000)]
g102<-G102[c(2001:10000)]
g103<-G103[c(2001:10000)]
g104<-G104[c(2001:10000)]
g105<-G105[c(2001:10000)]
g106<-G106[c(2001:10000)]
g107<-G107[c(2001:10000)]
g108<-G108[c(2001:10000)]

#########################################################
#    Collecting Spectra Data in a Matrix 90x8000   #
#########################################################

tabella<-matrix(0, nrow=90, ncol=8000)

tabella[1,]<-g19
tabella[2,]<-g20
tabella[3,]<-g21
tabella[4,]<-g22
tabella[5,]<-g23
tabella[6,]<-g24
tabella[7,]<-g25
tabella[8,]<-g26
tabella[9,]<-g27
tabella[10,]<-g28
tabella[11,]<-g29
tabella[12,]<-g30
tabella[13,]<-g31
tabella[14,]<-g32
tabella[15,]<-g33
tabella[16,]<-g34
tabella[17,]<-g35
tabella[18,]<-g36
tabella[19,]<-g37
```

```
tabella[20,]<-g38
tabella[21,]<-g39
tabella[22,]<-g40
tabella[23,]<-g41
tabella[24,]<-g42
tabella[25,]<-g43
tabella[26,]<-g44
tabella[27,]<-g45
tabella[28,]<-g46
tabella[29,]<-g47
tabella[30,]<-g48
tabella[31,]<-g49
tabella[32,]<-g50
tabella[33,]<-g51
tabella[34,]<-g52
tabella[35,]<-g53
tabella[36,]<-g54
tabella[37,]<-g55
tabella[38,]<-g56
tabella[39,]<-g57
tabella[40,]<-g58
tabella[41,]<-g59
tabella[42,]<-g60
tabella[43,]<-g61
tabella[44,]<-g62
tabella[45,]<-g63
tabella[46,]<-g64
tabella[47,]<-g65
tabella[48,]<-g66
tabella[49,]<-g67
tabella[50,]<-g68
tabella[51,]<-g69
tabella[52,]<-g70
tabella[53,]<-g71
tabella[54,]<-g72
tabella[55,]<-g73
tabella[56,]<-g74
tabella[57,]<-g75
tabella[58,]<-g76
tabella[59,]<-g77
tabella[60,]<-g78
tabella[61,]<-g79
tabella[62,]<-g80
tabella[63,]<-g81
tabella[64,]<-g82
tabella[65,]<-g83
tabella[66,]<-g84
tabella[67,]<-g85
tabella[68,]<-g86
tabella[69,]<-g87
tabella[70,]<-g88
tabella[71,]<-g89
tabella[72,]<-g90
tabella[73,]<-g91
tabella[74,]<-g92
tabella[75,]<-g93
tabella[76,]<-g94
tabella[77,]<-g95
tabella[78,]<-g96
tabella[79,]<-g97
tabella[80,]<-g98
tabella[81,]<-g99
tabella[82,]<-g100
tabella[83,]<-g101
tabella[84,]<-g102
tabella[85,]<-g103
tabella[86,]<-g104
tabella[87,]<-g105
```

```
tabella[88,]<-g106
tabella[89,]<-g107
tabella[90,]<-g108
#################################################################
# PCA auotoscaled on a Raw Spectral Data Matrix with Simbols#
#################################################################

analisi<-prcomp(tabella[,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorigruppi<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))

plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2,
col=colorigruppi)

title(xlab=list("PC1 (18.4%)",cex=1.3, font=1))
title(ylab=list("PC2 (15%)",cex=1.3, font=1))
title(main=list("PCA on Raw Data", cex=2, font=1))

legend(locator(2), c("T-WT", "T-GM1","S-WT","S-GM1", "S-GM2"),
pch=c(16,1,17,6,22),cex=1.3,col=c("dark
green","orange","red","blue", "purple"))

#######################################
# Evaluation of % Cumulative Variance #
#######################################

x<-0
varcum<-rep(0,90)
propvarcum<-rep(0,90)
for (i in 1:90) {
j<-i-1
varcum[i]<-analisi$sdev[i]^2+x
x<-varcum[i]
propvarcum[i]<-(varcum[i]/sum(analisi$sdev^2))*100
}
plot(propvarcum, xlab=" PCs", ylab="Cumulative % variance")
propvarcum

#################################################################
# Refer all Spectra on Beta Anomeric Glucose                    #
#################################################################

ncamp<-90
punti<-8000
centro<-4265 #(refered to the best representativ spectrum N°2)
maxvet<-matrix(data=0, ncol=1, nrow=ncamp)
reganomer<-c(4255:4280)
for (i in 1:ncamp){
for (j in reganomer){
if (tabella[i,j]==max(tabella[i,reganomer])) maxvet[i,]<- centro-
j
}
}
allineati<- matrix(data=0, ncol=punti, nrow=ncamp)
for (i in 1:ncamp) {
shift<-maxvet[i,1]
if (shift>0) allineati[i,(1+shift):punti]<-tabella[i,1:(punti-
shift)]
if (shift==0) allineati[i,]<-tabella[i,]
if (shift<0) allineati[i,1:(punti+shift)]<-tabella[i,(1-
shift):punti]
```

```
}
plot(maxvet)

##########################
#   Plotting All Spectra #
##########################

z<-c(4200:4300)
plot(allineati[1,z], type="l", ylim=c(0,110000),
ylab="Intensity", xlab="number of points")
for (i in 1:ncamp) {
lines(allineati[i,z])
}

z<-c(5600:8000)
plot(allineati[1,z], type="l", ylim=c(0,1000))
for (i in 1:ncamp) {
lines(allineati[i,z])
}

#####################################################################
# PCA auotoscaled on Aligned Spectral Data Matrix with Simbols #
#####################################################################

analisi<-prcomp(allineati[,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorigruppi<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))

plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2,
col=colorigruppi)

title(xlab=list("PC1 (18.6%)",cex=1.3, font=1))
title(ylab=list("PC2 (15.2%)",cex=1.3, font=1))
title(main=list("PCA on Aligned Data", cex=2, font=1))

legend(locator(2), c("T-WT", "T-GM1","S-WT","S-GM1", "S-GM2"),
pch=c(16,1,17,6,22),cex=1.3,col=c("dark
green","orange","red","blue", "purple"))

#############################################################
#   Normalization on a Significant Peak Area          #
#############################################################

#riferiti<-allineatibis
riferiti<-allineati
coeff<-rep(1,ncamp)
area2<-sum (riferiti[2,4500:5500])

for (i in 1:ncamp) {
somma<-sum(riferiti[i,4500:5500])
coeff[i]<-somma/area2
}
for (i in 1:ncamp) {
riferiti[i,]<-riferiti[i,]/coeff[i]
}

#############################################################
# Coeficient of Normalization on Total Spectrum Area   #
#############################################################

ncamp<-nrow(allineati)
```

```
coefmat<-matrix(data=0, nrow=ncamp, ncol=1)
for(i in 1:ncamp){
coefmat[i,]<-sum(allineati[i,])
}
media<-sum(coefmat)/ncamp

########################################
#  Normalization on Total Specrtum Area  #
########################################

ncamp<-nrow(allineati)
npunti<-ncol(allineati)
riferiti<-matrix(data=0, ncol=npunti, nrow=ncamp)
area<-matrix(data=0, ncol=1, nrow=ncamp)
for (j in 1:ncamp){
area[j,]<-sum(allineati[j,])
for (i in 1:npunti){
riferiti[j,i]<-allineati[j,i]/area[j,]
}
}
riferiti<-riferiti*media

#########################
#   Plotting All Spectra #
#########################

z<-c(4200:4300)
plot(riferiti[1,z], type="l", ylim=c(0,110000), ylab="Intensity",
xlab="number of points")
for (i in 1:ncamp) {
lines(riferiti[i,z])
}

z<-c(5600:8000)
plot(riferiti[1,z], type="l", ylim=c(0,1000))
for (i in 1:ncamp) {
lines(riferiti[i,z])
}

#####################################################################
#PCA auotoscaled on Normalized Spectral Data Matrix with Simbols#
#####################################################################

analisi<-prcomp(riferiti[,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorigruppi<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))

plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2,
col=colorigruppi)

title(xlab=list("PC1 (19%)",cex=1.3, font=1))
title(ylab=list("PC2 (9%)",cex=1.3, font=1))
title(main=list("PCA on Normalized Data", cex=2, font=1))

legend(locator(2), c("T-WT", "T-GM1","S-WT","S-GM1", "S-GM2"),
pch=c(16,1,17,6,22),cex=1.3,col=c("dark
green","orange","red","blue", "purple"))

########################################
#        More PCs Combination Plotting    #
########################################
```

```
pcvalues<-as.data.frame(analisi$x)
pairs(pcvalues[1:4],col=colorigruppi)

#oppure solo riferito ai gruppi
analisi<-prcomp(riferiti,center=T,scale=T)
y<-c(1,3)
plot(analisi$x[,y])
plot(analisi$x[,y], main="PC1 vs PC3", pch="*")
text(analisi$x[,y], labels=c("*"),col=colorigruppi)
pcvalues<-as.data.frame(analisi$x)
pairs(pcvalues[1:3],col=colorigruppi, pch="*")

#############
# Smoothing #
#############

ncamp<-nrow(riferiti)
sm<-edit (riferiti [1:5,1:10])

smooth<-matrix(data=0, nrow=ncamp, ncol=8000)
for (z in 1:ncamp) {
for (i in 3:(ncol(riferiti)-2)) {
w1<-(i-2)
w2<-(i+2)
w<-c(w1:w2)
a<-mean(riferiti[z,w])
smooth[z,i]<-a
}
}
tabella<-smooth
sm<-edit (tabella [1:5,1:10])

############################
#   Specral Derivate       #
############################

ncamp<-nrow(tabella)

derivata<-matrix(data=0, nrow=ncamp, ncol=8000)
for (z in 1:ncamp) {
for (i in 1:(ncol(tabella)-1)){
j<-i+1
a<-(tabella[z,j]-tabella[z,i])
derivata[z,i]<-a
}
}
edit (derivata[1,1:10])
windows()
plot(derivata[1,], type="l")

#########################
#  PeakPick Matrix      #
#########################

ncamp<-nrow(derivata)
sm<-edit(ncamp)
peakpick<-matrix(data=0, nrow=ncamp, ncol=8000)
i<-0
j<-0
for (z in 1:ncamp) {
for (i in 5:(ncol(derivata)-6)){
j<-i+1
a<-(derivata[z,i])
b<-(derivata[z,j])
c<-(tabella[z,j])
d<-derivata[z,(j+4)]-derivata[z,(i-4)]
if (a>0)
```

```
if (a*b<0)
if (abs(d)>=2)
peakpick[z,j]<-c
}
}
plot(peakpick[1,], type="l")

##############################
# Binning Spectra 100 Points #
##############################

#length(peakpick[1,])
ncamp<-nrow(riferiti)
intervallo<-matrix(data=0, nrow=ncamp, ncol=80)
for (z in 1:ncamp) {
i<-1
for (i in 1:80){
for (j in 1:100){
k<-((i-1)*100)+j
a<-riferiti[z,k]# Here the name of the matrix to be binned
intervallo[z,i]<-intervallo[z,i]+a
}
intervallo[z,i]<-intervallo[z,i]/100
}
}

#####################################################################
# PCA auotoscaled on Binned Spectral Data Matrix with Symbols #
#####################################################################

analisi<-prcomp(intervallo[,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorigruppi<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))

plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2,
col=colorigruppi)

title(xlab=list("PC1 (26%)",cex=1.3, font=1))
title(ylab=list("PC2 (12%)",cex=1.3, font=1))
title(main=list("PCA on Binned Data", cex=2, font=1))

legend(locator(2), c("T-WT", "T-GM1","S-WT","S-GM1", "S-GM2"),
pch=c(16,1,17,6,22),cex=1.3,col=c("dark
green","orange","red","blue", "purple"))

########################################
# Evaluation of % Cumulative Variance #
########################################

x<-0
varcum<-rep(0,80)
propvarcum<-rep(0,80)
for (i in 1:80) {
j<-i-1
varcum[i]<-analisi$sdev[i]^2+x
x<-varcum[i]
propvarcum[i]<-(varcum[i]/sum(analisi$sdev^2))*100
}

y<-c(propvarcum)
x<-c(1:80)
```

```
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ")
title(xlab=list("Number of PC",cex=1.3, font=1))
title(ylab=list("Cumulative Variance (%)",cex=1.3, font=1))
abline(h= 78.57981, col="red")
abline(v=14, col="red")# to add a x=a line

#####################################################################
# PCA auotoscaled on Binned Spectral Silcora Data Matrix with       #
# Simbols                                                           #
#####################################################################

analisi<-prcomp(intervallo[37:90,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorisilcora<-c(rep("red",18),rep("blue",18),rep("purple",18))
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(17,18),
rep(6,18), rep(22,18)), cex=2, col=colorisilcora)

title(xlab=list("PC1 (25%)",cex=1.3, font=1))
title(ylab=list("PC2 (13%)",cex=1.3, font=1))
title(main=list("Silcora PCA Binned Data", cex=2, font=1))

legend(locator(2), c("S-WT","S-GM1", "S-GM2"), pch=c(17,6,22),
cex=1.3, col=c("red","blue", "purple"))

###################################
#   BarPlot Loadings Values PC1   #
###################################

barplot(analisi$rotation[,1], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Silcora PC1 Loadings", cex=2, font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.2)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values  PC1  #
#########################################################

a<-rep(0,5)
b<-analisi$rotation[6:10,1]
c<-rep(0,11)
d<-analisi$rotation[22:26,1]
e<-rep(0,32)
f<-analisi$rotation[59:61,1]
g<-rep(0,19)
i<-c(a,b,c,d,e,f,g)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Silcora   Main   Loandings   PC1   Bins",   cex=2,
font=1))
abline(h= 0.1)#to add a y=a line
abline(h=-0.11)#to add a second y=-a  line

###################################
#   BarPlot Loadings Values PC2   #
###################################

barplot(analisi$rotation[,2], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
```

```
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Silcora PC2 Loadings", cex=2, font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.16)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values PC2  #
#########################################################

a<-rep(0,39)
b<-analisi$rotation[40,2]
c<-rep(0,40)
i<-c(a,b,c)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Silcora   Main   Loandings   PC2   Bins",   cex=2,
font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.16)#to add a second y=-a  line

#####################################################################
# Representation of the mean Area for Each bins Silcora Samples #
#####################################################################

z<-c(3900:4000)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

mean(integrale[37:54,1])
mean(integrale[55:72,1])
mean(integrale[73:90,1])

sd3<-sd(integrale[37:54,1])
sd4<-sd(integrale[55:72,1])
sd5<-sd(integrale[73:90,1])

sd3
sd4
sd5

#####################################################################
# PCA auotoscaled on Binned Spectral Thompson Data Matrix       #
# with Simbols                                                  #
#####################################################################

analisi<-prcomp(intervallo[1:36,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colorithompson<-c(rep("dark green",18),rep("orange",18))
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18)), cex=2, col=colorithompson)

title(xlab=list("PC1 (15%)",cex=1.3, font=1))
title(ylab=list("PC2 (12%)",cex=1.3, font=1))
title(main=list("Thompson PCA Binned Data", cex=2, font=1))
legend(locator(2), c("T-WT","T-GM1"), pch=c(16,1), cex=1.3,
col=c("dark green","orange"))
```

```
###################################
#   BarPlot Loadings Values PC1   #
###################################

barplot(analisi$rotation[,1], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Thompson PC1 Loadings", cex=2, font=1))
abline(h= 0.095)#to add a y=a line
abline(h=-0.175)#to add a second y=-a  line


###################################
#   BarPlot Loadings Values PC2   #
###################################

barplot(analisi$rotation[,2], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Thompson PC2 Loadings", cex=2, font=1))
abline(h= 0.18)#to add a y=a line
abline(h=-0.19)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values PC1   #
#########################################################

a<-rep(0,23)
b<-analisi$rotation[24:25,1]
c<-rep(0,9)
d<-analisi$rotation[35:37,1]
e<-rep(0,21)
f<-analisi$rotation[59:61,1]
g<-rep(0,19)
i<-c(a,b,c,d,e,f,g)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Thompson   Main   Loandings   PC1   Bins",   cex=2,
font=1))
abline(h= 0.095)#to add a y=a line
abline(h=-0.175)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values PC2  #
#########################################################

a<-rep(0,38)
b<-analisi$rotation[39:41,2]
c<-rep(0,39)
i<-c(a,b,c)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Thompson   Main   Loandings   PC2   Bins",   cex=2,
font=1))
abline(h= 0.18)#to add a y=a line
abline(h=-0.19)#to add a second y=-a  line

#####################################################################
# Representation of the mean Area for Each bins Thompson            #
# Samples                                                           #
#####################################################################
```

```
z<-c(3800:4100)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

mean(integrale[1:18,1])
mean(integrale[19:36,1])

sd1<-sd(integrale[1:18,1])
sd2<-sd(integrale[19:36,1])

sd1
sd2

######################################################################
# PCA auotoscaled on Binned Spectral Data Matrix with Simbols   #
######################################################################

analisi<-prcomp(intervallo,center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])

plot(x,y)
colori<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2, col=colori)

title(xlab=list("PC1 (12%)",cex=1.3, font=1))
title(ylab=list("PC2 (26%)",cex=1.3, font=1))
title(main=list("PCA Binned Data all Samples", cex=2, font=1))

legend(locator(2), c("T-WT", "T-GM1", "S-WT","S-GM1", "S-GM2"),
pch=c(16,1,17,6,22), cex=1.3, col=c("dark green",
"orange","red","blue", "purple"))

###################################
#   BarPlot Loadings Values PC1   #
###################################

barplot(analisi$rotation[,1], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("All Samples PC1 Loadings", cex=2, font=1))
abline(h= 0.17)#to add a y=a line
abline(h=-0.18)#to add a second y=-a  line

###################################
#   BarPlot Loadings Values PC2   #
###################################

barplot(analisi$rotation[,2], axes=FALSE, col=c(rep("green",37),
rep("light blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("All Samples PC2 Loadings", cex=2, font=1))
abline(h= 0.1)#to add a y=a line
abline(h=-0.14)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values PC1   #
#########################################################
```

```
a<-rep(0,23)
b<-analisi$rotation[24:25,1]
c<-rep(0,33)
d<-analisi$rotation[59:61,1]
e<-rep(0,19)
i<-c(a,b,c,d,e)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("All  Samples  Main  Loandings  PC1  Bins",  cex=2,
font=1))
abline(h= 0.17)#to add a y=a line
abline(h=-0.18)#to add a second y=-a  line

#########################################################
#   BarPlot most representative Loadings Values PC2  #
#########################################################

a<-rep(0,5)
b<-analisi$rotation[6:10,2]
c<-rep(0,70)
i<-c(a,b,c)
barplot(i, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("All  Samples  Main  Loandings  PC2  Bins",  cex=2,
font=1))
abline(h= 0.1)#to add a y=a line
abline(h=-0.14)#to add a second y=-a  line

#####################################################################
# Representation of the mean Area for Each bins All Samples     #
#####################################################################

z<-c(4700:5000)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

mean(integrale[1:18,1])
mean(integrale[19:36,1])
mean(integrale[37:54,1])
mean(integrale[55:72,1])
mean(integrale[73:90,1])

sd1<-sd(integrale[1:18,1])
sd2<-sd(integrale[19:36,1])
sd3<-sd(integrale[37:54,1])
sd4<-sd(integrale[55:72,1])
sd5<-sd(integrale[73:90,1])

sd1
sd2
sd3
sd4
sd5

############################
# T Student or T-test     #
############################

plotpc<-matrix(data=0, ncol=5, nrow=18)

plotpc[,1]<-analisi$x[1:18,2] # on PC2
```

```
plotpc[,2]<-analisi$x[19:36,2]
plotpc[,3]<-analisi$x[37:54,2]
plotpc[,4]<-analisi$x[55:72,2]
plotpc[,5]<-analisi$x[73:90,2]

# Confronto tra Varianza di due Popolazioni

var(plotpc[,1])
var(plotpc[,2])
var(plotpc[,3])
var(plotpc[,4])
var(plotpc[,5])

#Var test per Confermare l'Uguaglianza tra le Varianze: Calcolo
dell'intervalo di Confidenza al 95%

var.test(plotpc[,1],plotpc[,2])$conf.int
var.test(plotpc[,3],plotpc[,4])$conf.int
var.test(plotpc[,3],plotpc[,5])$conf.int

#se l'Intervallo di confidenza contiene 1 Allora le 2 Varianze
sono Uguali

t.test(plotpc[,1],plotpc[,2], var.equal=T) # per Var Uguali
t.test(plotpc[,3],plotpc[,4], var.equal=T) # per Var Uguali
t.test(plotpc[,3],plotpc[,5], var.equal=T) # per Var Uguali

#t.test(plotpc[,1],plotpc[,2], var.equal=F) #Welch test per Var
Diverse

qt(o.975, 34) # Quartile con confidenza dello 0,05%

#########################
#     ANOVA One Way     #
#########################

livelli<-rep(c("A","B","C","D","E"), c(18,18,18,18,18))
livelli<-as.factor(livelli)
levels(livelli)
table(livelli)

sm<-edit (analisi$x[,1])

y<-c(analisi$x[,2]) # referred to PC1.
mean(y)

tapply(y,livelli,mean)
tapply(y,livelli,length)
boxplot(y~livelli)

plot.design(y~livelli) # To see distances's groups

mod<-lm(y~livelli)
summary(mod)
anova(mod)

###########
#  LDA    #
###########

#Download MASS package

pc<-c(1,2,4,5,8,9,10)# Main PCs chosen from ANOVA
colori<-c(rep("dark
green",18),rep("orange",18),rep("red",18),rep("blue",18),rep("pur
ple",18))

pcvalues<-as.data.frame(analisi$x[,pc])
```

```
attach(pcvalues)
livelli<-rep(c("T-WT","T-GM1","S-WT","S-GM1","S-GM2"),
c(18,18,18,18,18))
livelli<-as.factor(livelli)
levels(livelli)
table(livelli)
ldauva<-lda(livelli~.,pcvalues, prior=c(1,1,1,1,1)/5) # con
"prior" dò probabilità a priori uguale a tutti i gruppi

lduva<-predict(ldauva) #al predict assegno un nome diverso
plot(lduva$x,type="n")
text(lduva$x, labels=as.character(livelli), col=colori,
main="LDA")

plot(ldauva, dimen=4, col= colori)

plot(ldauva, dimen=2, col= colori)

########################################
# Coefficients of linear discriminants #
########################################

# To find out the PC chosen coefficients for each LD

ldauva

######################
# Plot LD1 vs LD2    #
######################

LD1<-predict(ldauva)$x[,1]
LD2<-predict(ldauva)$x[,2]

plot(LD1,LD2, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2, col=colori)

title(xlab=list("LD1",cex=1.3, font=1))
title(ylab=list("LD2",cex=1.3, font=1))
title(main=list("LDA Binned Data all Samples", cex=2, font=1))

######################
# Plot LD3 vs LD4    #
######################

LD3<-predict(ldauva)$x[,3]
LD4<-predict(ldauva)$x[,4]

colorigruppi2<-c(rep("dark green",18),rep("dark
orange",18),rep("grey",18),rep("grey",18),rep("grey",18))

plot(LD3,LD4, cex.axis=1.3, xlab=" ",ylab=" ", pch=c(rep(16,18),
rep(1,18), rep(17,18), rep(6,18), rep(22,18)), cex=2,
col=colorigruppi2)

title(xlab=list("LD3",cex=1.3, font=1))
title(ylab=list("LD4",cex=1.3, font=1))
title(main=list("LDA Binned Data all Samples", cex=2, font=1))

######################
#   LD Box Plot      #
######################

livel<-rep(c("0-T-WT","1-T-GM1","2-S-WT","3-S-GM1","4-S-GM2"),
c(18,18,18,18,18))
y<-c(lduva$x[,2])
colbox<-c(rep("dark
green",1),rep("orange",1),rep("red",1),rep("blue",1),rep("purple"
,1))
```

```
boxplot(y~livel,col=colbox, cex.axis=1.3)
title(xlab=list("Groups",cex=1.3, font=1))
title(ylab=list("Variance",cex=1.3, font=1))
title(main=list("LD4 Representation", cex=2, font=1))

########################################
#    Contribution of each PC on each LD #
########################################

#xLD1: PCn (all samples) for n coefficient obtained by the com-
mand ldauva

PC1LD1<-analisi$rotation[,1]*-1.19688174
PC2LD1<-analisi$rotation[,2]* 0.09604888
PC4LD1<-analisi$rotation[,4]* 0.17389633
PC5LD1<-analisi$rotation[,5]* -0.16385861
PC8LD1<-analisi$rotation[,8]* -0.50760487
PC9LD1<-analisi$rotation[,9]* -0.23438227
PC10LD1<-analisi$rotation[,10]* -0.01937333

summa<-PC1LD1+PC2LD1+PC4LD1+PC5LD1+PC8LD1+PC9LD1+PC10LD1

barplot(summa, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin on LD1", cex=1.3,
font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.2)#to add a second y=-a  line

#xLD2

PC1LD2<-analisi$rotation[,1]* 0.05578327
PC2LD2<-analisi$rotation[,2]* 0.65596033
PC4LD2<-analisi$rotation[,4]* 0.11496348
PC5LD2<-analisi$rotation[,5]* -0.54232059
PC8LD2<-analisi$rotation[,8]* -0.38127295
PC9LD2<-analisi$rotation[,9]* 0.16122044
PC10LD2<-analisi$rotation[,10]* 0.04063683

summa<-PC1LD2+PC2LD2+PC4LD2+PC5LD2+PC8LD2+PC9LD2+PC10LD2

barplot(summa, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin on LD2", cex=1.3,
font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.2)#to add a second y=-a  line

#xLD3

PC1LD2<-analisi$rotation[,1]* 0.02425739
PC2LD2<-analisi$rotation[,2]* -0.11150008
PC4LD2<-analisi$rotation[,4]* -0.37502033
PC5LD2<-analisi$rotation[,5]* -0.30811770
PC8LD2<-analisi$rotation[,8]* -0.48677809
PC9LD2<-analisi$rotation[,9]* -0.42430681
PC10LD2<-analisi$rotation[,10]* -0.27781272

summa<-PC1LD2+PC2LD2+PC4LD2+PC5LD2+PC8LD2+PC9LD2+PC10LD2

barplot(summa, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin on LD3", cex=1.3,
font=1))
```

```
abline(h= 0.2)#to add a y=a line
abline(h=-0.1)#to add a second y=-a  line

#xLD4

PC1LD2<-analisi$rotation[,1]* 0.02478813
PC2LD2<-analisi$rotation[,2]* -0.01066208
PC4LD2<-analisi$rotation[,4]* 0.25615319
PC5LD2<-analisi$rotation[,5]* 0.28685386
PC8LD2<-analisi$rotation[,8]* -0.49535749
PC9LD2<-analisi$rotation[,9]* 0.09231256
PC10LD2<-analisi$rotation[,10]* -0.46917014

summa<-PC1LD2+PC2LD2+PC4LD2+PC5LD2+PC8LD2+PC9LD2+PC10LD2

barplot(summa, axes=FALSE, col=c(rep("green",37), rep("light
blue",19), rep("red",24)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin on LD4", cex=1.3,
font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.1)#to add a second y=-a  line

#########################################################
# Representation of the mean Area for Organic Acids  #
#########################################################

z<-c(5800:6200)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

mean(integrale[1:18,1])
mean(integrale[19:36,1])
mean(integrale[37:54,1])
mean(integrale[55:72,1])
mean(integrale[73:90,1])

sd1<-sd(integrale[1:18,1])
sd2<-sd(integrale[19:36,1])
sd3<-sd(integrale[37:54,1])
sd4<-sd(integrale[55:72,1])
sd5<-sd(integrale[73:90,1])

sd1
sd2
sd3
sd4
sd5

##############################################
# comparison between Malic and Citric Acids #
##############################################

# Malic Acid
z<-c(5870:5970,6070:6170)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}
mean(integrale[,1])
sd1<-sd(integrale[,1])

sd1
```

```
# Citric Acid

z<-c(5970:6070)
riferitibis<-riferiti[1:90,]
integrale<-matrix(data=0, nrow=90, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}
mean(integrale[,1])
sd1<-sd(integrale[,1])

sd1
```

Appendix B.1 Thompson Seedless Wild Type (T-WT) spectrum

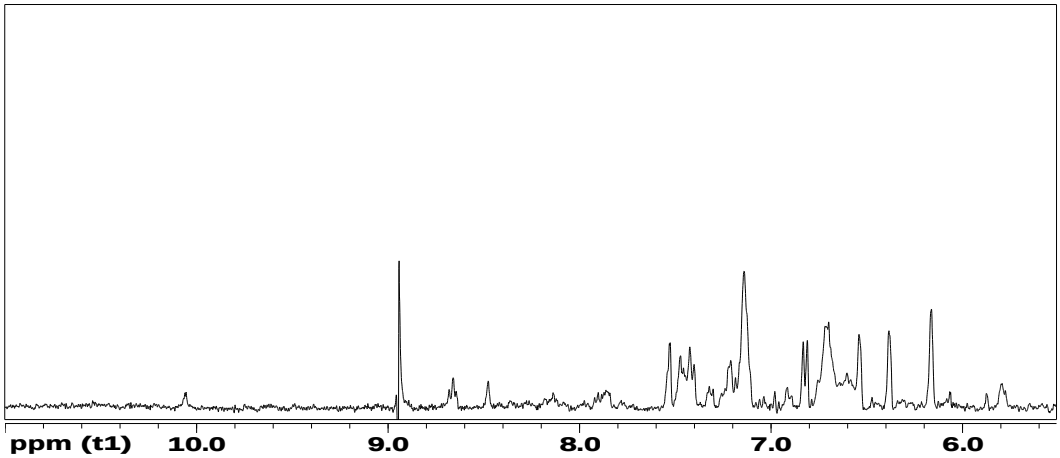


Figure Appendix B1: T-WT: Downfield Region (> 5.5 ppm)

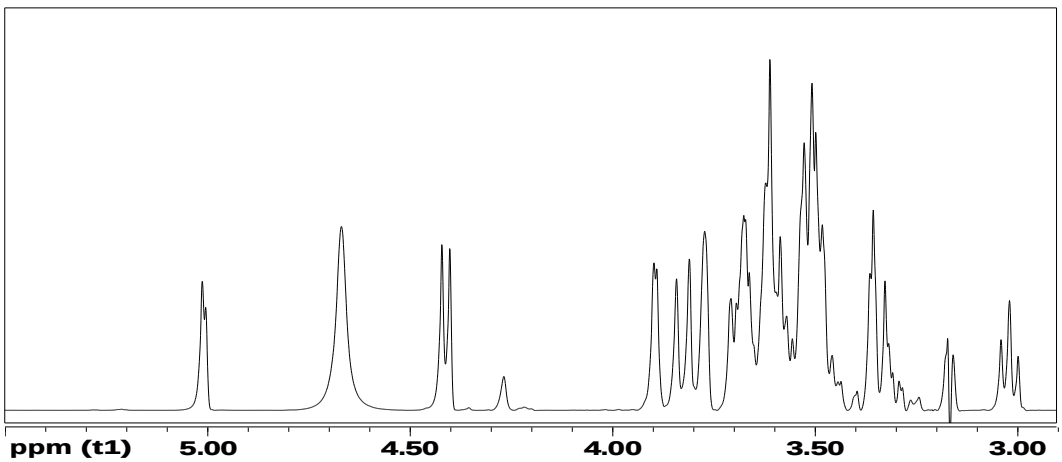


Figure Appendix B2: T-WT: Midfield Region (5.5 < ppm < 2.9)

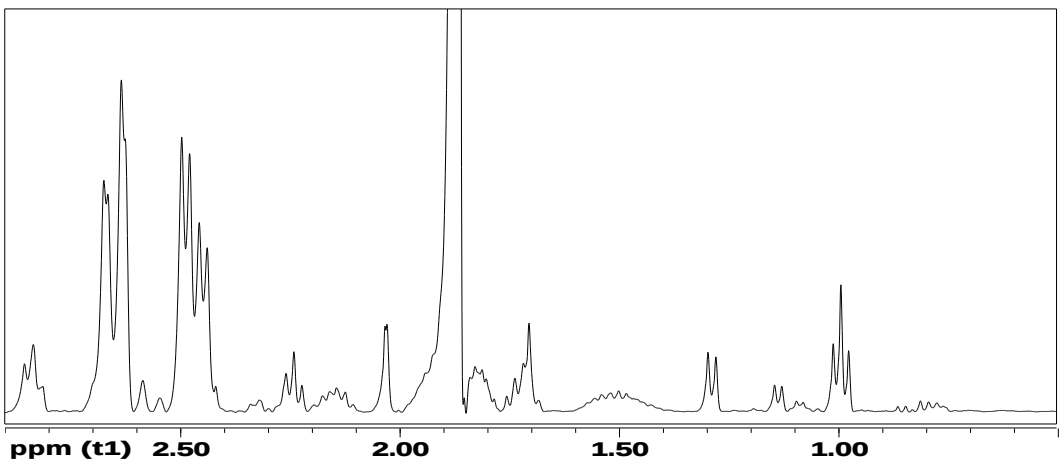


Figure Appendix B3: T-WT: Upfield Region (< 2.9 ppm)

Appendix B.2 Thompson Seedless Genetically Modified (T-GM1) spectrum

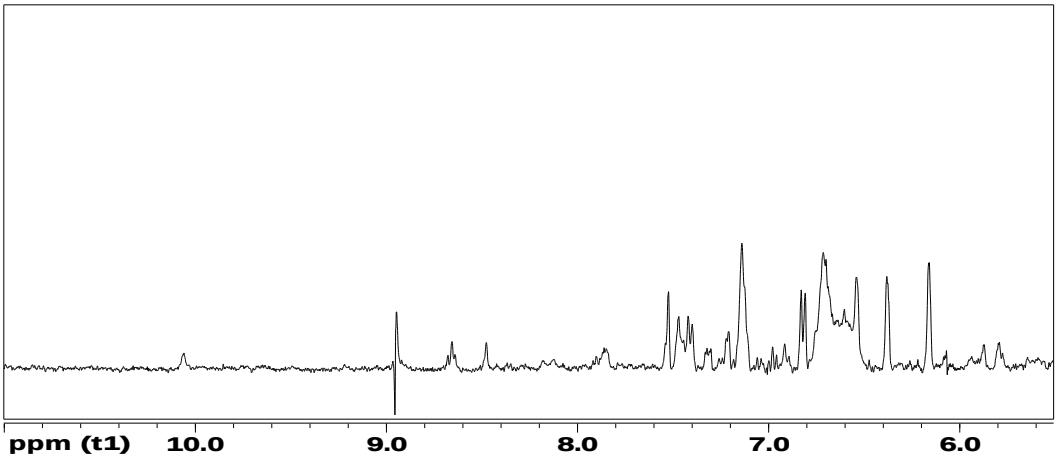


Figure Appendix B2: T-GM1: Downfield Region (> 5.5 ppm)

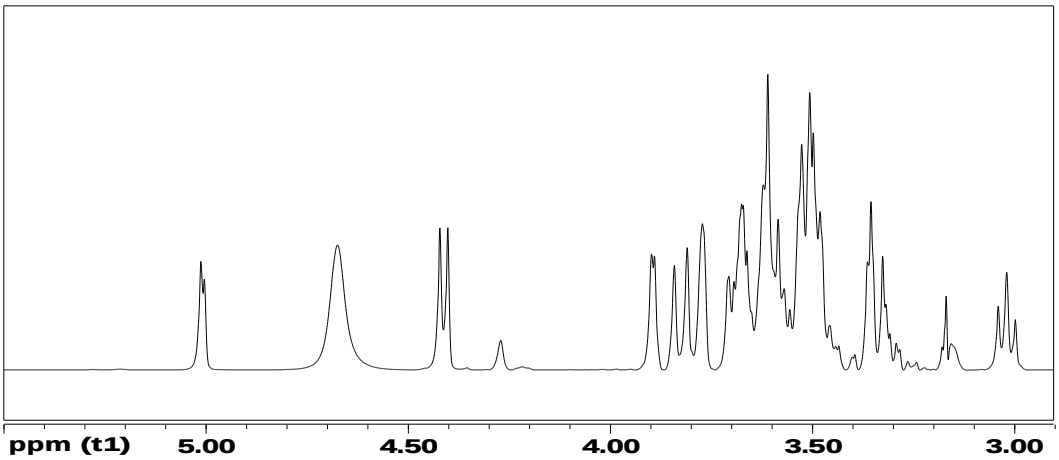


Figure Appendix B2: T-GM1: Midfield Region (5.5 < ppm < 2.9)

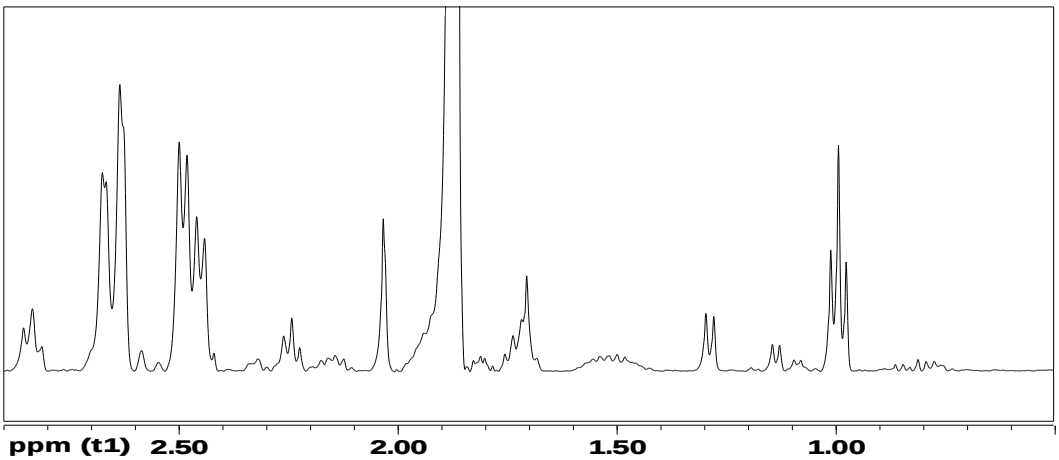


Figure Appendix B2: T-GM1: Upfield Region (< 2.9 ppm)

Appendix B.3 Silcora Wild Type (S-WT) spectrum

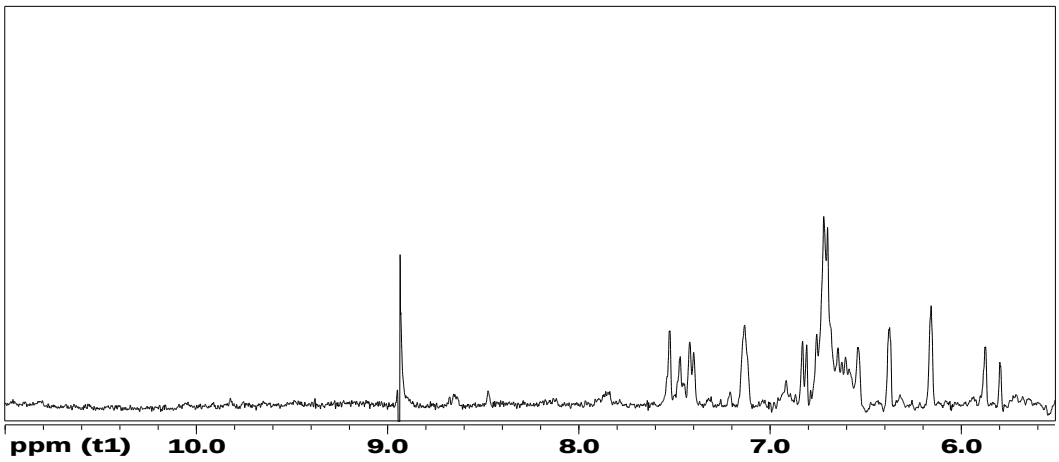


Figure Appendix B3: S-WT: Downfield Region (> 5.5 ppm)

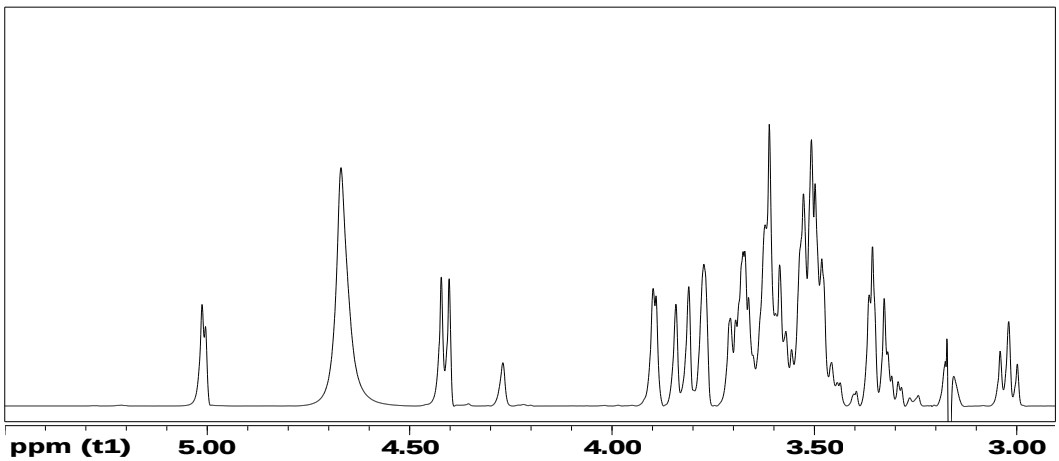


Figure Appendix B3: S-WT: Midfield Region (5.5 < ppm < 2.9)

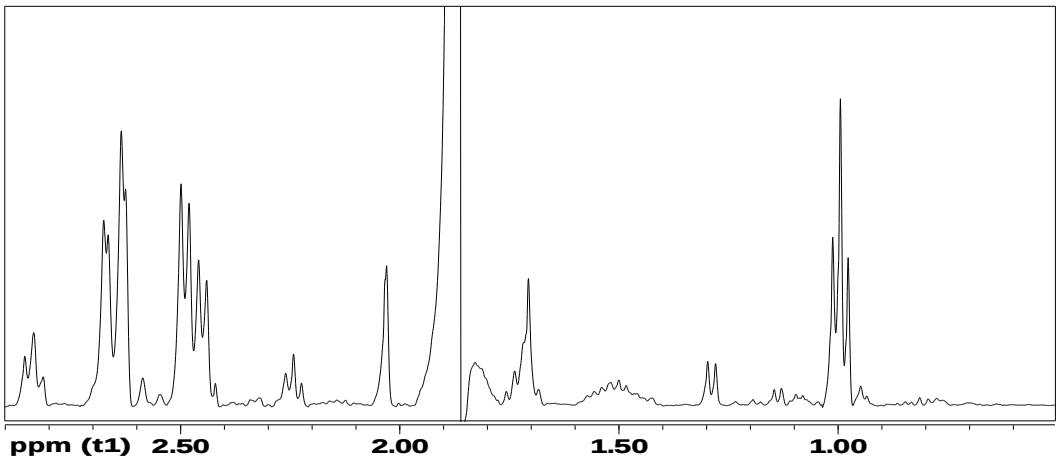


Figure Appendix B3: S-WT: Upfield Region (< 2.9 ppm)

Appendix B.4 Silcora Genetically Modified 1 Type (S-GM1) spectrum

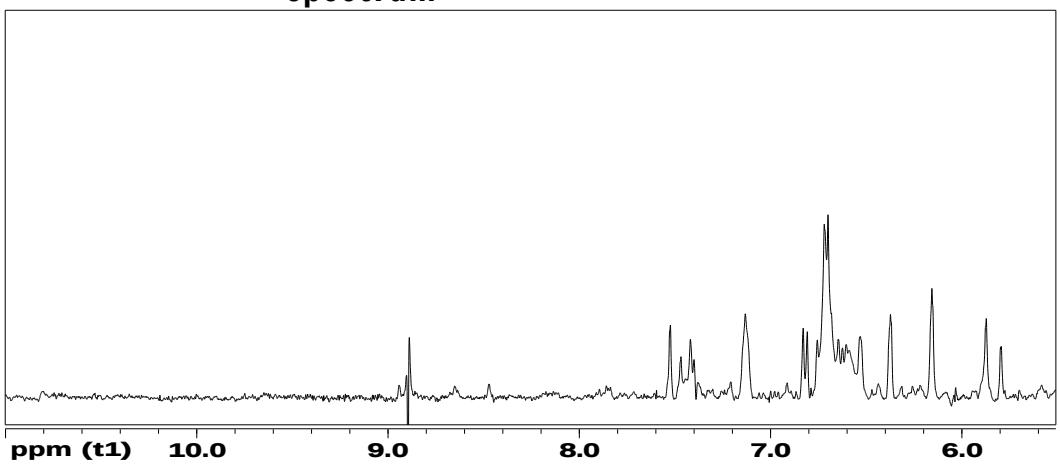


Figure Appendix B4: S-GM1: Downfield Region (> 5.5 ppm)

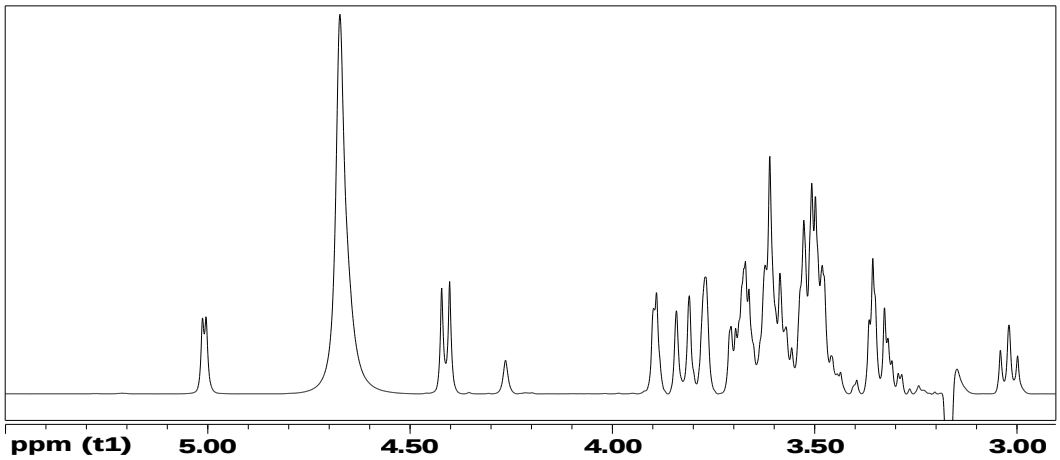


Figure Appendix B4: S-GM1: Midfield Region (5.5 < ppm < 2.9)

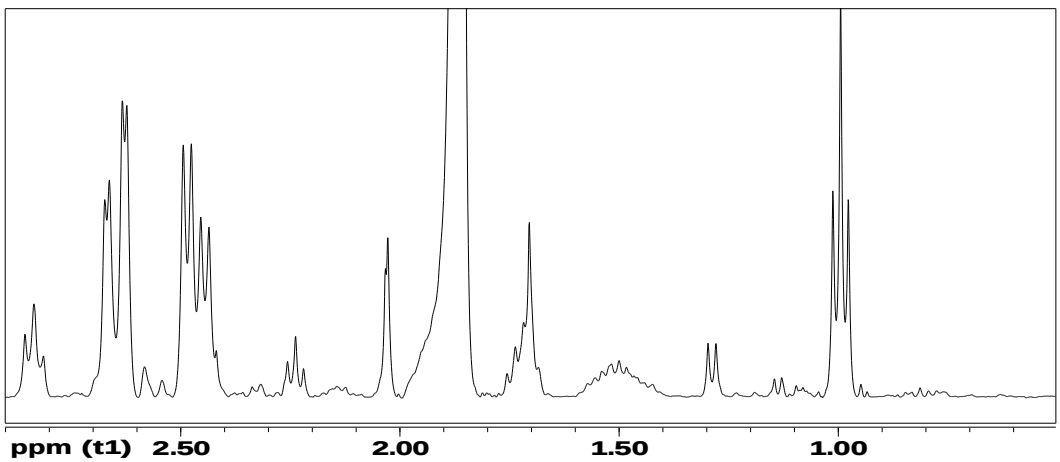


Figure Appendix B4: S-GM1: Upfield Region (< 2.9 ppm)

Appendix B.5 Silcora Genetically Modified 2 Type (S-GM2) spectrum

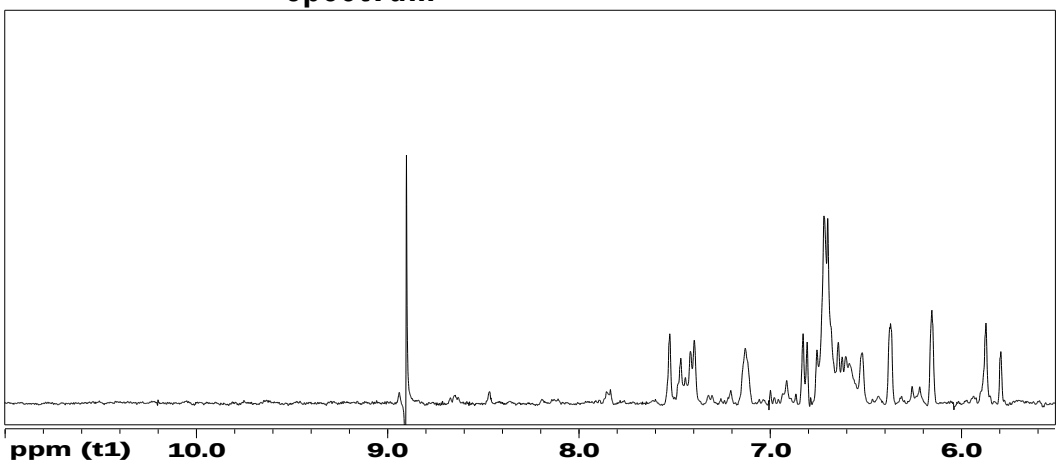


Figure Appendix B5: S-GM2: Downfield Region (> 5.5 ppm)

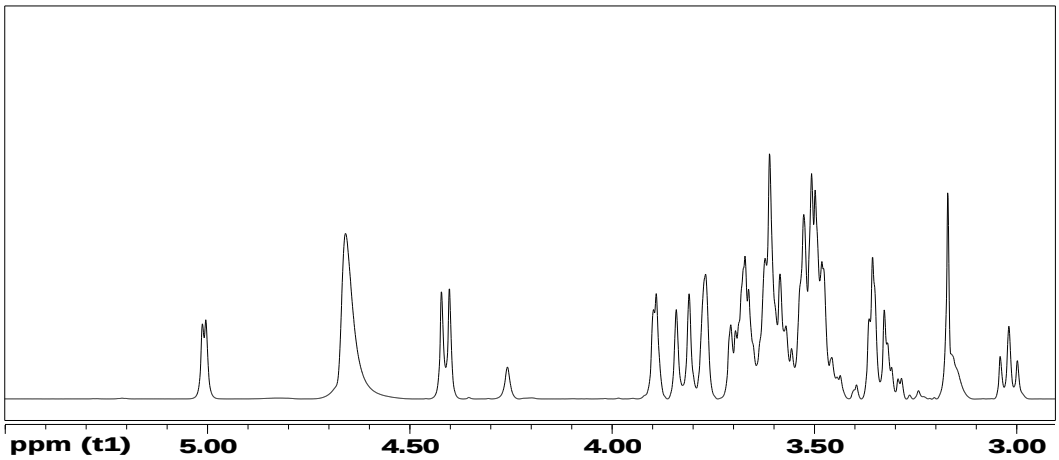


Figure Appendix B5: S-GM2: Midfield Region (5.5 < ppm < 2.9)

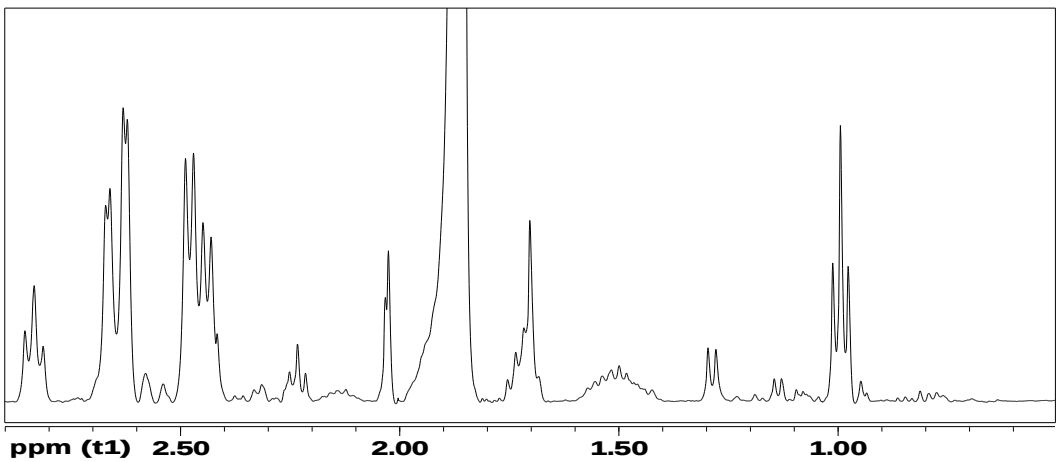
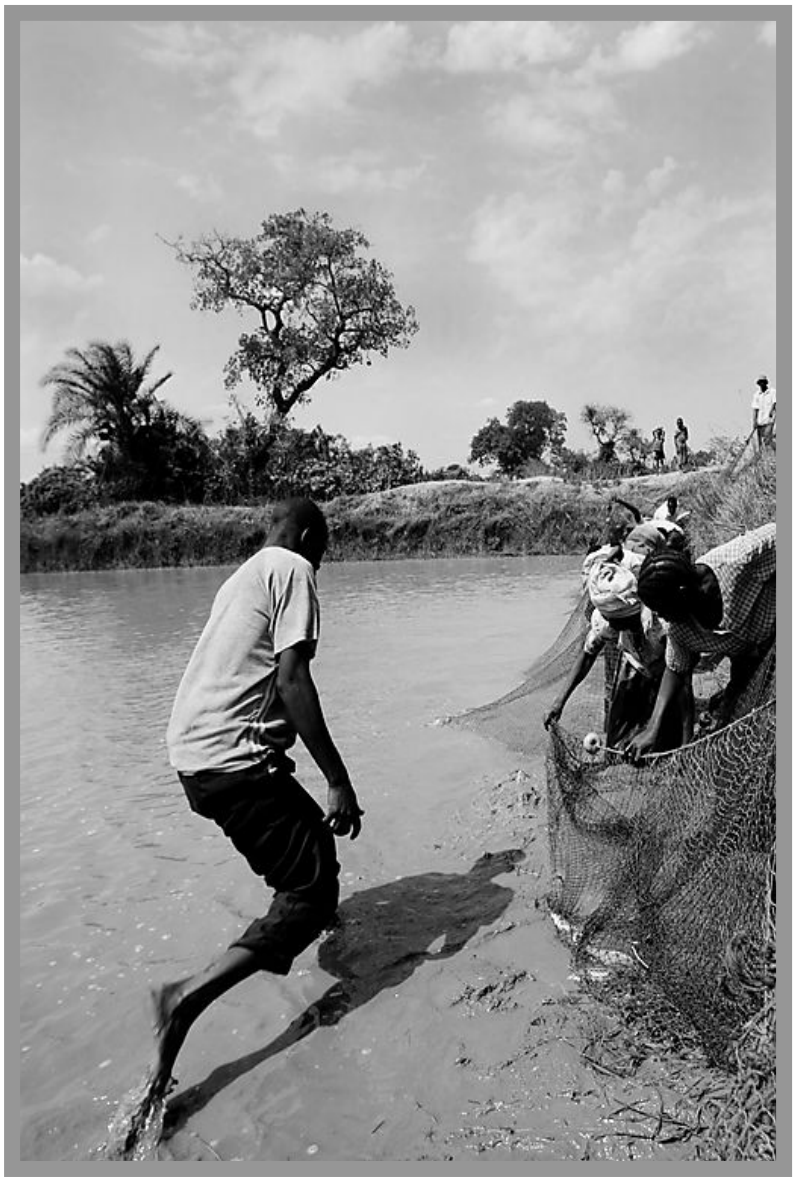


Figure Appendix B5: S-GM2: Upfield Region (< 2.9 ppm)

CHAPTER FOUR
- FISH FARMING -



4 METABONOMICS AS A TOOL OF INVESTIGATION OF QUALITY IN FISH FARMING

4.1 INTRODUCTION

Since last decade, several scientific studies on fish and seafood, including wild and farmed fish and shellfish, both of marine and freshwater origin, have shed light on the importance of these products in human diet as healthy products (C. M. Oomen et *al.*, 2000; H. Senzaki et *al.*, 2001; M. De Lorgeril et *al.*, 2002; C. Cahu et *al.*, 2004).

The main quality aspect derives from the nutritional value due especially to the presence of essential amino acids (arginine, histidine, isoleucine, leucine, lysine, methionine, plenylalanine, threonine, tryptophan and valine), highly digestible proteins (M. M. Friedman, 1997), vitamins (A, D and B complex), minerals and a high content of polyunsaturated fatty acids (PUFA). These latter represent the only source of ω-3 fatty acids, whose importance in prevention of human diseases has been largely discussed and abundantly presented in literature (A. L. Stoll et *al.*, 1999; C. H. S. Ruxton et *al.*, 2004; A. Bhise et *al.*, 2005; R. S. Pardini et *al.*, 2005).

The importance of seafood becomes so relevant that in less than fifty years the global demand for fish product had doubled and up today it is still rising (E. Engelhaupt, 2007). This aspect together with the development of new technologies in fishing activities and massive production through selective breeding, hybridization, and the application of biotechnology (D. W. Cole et *al.*, 2008) leads to substitute the traditional fishing with the fish farming.

For its increasing importance on the global economical scene, aquaculture becomes an important subject on which focus particularly attention especially from both safety and quality point of view, as the quality of farmed fish depends not only on its intrinsic characteristics such as species, age and sex, but also on factors such as the developmental phase, environmental temperature, feed regime and composition of lipids in the diet (M. Krajnovic-Ozretic et *al*., 1994).

In this way, the global quality of fish is determined also by the harvesting and post-harvesting procedures (aquaculture) as they represent today the higher percentage of the total amount of fish consumed (L. Senso et *al*., 2007). The definition of quality related to seafood, thus, becomes really complex.

Generally it can be defined as "**a combination of such characteristics as wholesomeness, integrity and freshness**" (R. Martin, 1988). These three characteristics include only an objective point of view of the term quality, and are specific to the seafood fit to eat (K. Grigorakis, 2007). Low in microbiological (B. T. Lunestad et *al*., 2008) and inorganic contamination (C. R. Santerre et *al*., 2001; R. A. Snyder, 2008), packaging and storage in a sanitary environment, taste, texture and appearance are all intrinsic parameters that define the quality.

From a subjective point of view, the quality perceived by the consumers is strictly connected to the nutritional value, flavor of and others sensory components (R. Jonsdottir et *al*., 2004) based on the chemical composition of fish. Additionally, the complexity of fish quality increases even more if the different practice of fish farming are taken in consideration. Thus, it becomes really important, in order to evaluate such a different increase in quality, to compare the whole metabolic profile in fish muscle between wild and farmed fish and among different methods of farming.

The Gilthead seab ream (*Sparus aurata*) is an example of economical fish species whose market has rapidly increased in the last decade. It has traditionally been cultured in Mediterranean coastal lagoons and salt water ponds or in the northern Adriatic lagoons (a.k.a. *valle)* in Italy (L. Sola et *al*., 2007). Due to the increasing demand, sea breams are farmed, up to day, extensively in lagoons, or intensively in tanks or cages (L. Sola et *al*., 2007).

This intensive production, especially in "artificial" conditions, has raised the problem of quality of farmed fish in comparison with wild fish. In the same time, because of different aquacultures' systems, the attention is also focused on these new rearing systems, trying to find which is the best one from the nutritional point of view.

Several works in literature report the composition of the fish lipid fraction through classical analysis like gas chromatography (C. Alasalvar et *al*., 2005) and, in the last decade, by NMR spectroscopy. The latter has been exploited for the characterization of fishes with an high economical relevance like Atlantic salmon (*Salmon salar*) (M. Aursand et *al*., 1992), Atlantic halibut (*Hippoglossus hippoglossus L*) (B. Sitter et *al*., 1999) and cod (*Gadus morhua*) (I. Martinez et *al*., 2005).

Despite Gilthead Sea bream (*Sparus aurata*) is an important and largely marked fish (L. Sola et *al*., 2007), only a few works have been carried aiming at characterizing its composition (S. D. Klaoudatos, 1989; N. Kalogeropoulos et *al*., 1992; C. Alasalvar et *al*., 2005; S. Rezzi et *al*., 2007). Thus, the work here introduced try to fill up the description of the metabolic profile of this fish species, by analyzing the composition of perchloric acid extract and by evaluating the metabolic changes due to different kind of aquaculture conditions and to different time of storage.

NMR spectroscopy is considered a key tool for understanding the

effect of perturbations on metabolic processes in living systems (F. F. Brown et *al*., 1997) as defined by the *metabonomics* concept (J. K. Nicholson et *al*., 1999). Conversely, *metabolomics* is "targeted at identifying and quantifying all metabolites in a given biological context" (T. Kind et *al*., 2006) and it is preferably based on mass spectrometry.

The NMR technique generates spectral data, representing a "unique metabolic fingerprint for each complex biological mixture" (F. F. Brown et *al*., 1997). In fact, if the physiologic equilibrium of an organism changes, its unique metabolic fingerprint reflects this change (F. F. Brown et *al*., 1997). This work is based on data collected by application of NMR spectroscopy, subsequently subjected to multivariate statistical analysis in order to evaluate the metabolic changes as an effect of rearing conditions and storage time. In this chapter, three kinds of aquaculture practices have been considered, namely cage (**GB**), sink (**VS**) and lagoon (**VL**) environments. For each one, the effect of ice storage time on the metabolites evolution has been considered. In fact, samples captured, sacrificed and immediately stored under ice (t_0) have been compared to samples stored under ice for sixteen days after fish sacrifice (t_{16}).

4.2 MATERIALS AND METHODS

4.2.1 The Gilthead sea bream (Sparus aurata)

The Gilthead sea bream (*Sparus aurata*) is a fish of the bream family *Sparidae* (A. Davidson, 1972). It shows a typical oval body, rather deep and compressed of silvery grey colour. A large black blotch at origin of lateral line extends on upper margin of opercle (the uppermost and largest of the bones that form the gill cover). A golden frontal band between eyes edged by two dark areas and a dark longitudinal lines often present on sides of body (Figure 4.1).

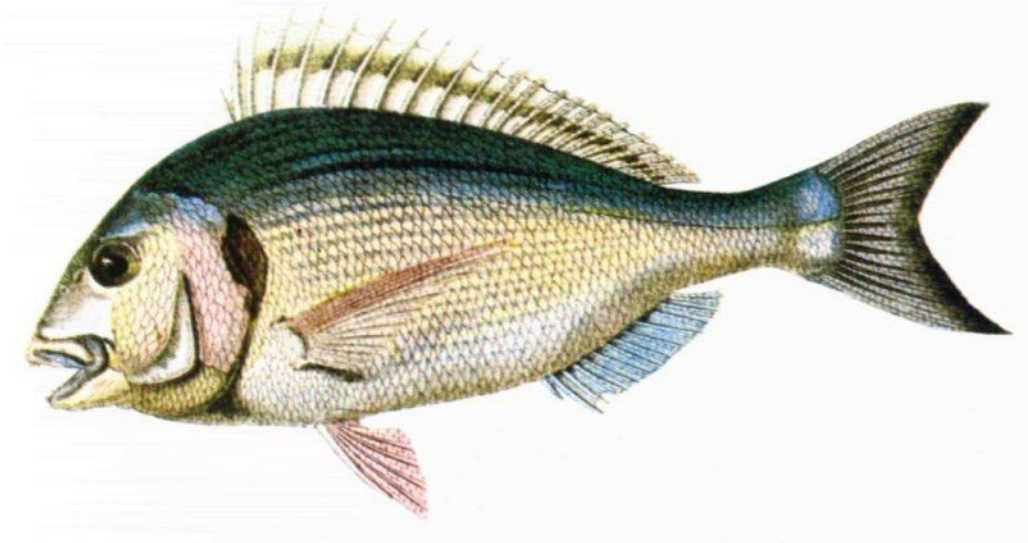


Figure 4.1: representation of Gilthead sea bream (*Sparus aurata*). The picture was adapted from http://www.fishingcy.com/

It grows to about 60 cm long and is generally considered the best-tasting of the breams and has given the whole family of Sparidae its name. Found in sea grass beds and sandy bottoms as well as in the surf zone commonly to depths of about 30 m, but adults may occur to 150 m depth. A sedentary fish, either solitary or in small aggregations. In spring, they often occur in brackish water coastal lagoons and estuaries (A., Davidson, 1972).

It is mainly carnivorous, feeding on shellfish, including mussels and oysters.

It can be found wildly in the Mediterranean Sea and the eastern coastal regions of the North Atlantic Ocean (Figure 4.2) (R. Froese et al., 2006), but up today, it represents one of the most important fishes in saline and hypersaline aquaculture, with a global production around 90,995 tonnes (L. Sola et al, 2007), especially in Mediterranean coastal lagoons where they are extensively farmed in lagoons, or intensively in tanks or cages and brackish/salt water ponds in the northern Adriatic *valli* in Italy and the Egyptian *hosha*.

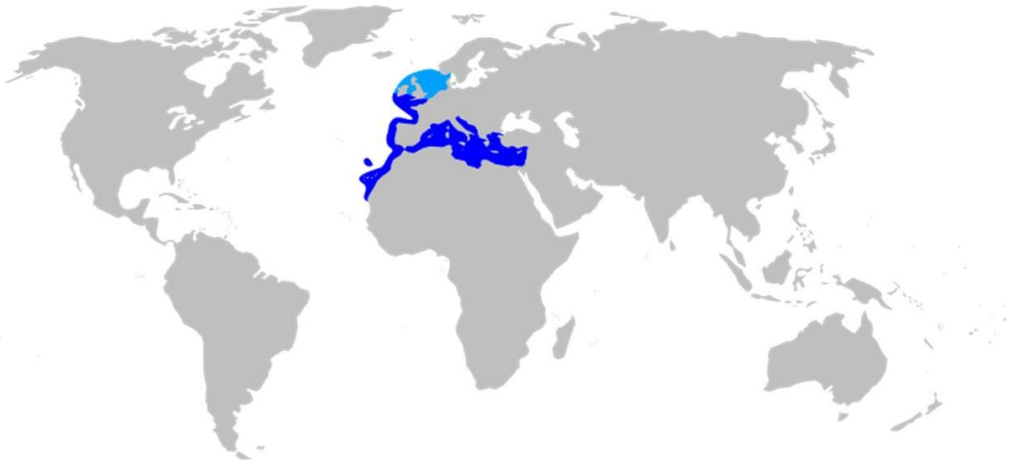


Figure 4.2: representation of Gilthead sea bream (*Sparus Aurata*). It occurs naturally in the Mediterranean and the Black Sea (rare), and in the Eastern Atlantic, from the British Isles, Strait of Gibraltar to Cape Verde and around the Canary Islands. The picture was adapted from http://www.fishingcy.com/

At present, most production is from intensive farming, whilst extensive farming, though still remains a traditional activity in some regions, has a very low impact on the market

4.2.2 Biological Material for the experimental work

Only the white right muscle coming from fish samples were analyzed. The fishes are the same used in a previous research work performed by the group of Aquaculture of the University of Bologna, leaded by Professor Anna Badiani.

Three groups of farmed Gilthead sea bream (*Sparus aurata*) were studied, each one coming from a different aquaculture system:

1. **sink for intensive fish farming:** developed at the laboratories of Aquaculture in Cesenatico (Italy) (Figure 4.3). All batches consisted of three, two-years old, individuals, fed with a commercial feed composed by 46% proteins and 21% lipids. Fishes were sacrificed in October 2006, when their body weight was ranging between 300 and 400 grams (commercial size). Before sacrifice, all fishes were fasting for 24 hours;

Figure 4.3: sinks for intensive fish farming located in the aquaculture's laboratories of the University of Bologna, located in Cesenatico, via dei Mille, 160, Italy (Figure is adapted from

2. **lagoon or "*valle*" for extensive fish farming:** developed at Valle Smarlacca S.r.l. in Ravenna (Italy) (Figure 4.4). All batches consisted of three, two-years old, individuals having the above mentioned commercial size. These fishes were not fed by a commercial feeding, but were allowed to feed with natural environmental resources, mainly constituted by *benthos* (K. Govindan, 2002);

Figure 4.4: lagoons or "valli" extensive fish farming (Figure and legend are adapted from www.venetoagricoltura.org)

3. **cage for extensive fish farming:** gilthead sea bream were fished from cages in Monfalcone (Italy) (Figure 4.5). Samples were excised from two-years old fishes fed with commercial feed (composed by a 45% protein and 16-18% lipids). In order to avoid *Winter disease*, due to low water temperature, fishes were fasting for two months before their capture, occurred in January 2006 (M. Manera et *al.*, 2003).

Figure 4.5 cages extensive fish farming (Figure and legend are adapted from http://cache.daylife.com)

After capture all samples underwent the same treatment and stored under ice before excision of samples that were finally stored in -80°C refrigerators.

4.2.3 Experimental design for NMR analysis

Fifty-four NMR samples were prepared from the extracts of the white right caudal muscle's tissue, excised from Gilthead sea bream (*Sparus aurata*), reared according to the three kinds of aquaculture (sink, lagoon and cage) systems already described, and at two different time of storage (t_0 and t_{16}).

In this work, two thesis consisting of different ice storage time (T_0 and T_{16}) included 9 animals, whose three for each aquaculture type (Figure 4.6).

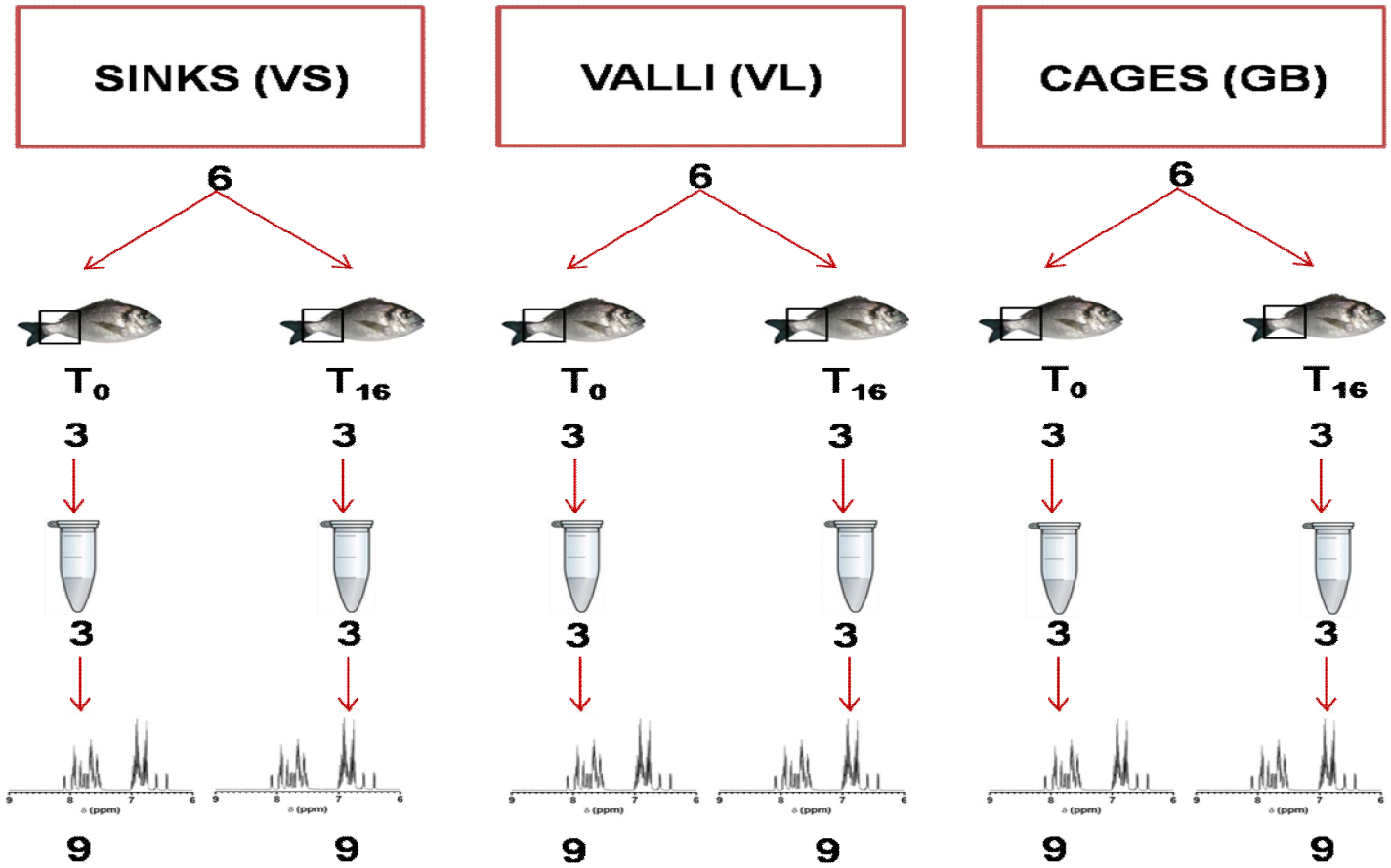


Figure 4.6: Gilthead sea bream sampling. For each time, 9 animals were required: three for each aquaculture thesis (**VS**: sink fish farming; **VL**: lagoons fish farming and **GB**: cages fish farming). At the end, a total of 54 NMR samples were prepared in order to perform the experimental work

4.2.4 Sample preparation and NMR data acquisition

All the steps were performed in cold room (4°C) in order to avoid degradation phenomena.

Four grams of white right caudal muscle from each sample were pulverized in a mortar under liquid nitrogen. Eight milliliters of 7% perchloric acid (i.e., 2 ml per gram of sample) were added to the muscle and mixed until complete homogenization is reached (I. S. Gribbestad et *al.*, 1994). The homogenized, transferred into a centrifuge tubes, was neutralized with 9M KOH to a final pH of 7.8. Samples were centrifuged to remove potassium perchlorate at 8000 rpm, for 20 minute at 4°C. The resulting supernatant was dispensed in 1 ml aliquot in Eppendorf tubes and stored at -80°C. The solutions were thawed just before NMR analysis.

Before Fourier transformation, a line broadening of 1.60 Hz was applied to each FID. All spectra were aligned each other by referring to the Creatine signal, the latter calibrated to 3.04 ppm with respect to TSP (0.00 ppm). MestReC Software (version 4.9.8.0) was used to phase and correct the baseline of all the spectra before performing the successive statistical analysis.

Some spectral regions were eliminated because not including signals above the noise or being affected by the presence of the solvent signals which can interfere in the statistical analysis. The removed ranges are chosen between 12.60 and 10.16 ppm, between 4.89 and 4.40 ppm, and between -0.52 and -3.38 ppm. In this way, spectral data points were reduced from 32 K (32768) to 18000 points.

The spectral information was further condensed by subdividing the spectra into 150 bins, each integrating 120 data points. The resulting binned spectra were saved as ASCII file for the subsequent statistical analysis.

4.2.5 Chemometric techniques

Except for several classic chemometric tools, described in Chapter 1 par. 1.2.4, as PCA and PLS-DA and applied on the processed matrix using home-made algorithms (Appendix A) written in the R program language (version 2.4.0), also advanced tool have been used for extracting relevant information out of the acquired NMR data and they are briefly described below:

Interval Partial Least Square (iPLS). *i*PLS (L. Nørgaard et *al*., 2000) is based on a recursive algorithm that builds PLS models of user-defined intervals in which the dataset has been split. Performances of these interval models are then compared to each other and with that of the model built using the full-length data (global model), by the *i*PLS plot. The method is particularly useful to identify important regions in NMR spectroscopic data (F. H. Larsen et *al.*, 2006) with large intensity differences and it therefore becomes a powerful tool for variables (regional) selection. In the present application, the entire dataset was subdivided into 100 segments of equal size. It is important to point out that this segmentation does not lead to a reduction of data, like binning does, but provides "an overall picture of the relevant information in different spectral subdivisions, focusing on important spectral regions" (L. Nørgaard et *al*., 2000). *i*PLS models were carried out in Matlab using the *i*PLS toolbox available at http://www.models.life.ku.dk/source/.

Extended Canonical Variable Analysis (ECVA) and Interval Extended Canonical Variable Analysis (iECVA). ECVA (L. Nørgaard et al. 2006) is a recent chemometric classification tool representing a new approach for grouping samples based on the standard Canonical Variates Analysis, but with a PLS engine underneath. It is able to cope with several different classes yielding powerful separations. Careful validation is required in order to avoid overfitting. *i*ECVA (L. Nørgaard et *al*., 2000) is an extension of the *i*PLS concept to ECVA designed to provide meaningful

information about which spectral regions hold the main relevance responsible for the separation among groups. In the present research work ECVA and *i*ECVA were applied for the classification of the three aquaculture groups and were carried out in Matlab using the ECVA toolbox available at http://www.models.life.ku.dk/source/. A Venetian blinds cross validation system was used in both the cases.

4.2.6 Chemicals

All chemical reagents, purchased from SIGMA-ALDRICH, Inc.(St. Louis, MO), were of analytical grade and are listed in Chapter 2.

4.3 RESULTS AND DISCUSSION

4.3.1 NMR Spectra and identification of samples' metabolic compounds

The ^{1}H-NMR spectrum of the perchloric acid extract of right white caudal muscle of Gilthead Sea bream recorded at 400 MHz, is shown in Figure 4.7. Appendix B reports the spectra of some representative samples used in this work, one for each category (t_0, t_{16} x VS, VL, GB), each spectrum being shown with the downfield, midfield and upfield spectral regions presented separately for a better vision. In order to visually appreciate signals belonging to minority species the upfield region (<2.9 ppm) and the downfield region (>5 ppm) were expanded on the vertical scale, and Table 4.1 lists the assignments of the NMR peaks. The assignment attained in this research work is the first one reported for an aqueous extract of Gilthead Sea bream, since only NMR spectra recorded on the lipid fraction have been published up to now (S. Rezzi et *al*., 2007).

Moreover, this study reports on the evolution of fish hydrosoluble

metabolites, mainly nucleotides, during storage under ice. Assignment of twenty-nine peaks in the NMR spectra were done by adding standard compounds to the samples and by comparing chemical shifts with previous published data collected on different fish species (I. Martinez et *al*., 2005).

Most intense signals belong to lactate (1.33 ppm), creatine and phosphocreatine (3.04 and 3.94 ppm, both) and trimethylamine n-oxide (TMAO; 3.27 ppm).

Table 4.1: list of metabolites assigned in perchloric acid extracts of Gilthead Sea bream. Their chemical shift are referred to TSP (0,00 ppm).

Peak	^{1}H Chemical Shift (ppm)	Compound
1	0.96	Leucine
2	1.33	Lactate
3	1.48	Alanine and Lysine
4	1.72	Lysine and Leucine
5	1.89	Lysine
6	2.89	TMA
7	3.04	Creatine and Phosphocreatine
8	3.12	Histidine
9	3.24	Taurine
10	3.27	TMAO
11	3.42	Taurine
12	3.56	Glycine
13	3.77	Glutamate
14	3.90	Betaine
15	3.94	Creatine and Phosphocreatine
16	4.12	Lactate
17	6.10	Inosine
18	6.14	Inosine MP
19A	7.00	Histidine in Anserine
19B	7.05	Histidine
20A	7.73	Histidine in Anserine
20B	7.77	Histidine
21	8.19	Hypoxantine
22	8.22	Hypoxantine
23	8.22	Inosine and Inosine MP
24	8.27	ADP
25	8.34	Inosine
26	8.46	Formiate
27	8.52	ADP
28	8.57	Inosine MP
29	8.61	ADP

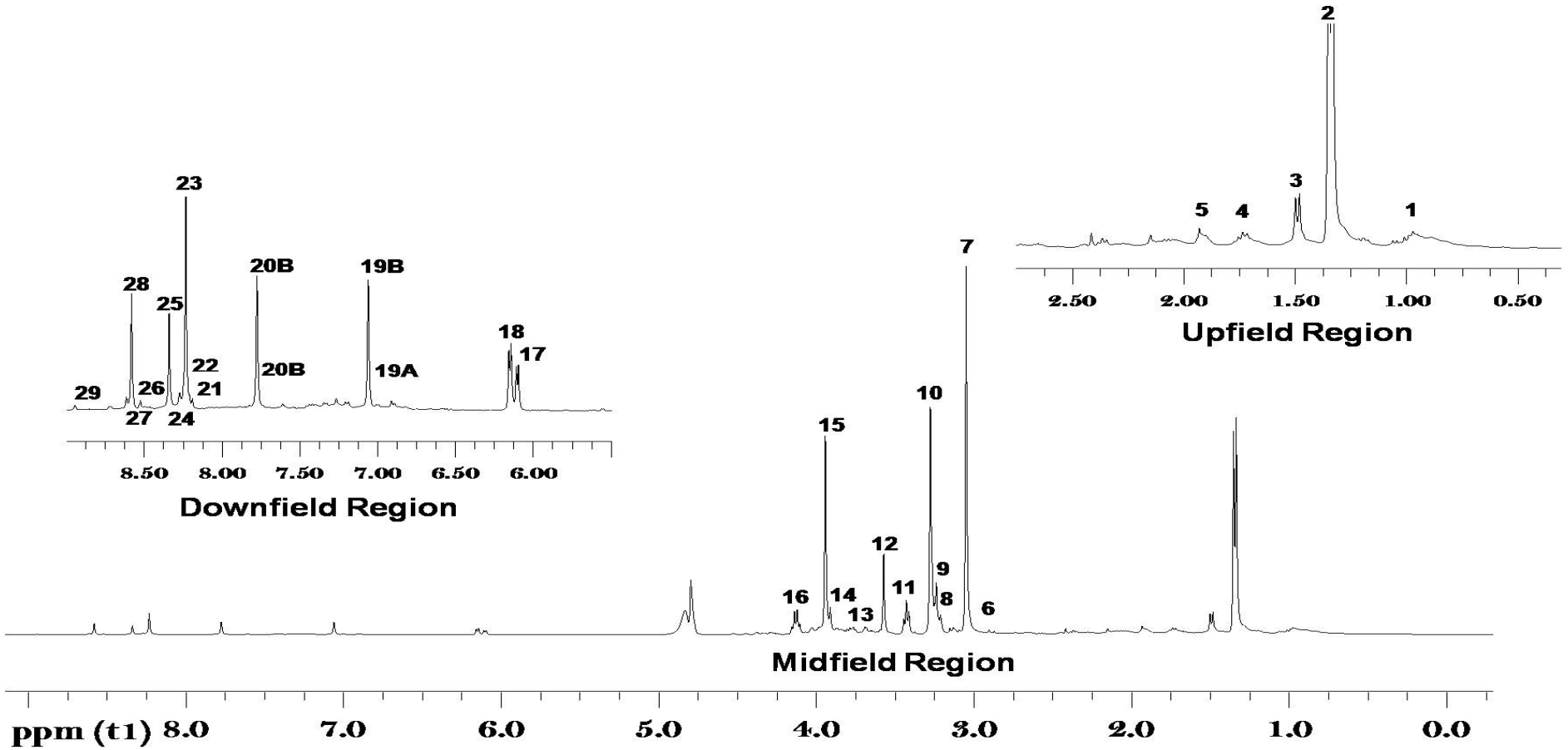


Figure 4.7: ^{1}H-NMR spectrum of perchloric acid extract of white right caudal muscle of Gilthead Sea bream (*Sparus aurata)* after 16 days of storage under ice. Downfield and upfield regions were expanded on the vertical scale in order to appreciate the presence of small signals. All labelled signals are referred to TSP (0 ppm) (the figure was performed using MestRec software)

White muscle represents the main muscle in fish and it is characterized by a low concentration of lipids and a high concentration of amino acids. In fact, the ^{1}H-NMR shows several signals belong to amino acids, including leucine, lysine, alanine, glycine, glutamine and histidine. Most of the amino acids are free although some might be bound to peptides. Large proteins are not visible because their slow tumbling gives origin to broad signals. Other signals like taurine (3.24 and 3.42 ppm), betaine (3.90 ppm) and nucleotides like ADP were assigned.

Creatine and phosphocreatine represent an important energy deposit in skeletal muscle. Phosphocreatine is used to anareobically generate ATP from ADP, forming creatine. This reversible reaction is catalyzed by the phosphocreatine kinase (Figure 4.8).

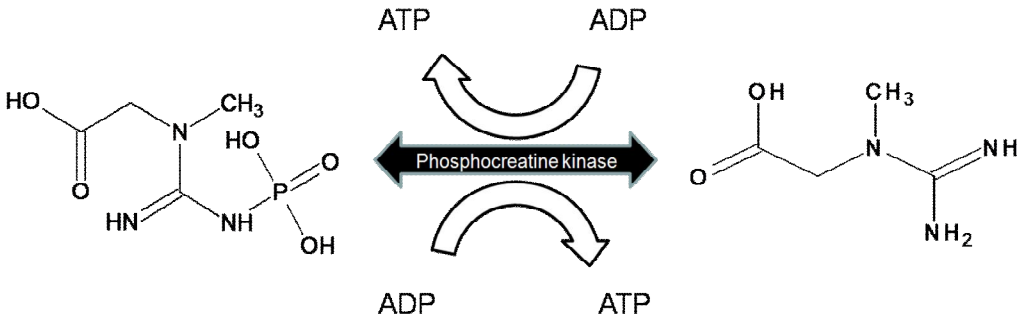


Figure 4.8: the phosphocreatine kinase reaction. Phosphocreatine kinase catalyzed the transfer of the phosphoryl group from phosphocreatine to ADP, originating ATP and creatine

In Figure 4.7, signals labelled with #7 an #15, at 3.04 and 3.94, respectively, are due mainly to creatine because of a rapid conversion of phosphocreatine to creatine occurring early during post mortem phase. Also taurine is important due to its relative abundance in fish metabolic profile. In fact, aquatic food are the major source of this metabolite, essential for the growth and it has been demonstrated that it is closely related to cholesterol metabolism (H. Yokogoshi et *al*., 1999; H. Yokogoshi et *al*., 2002). It is important to underline a low presence of TMA (2.89 ppm) in spectra acquired on samples stored for 16 days under ice. TMA is the degradation product of

TMAO and it confers the classical "fish malodor". Its presence is due to both bacterial activity and enzymatic reaction. We hypothesize that bacterial activity is minimized under ice storage; thus, the endogenous TMAOase enzyme is mainly responsible for the TMAO metabolization to TMA.

4.3.2 Data pre-treatment

Prior to chemometric data analysis, the data matrix (54x18K) consisting of raw NMR spectra, underwent to a harmonizing pre-treatment in order to reduce as low as possible the artefacts due to sample preparation and instrumental errors. The first step is a normalization of the data to minimize all unwanted sources of variation due to slightly different instrumental conditions as small differences in the tuning conditions of the spectrometer, imperfect baseline and phase corrections, and sample preparation artefacts due to dilution errors during sample preparation. Further pre-treatment, like alignment of the data and spectral data binning, proved to be necessary to reduce variations of chemical shift among different spectra, generated by small differences in the pH of the extract being analyzed. The binning application provides both the reduction of the pH effect on spectrum's chemical shift and the reduction of the data points to facilitate the chemometrics analysis.

4.3.3 Multivariate data analysis

Principal component analysis (PCA) (H. Hotelling, 1993) is a fundamental method in chemometrics and it has been chosen as the unsupervised method able to describe the total sample variance by projecting it in a condensed space. In PCA, the data collected on a set of samples is located in a space whose dimensions are defined as principal components (H. Winning et *al*., 2008). The first principal component describes most of the variance, the second principal component is the profile describing the second most of the variance orthogonal to the first one, and so on; at the end, the last components

describe negligible variance, mainly noise, and for this reason they are not taken in consideration. The principal components are described by scores and loadings. In the data set, loadings represent the condensed information on the NMR variables (spectral point or chemical shifts or ppm), allowing to find out which bins of the spectrum account for the main variation amongst the samples. On the other hand, scores summarized all the information on samples composition, providing information about the extent to which the spectral information, represented by the loadings, are high or low for each particular sample (H. Winning et *al.*, 2008). In this work, the PCA screening was applied using a correlation method in which each single data bin is standardised by centring its area integrals with respect to the mean value among all spectra (mean centring). Since the present study is aimed at evaluating the effect of the aquaculture system on the metabolic profile of fish and, simultaneously, the effect of the storage time on its biochemical profile, we firstly performed a global PC analysis on spectra recorded on all samples.

4.3.4 PCA of the different fish farming at both storage times

The PC plots, originated by the principal component analysis performed on the whole sample set, are shown in Figures 4.9 and 4.10. In the first case (Figure 4.9), the PC plot is the result of the analysis on the normalized binned spectral data. The combination of the first two PCs explains 80% of the total variance, and it is able to spontaneously cluster the samples according to the farming conditions. A good discrimination is observed among the three aquaculture groups, both for fresh samples (open coloured symbols) and for those stored under ice for 16 days (filled coloured symbols). By taking PC1 into account, a good separation is obtained between the t_0 samples group belonging to lagoon category (VL-T0) and all the other t_0 samples (VS-T0 and GB-T0). On the other hand, along PC2 sink farming system (VS-T0) is separated from lagoon and cage categories (VL-T0 and GB-T0). Thus, a combination of PC1 and PC2

allows discrimination of farming categories at t_0. The inspection of the PC plot gives rise also to similar consideration for samples at t_{16}. In fact, PC1 better discriminates the cage group from the other two categories, while PC2 separates sink farming conditions from the other two categories.

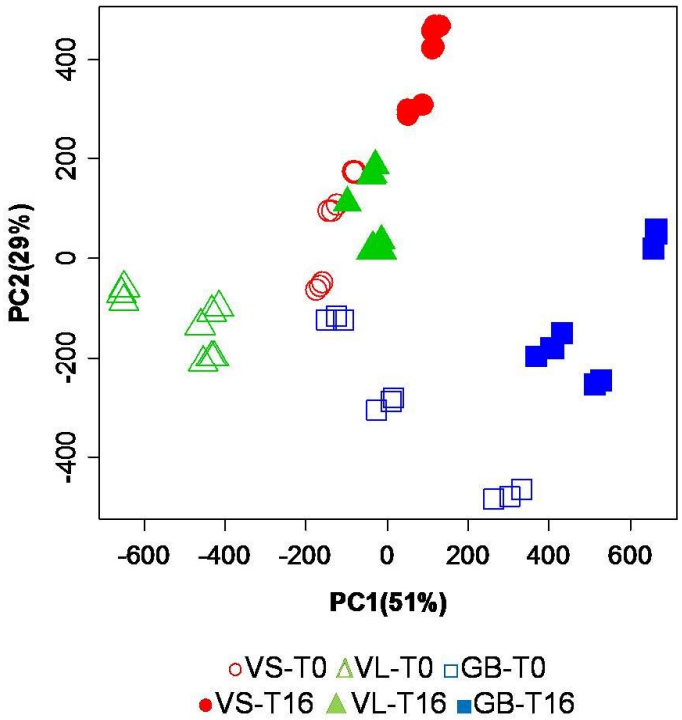


Figure 4.9: mean centered PCA of binned spectral data set, including all fish samples. According to the total variance, PC1 and PC2 explain 51% and 29% respectively. As it can be seen, all kinds of aquaculture systems are differentiated (the Figure was prepared by using the R program; see Appendix A)

The PC plot shown in Figure 4.9 gives rise to some considerations in terms of general quality of fish, as affected by storage time and farming conditions. First of all, time evolution of metabolites shifts fish scores towards a right-up direction (on the bisector of PC1-PC2 axes). The same dislocation is observed between lagoon and sink categories at t_0, as if the latter had the same metabolic composition of older samples of the first one. Conversely, cage category at t_0 is located in a position of the PC plot that may reflect a metabolic composition not related to time-dependent degradation. This result is compatible with the fact that they are also fed with different feeding. As the midfield region is characterized by the presence of much more

intense signals, compared to those falling in the other two regions (i.e., downfield and upfield ones) we decided to apply a regional normalization algorithm to the spectral dataset. The algorithm operates in such a way that the most intense peak in the upfield region is scaled to the same intensity of the most intense peak in the midfield region. The same operation is applied also to the downfield region, which is again scaled on the intensity of the highest peak in the midfield region. This regional scaling is applied because of the necessity to give the same importance to minor and major components, since there is no reason to assign greater nutritional value to substances only because are more abundant. Before application of the PC analysis to the regionally scaled data set, we performed a preliminary step consisting of the binning operation that transforms the spectra in a collection of 150 bins, each integrating 120 data points. The new regionally amplified binned data set underwent the PC analysis (Figure 4.10).

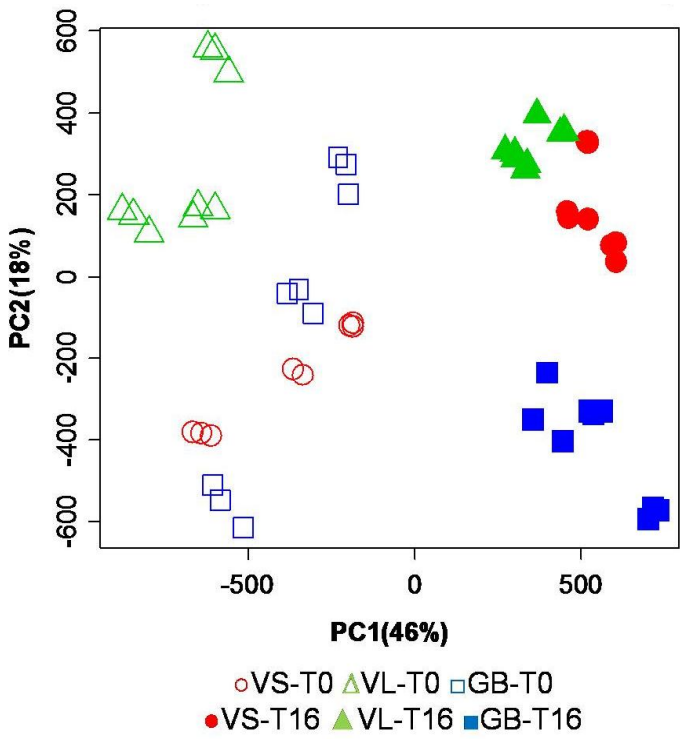


Figure 4.10: meancentered PCA of regionally amplified binned data set, including all fish samples. Downfield region and upfield region were amplified in order to have the same maximum value. According to the total variance, PC1 and PC2 explain 46% and 18% respectively. As it can be seen, all samples are differentiated on the base of time storage (the Figure was prepared by using R program; see Appendix A)

The new data transformation gives origin to a well defined time separation, clearly visible along PC1, where t_0 samples are located at negative values, while the t_{16} ones are in the positive side of the PC1 axis. Within these two major clusters, the distinction among different aquaculture systems for t_0 samples is less evident but, again, is confirmed for t_{16} samples belonging to the cage category.

As the comparison of ^{1}H-NMR spectra acquired one aqueous extracts of t_0 samples with those of t_{16} ones shows differences especially for the nucleotides content, whose signal are mainly located in the downfield spectral region, a different multivariate data analysis was applied to three different data sets, each containing separately, the three spectral regions (i.e., downfield, midfield and upfield). This application can highlight which part of the spectrum is mainly responsible for separation of clusters.

The PC plots in Figure 4.11 A is the result of the PCA on the downfield spectral region of all samples and points out that the signals here included (mainly belonging to nucleotides) allow separation of samples on the basis of the storage time. The total variance explained by the first two principal components generated on the downfield region is the lowest (74%) with respect to the variance expressed with the same number of PCs generated on the other two regions.

In Figure 4.11 B, the separation reflects what already seen whit the PCC performed on spectral data not regionally scaled, where the midfield region has the highest weight, and will not be further discussed. PC plot related to the upfield spectral region (Figure 4.11 C) points out that the samples are well differentiated on the basis of the farming conditions soon after sacrifice, while there is large overlap when analyzed after 16 days. It is worth noting here that such a region includes signals form organic acids, as lactic acid, and amino acids.

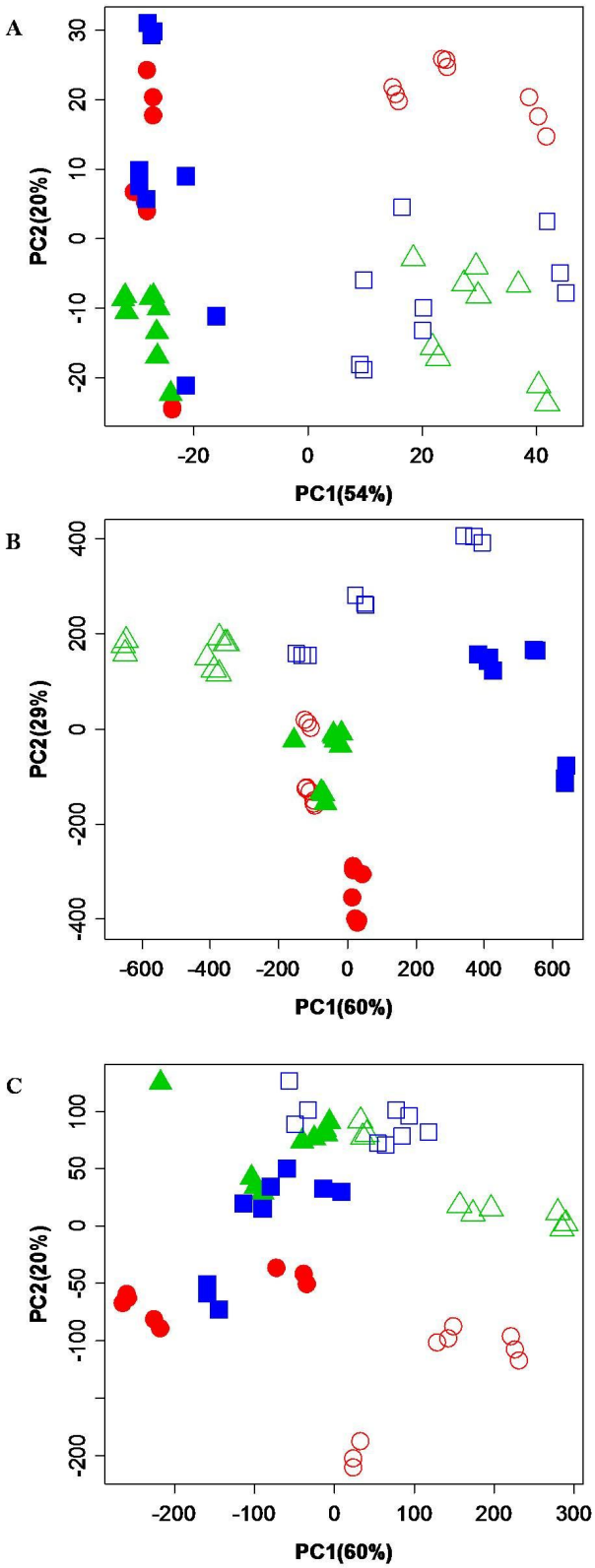


Figure 4.11: mean centered PCA of binned spectral data set divided into three main spectral regions: **A**) downfield; **B**) midfield and **C**) upfield region (the Figures were prepared by using R program; see Appendix A)

4.3.5 PCA applied to the downfield spectral region

More details arise from the analysis of PC loadings, by which it is possible to find out the bins of the spectrum, and then the groups of compounds, representing the main source of variation amongst categories. Thus, the representation reports the weight with which each bin influences the position of the sample along PC1 or PC2 axes. As above discussed with the preliminary PC analysis, mainly downfield and midfield regions may include spectral bins explaining the reasons of the separation. Figure 4.12 shows the loadings plot relative to the downfield region.

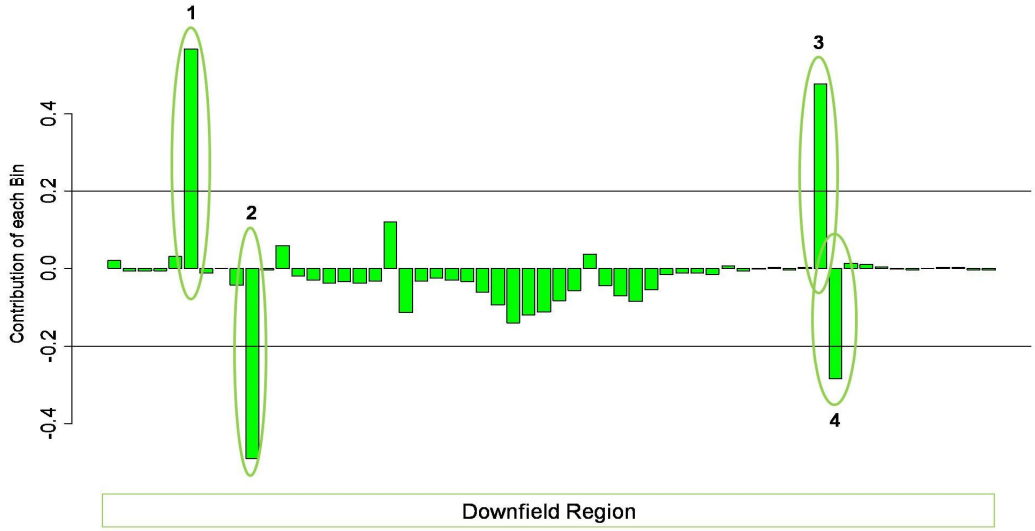


Figure 4.12: Downfield region loadings plot along PC1. Each bar corresponds to the importance of each spectral bin in determining the PC1 score. Both y=a and y=-b lines indicate, for convenience, the threshold of the absolute value of loadings adopted to choose important bins. Bins labelled with #2 and #4 belong to the same metabolite (IMP) as bins #1 and #3 to INO (the Figure was prepared by using the R program; see Appendix A)

According to the previous assignments (Table 4.1), bins labelled with #1 and #3, corresponding to the 6th and 47th bin in the whole spectrum, include the two signals of inosine monophosphate (IMP) falling at 8.57 ppm and 6.14 ppm. Bins labelled with #2 and #4, (corresponding to the 10th and 48th bin), include the inosine's (INO) signals falling at 8.34 and 6.14 ppm.

PC loadings have the ability to point to spectral features possessing high discriminative power. However, the actual amount of substances

responsible of differentiation is deducible by the integration of the signals included in the evidenced bin and by comparing such integrals among the spectra of different samples.

For this purpose, the areas of the bins labelled #1 and #2 in Figure 4.12, averaged separately on all t_0 and t_{16} samples, are reported in Figure 4.13 A and B, respectively. As bins #1 and #3 belong to the same metabolite (IMP), as well as bins #2 and #4 belongs to inosine, only average values of integrals relative to bin #1 and #2 are shown.

Inspection of Figure 4.13 points out that the IMP concentration is inversely proportional to the amount of INO, and there is a significant decrease of IMP concentration during ice storage, parallelized by the same increase exerted by INO. This is the obvious consequence of the fact that INO is a catabolite of IMP.

What cannot be trivially evinced without the inspection of the results is that the total amount of IMP and INO, together, remain constant during ice storage, because all the oxidative deamination of AMP, with formation of the hypoxanthine moiety of the nucleoside, is completely terminated just few hours after fish death (M. T. Veciana-Nogues et *al.*, 1996). The catabolism of ATP is controlled by native enzymes, such as those involved in the adenosine triphosphate (ATP) breakdown process (J. M. Kennish et *al.*, 1986), quickly occurring post mortem since there is no more reformation of ATP. For this reason only ADP is detected, at low concentration, in the NMR spectra, while AMP and ATP are not detectable even in the t_0 samples.

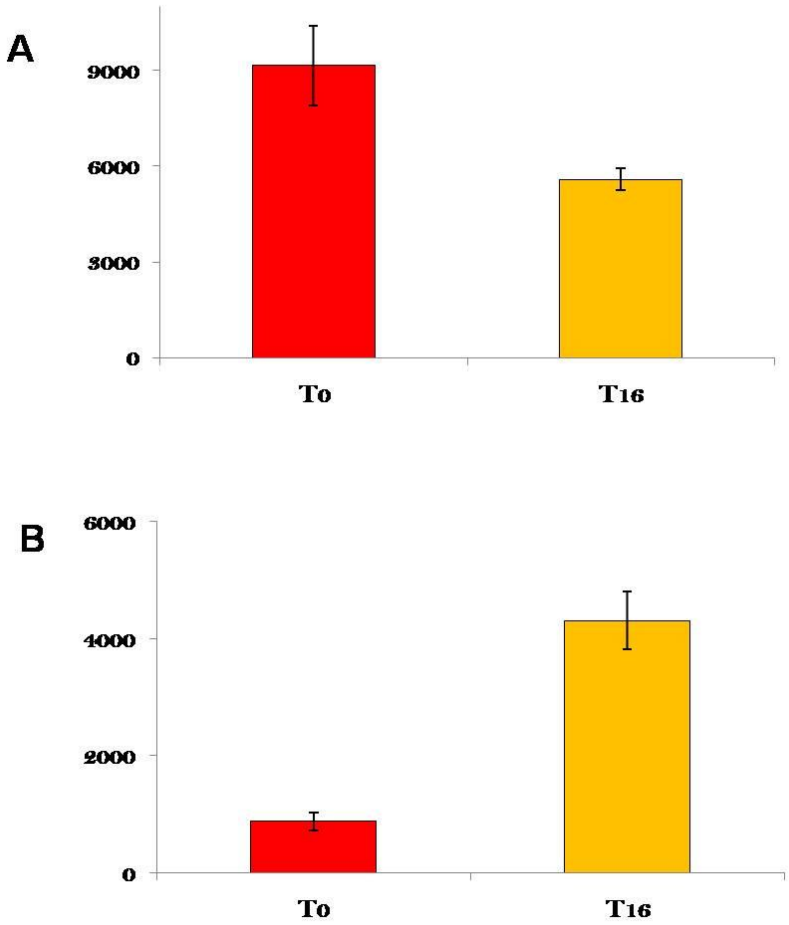


Figure 4.13: A) the histogram shows the absolute area of the bin labelled with #1 in PC1 loadings plot corresponding to the IMP integral peak; as it can be seen, immediately after the fishes' sacrifice, IMP content is higher and after 16 days most of the IMP is converted into INO as shown in figure **B)** that represents the absolute area of the bin labelled with #2 in PC1 loadings plot. It corresponds to INO integral peak. Standard errors are also shown as black lines on the top of each bar

IMP/INO ratio becomes a good indicator of fish freshness when it is stored under ice, since the enzymatic conversion of TMAO to TMA does not occur so rapidly at such temperature. The reason of change in the IMP/INO ratio, occurring during ice storage of fish tissues, is to be attributed to autolytic reactions. In fact, it is reported that IMP increases sharply around 5-24 h after death and decreases gradually by conversion to inosine (INO) (T. Saito et *al*., 1959; I. S. Park et *al*., 1999). IMP is known to contribute to the pleasant flavour of fresh fish.

The pathway of ATP catabolism in fish muscle has been extensively

documented as a degradative sequence to adenosine diphosphate (ADP), adenosine monophosphate (AMP), inosine monophosphate (IMP) and inosine (INO) (Figure 4.14).

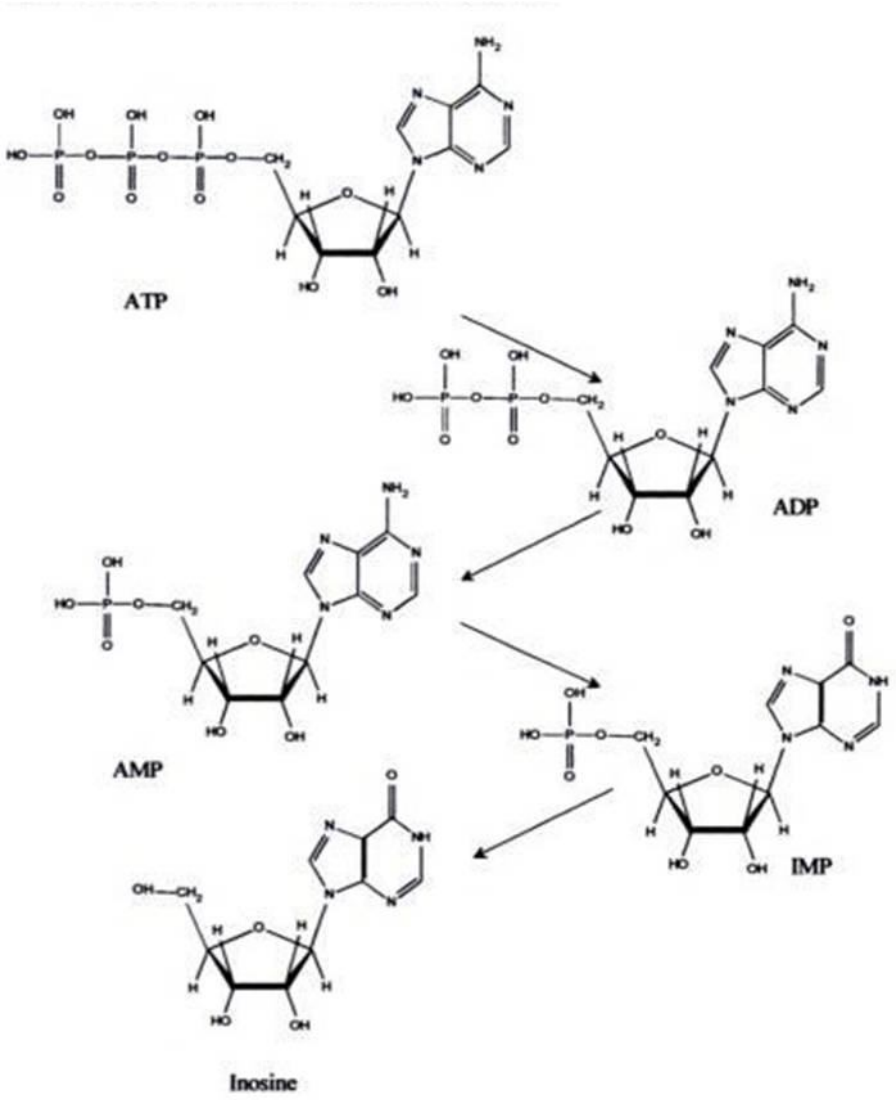


Figure 4.14: ATP enzymatic catabolism pathway in fish muscle. The post-mortem degradation products are important parameters for fish quality assessment (B. Sitter et *al*., 1999)

After time, a further conversion of INO to hypoxanthine (HX) occurs, which is strictly related and involved in the progressive loss of desirable fresh fish flavour (M. D. Huynh et *al*., 1992; F. Özoğul et *al*., 2002).

On account of this, nucleoside degradation products have been widely

used as indicators of storage age or freshness (R. Mendes et *al.*, 2001).

It is worth noting here that the rates and patterns of changes in the level of adenine nucleotides, and their related compounds, during storage time differ according to fish species (J. N. Dingle et *al.*, 1971), and, within the same species, also on storage conditions (H. Uchiyama et *al.*, 1970) and muscle types (M. Murata et *al.*, 1986; F. A. Vazquez-Ortiz et *al.*, 1997; J. M. Ryder et *al.*, 1984).

4.3.6 PCA applied on the midfield spectral region

According to Figure 4.10 B, the metabolites allowing the separation of samples on the basis of farming system lay in the midfield region. By looking at the PC plot, it appears that along PC1, lagoon samples (VL) are differentiated from the other two categories, at t_0, while are the cage samples (GB) to be separated from the other two groups when longer storage time (t_{16}) is considered.

PC2 gives a complementary discrimination so that sink system is differentiated from the other two groups, at both time of storage. Thus, a combination of PC1 and PC2 allows discrimination among farming system, when the same time of storage is considered.

PC1 loadings are illustrated in Figure 4.15 whilst PC2 loadings are reported in Figures 4.16. For both PCs, bins labelled with #1, #2, #3 and #4 (corresponding to the 77^{th} , 83^{rd} , 88^{th} , and 92^{nd} spectral bins) include metabolites involved in separation among the three farming systems.

On the basis of the previous peaks assignments, the corresponding metabolites are creatine and phosphocreatine (bin #1 and #4, centered at 3.94 ppm and 3.04 ppm, respectively), glycine (bin #2, centered at 3.56 ppm) and TMAO (bin #3, centrered at 3.27 ppm), (Figure 4.18).

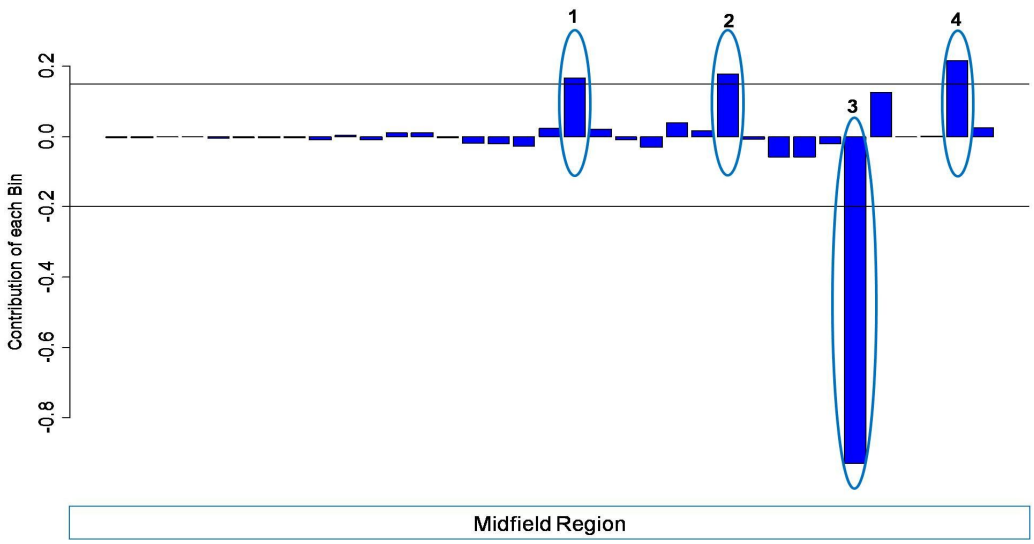


Figure 4.15: Bins located in the midfield region, responsible for samples separation along PC1 (the Figure was prepared by using R program; see Appendix A)

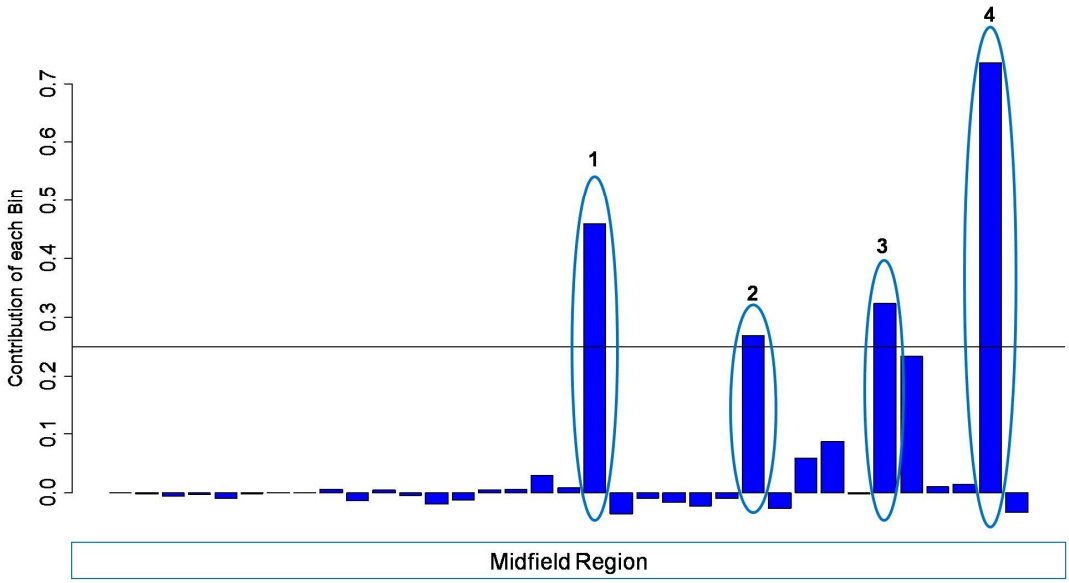


Figure 4.16: Bins located in the midfield region, responsible for samples separation along PC2 (the Figure was prepared by using R program; see Appendix A)

Since PC loadings are proportional to the importance of a bin to explain variance along the principal components, the further step consists of analyzing the integral of the highlighted bins in order to appreciate their change of concentration by passing from a category to another one. The average integrals of bins #1, #2 and #3, within each category, are reported in Figure 4.17, 4.18 and 4.19, respectively. Integrals of bin #4 are omitted since they refer to the

same metabolite (creatine) included in bin #1.

Creatine or phosphocreatine (having coincident signals in the ^{1}H-NMR spectrum) are able to discriminate among cage farming system and the other two groups, at both storage time, as seen by PC plot and PC2 loadings shown in Figure 4.11 B and 4.16, respectively. The average areas, as well as their standard deviations, for the creatine/phosphorcreatine signals are reported in Figure 4.17. Here is confirmed that the amount of creatine/phosphocreatine is higher and higher by passing along the sequence sink-lagoon-cage farming systems, and the same order remains constant after 16 days of time storage. The latter group, however, has lower content of this metabolite, suggesting a moderate degradation of creatine during storage under ice.

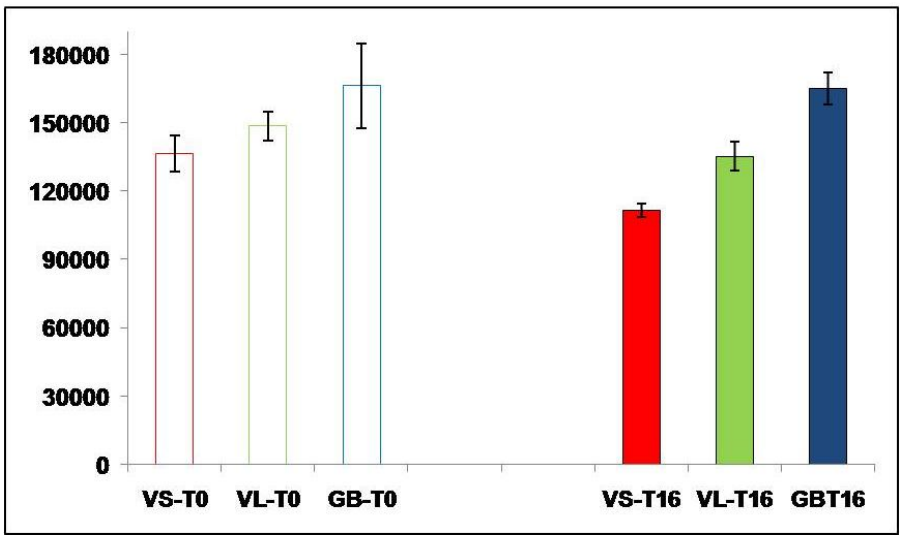


Figure 4.17: Average integral area of the bin labelled with #1 in the PC2 loadings plot reported in Figure corresponding to the creatine/phosphocreatine integral peak. Standard errors are also shown as black lines on the top of each bar

It has been reported that during post-mortem phase, also under ice storage, the amount of creatine phosphate decreases less rapidly then ATP, even though the latter starts to decrease only when its concentration equals the amount of creatine phosphate (S. Watabe et *al*, 1990). Since signals of creatine and creatine phosphate are coincident in the ^{1}H-NMR spectrum, it is not possible to quantify the

relative amount of each species. However, it is reported that creatine phosphate donates its phosphate group to the ADP to regenerate ATP (Figure 4.8), and converting itself in creatine (M. Wyss et *al.*, 2000). Since the amount of adenine containing nucleotides is very low in the analyzed fishes, it is presumable that all the creatine present in the samples is not bound to phosphate, according to the hypothesis that the muscles have already consumed their energy sources early during post-mortem.

Particularly interesting is the TMAO (trimethylamine oxide) changes occurring during storage (Figure 4.18). The relative bin has got a high weight along PC1 (Figure 4.15) and is responsible for the separation due to both time of storage and aquaculture system.

Usually, its amount is taken as parameter for evaluating the storage time; in fact, TMAO is converted to TMA (trimethylamine) by endogenous enzymes and later by bacteria when microbial activity begins (N. Church, 1998).

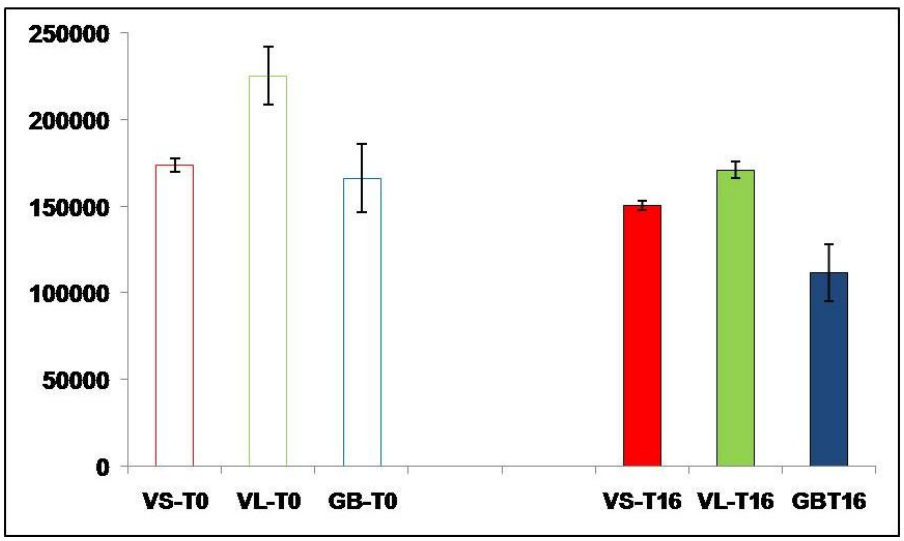


Figure 4.18: Absolute area of the bin labelled #2 in the PC1 loadings plot of Figure 4.15, corresponding to the TMAO integral peak. Standard errors are also shown as black lines on the top of each bar

According to Kyrana et *al.*, initial levels of TMAO are related to the composition of feed for reared fish, and this argument can explain the reasons of its concentration differences among groups (V. R. Kyrana

et *al.*, 2002). Furthermore, the values of TMAO decreased during storage in ice, maintaining anyway the same time difference among the three fish farming systems.

No significant differences in TMA levels have been detected between gropus, so that the small increase in TMA during the shelf life of Gilthead Sea bream prevent this parameter to be used as a freshness indicator for this species. Previous work reported the same results (T. Civera et *al.*, 1995; A. Huidobro et *al.*, 2001).

About glycine (bin #2), its integral area is reported in Figures 4.19. Except for samples farmed in sinks, the tendency of this free amino acid is to slightly increase during post-mortem under ice storage. The total amount of glycine, higher than the area of other amino acids, has been proposed to derive from the acidic hydrolysis consequent to the perchloric acid extraction step, that forms glycine starting from purine nucleotides and their derivates (ATP, ADP, AMP and INO, J. Arakaki et *al.*, 1974). However, this hypothesis does not explain the higher content of glycine presented by fishes reared in cage farming system, compared to the other two categories.

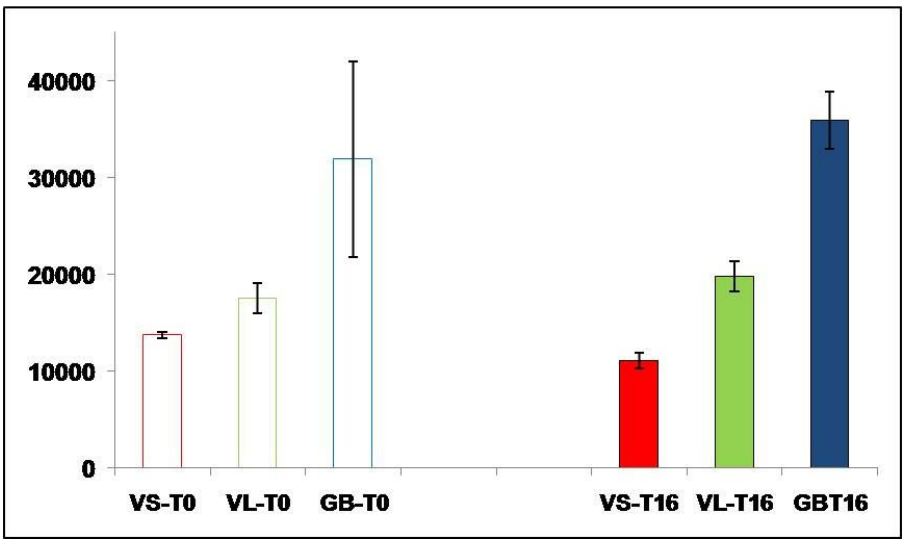


Figure 4.19: figure shows the absolute area of the bin labelled with #3 in PC1 loadings plot corresponding to the glycine integral peak. Standard errors are also shown as black lines on the top of each bar

For the latter observation, the explanation is attributable to the fish feeding activity.

In fact, fish starvation considerably reduces the uptake of that amino acid, and for this reason, the glycine ratio has been proposed to be assumed as an index related to the general nutritional state, instead of being considered a freshness parameter (C. M. Smagula et *al.*, 1980).

All results above discussed rely on binned spectral data. Binning application can solve most of the chemical shift problem due to pH variations but, at the same time, it generates overlap problems. In fact, many signals are collected in the same bin, covering 0.06 ppm, so that they are characterized by a broad range of intensities.

For this reason, some signals, although belonging to molecules with concentration highly depending on the category of fish farming and/or time of storage, can be hidden by other more intense signals that, in turn, belong to molecules not meaningful for the discrimination purposes.

To avoid this problem, raw matrix data set (54x18000) underwent to multivariate data analysis after a coshift application that eliminates the pH problems.

4.3.7 Multivariate data analysis of ^{1}H NMR spectra without the binning operation

After the referring of all the spectra to creatine signal, several shifts in peak positions were still observed, especially for those metabolites hardly dependent on the pH differences.

A correlation optimized shifting (Coshift) was applied to those intervals in which signals were not aligned, obtaining a considerable reduction of shifting (Figure 4.20).

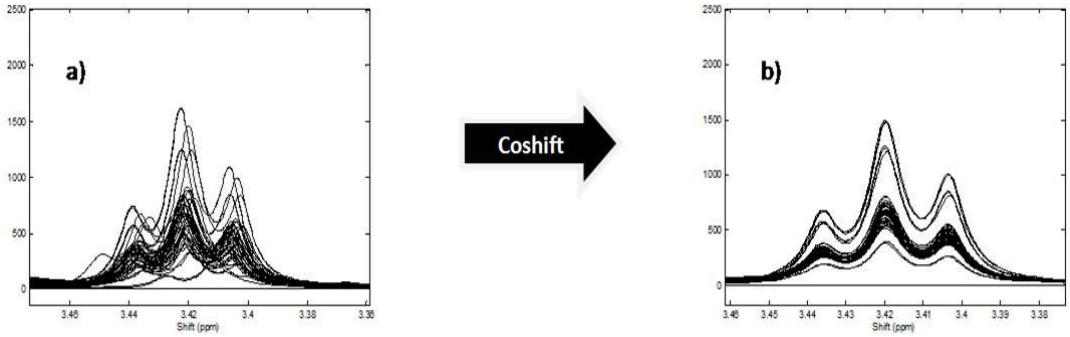


Figure 4.20: (a) zoom of taurine signal in uncorrected ^{1}H-NMR spectra to show the chemical shift problem due to the effect of lower pH variation on sensible metabolites and (b) the same zoom after coshifting optimization

After the application of the coshift method, it has been first considered the PCA applied on the whole spectral dataset, in order to evaluate the possibility of a discrimination based on the ice storage. From Figure 4.21 is possible to observe that PC1 and PC2 do not give a good separation.

For this reason, the subsequent step was the analysis of the loading in order to evaluate which part of the spectrum is able to discriminate all samples according to the time of ice storage.

The loadings analysis reported in Figures 4.22 A and 4.22 B suggests that the main discriminative spectral region is the downfield one, from 10.5 to 5.5 ppm, in which it is possible to find out several discriminative signals belonging to the nucleotide metabolites.

A PC analysis on this region exhibits a significant ability to separate samples according to time storage with an explained variance increased from 50.97% for PC1 in the whole spectral data set analysis to 61.32% considering only the downfield region (Figure 4.23)

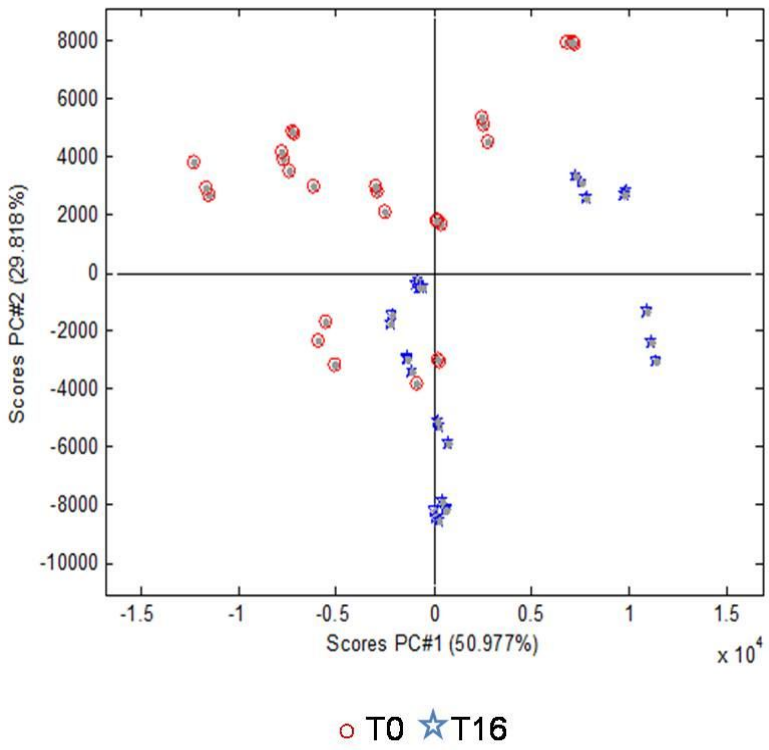


Figure 4.21: PCA scores plot of coshifted data set consisting of 54 ^{1}H-NMR sample spectra. PCA is not able to separate samples according to time storage (PCA was performed by using Latentix program)

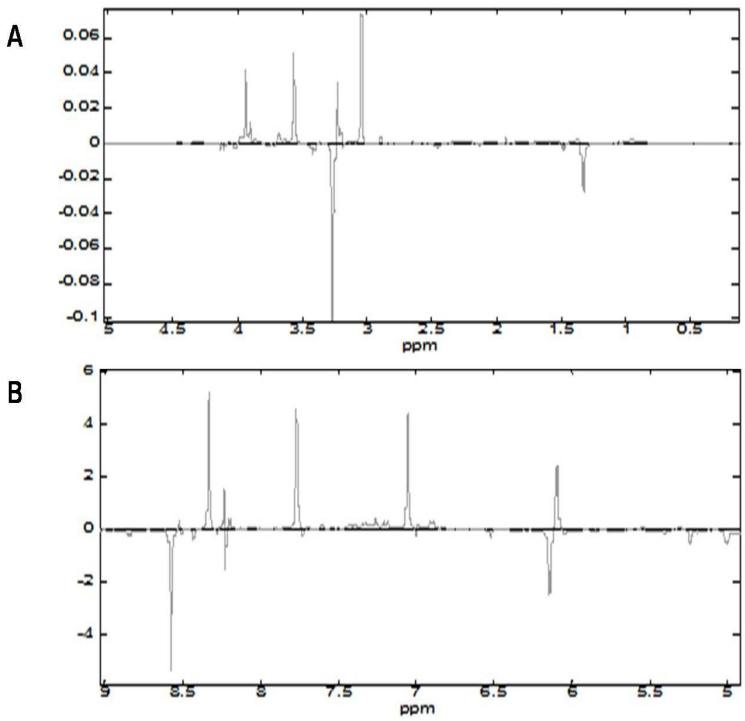


Figure 4.22: A) loadings plot of midfield and upfield region and **B**) loadings plot of downfield region. The most interesting region able to discriminate according to time of ice storage is the downfield one (PCA was prepared by using Latentix program)

.

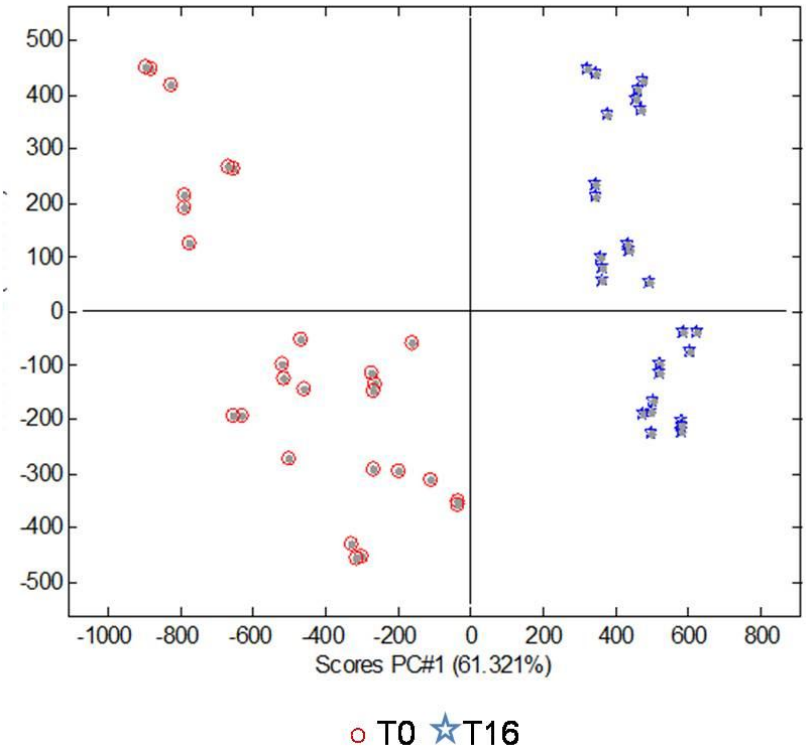


Figure 4.23: PCA scores plot on coshifted data set of 54 ^{1}H-NMR fish spectra considering only the midfield region. The analysis on this region exhibits a significant ability to separate samples according to time storage pushing samples at t_{16} on the positive side of PC1 axis and samples at t_0 towards the negative side of the same axis (PCA was prepared by using Latentix program)

The examination of PC1 loadings, explained in Figure 4.24, allows the identification of the metabolites responsible of the separation.

The positive values of the peaks in the loadings plot are related to metabolites with a high intensity in samples clustered along positive PC1. These peaks, clearly identifiable by plotting the opportune range of the spectrum, belong to the inosine, whose concentration increases during conservation (Figure 4.25).

On the contrary, at t_0 inosine monophosphate becomes to be the main metabolite in samples.

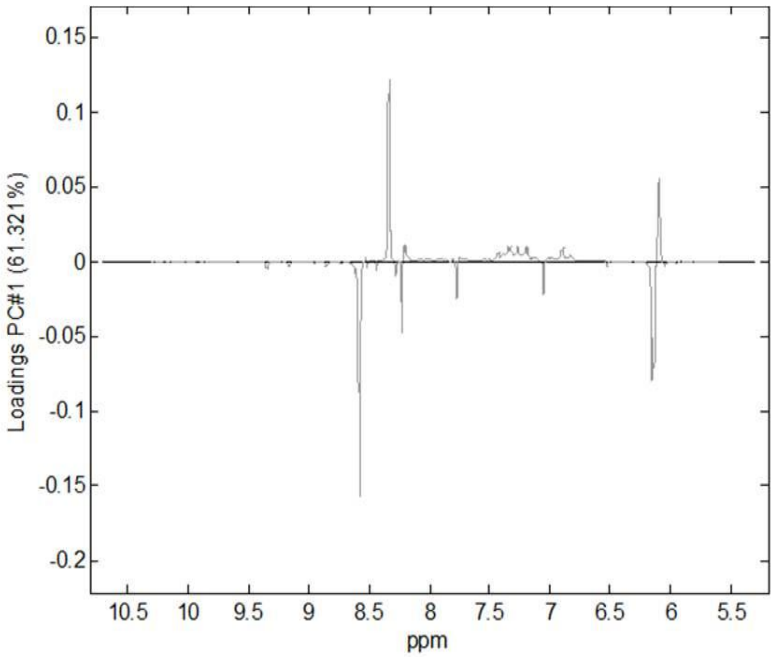


Figure 4.24: PCA loadings plot of downfield region

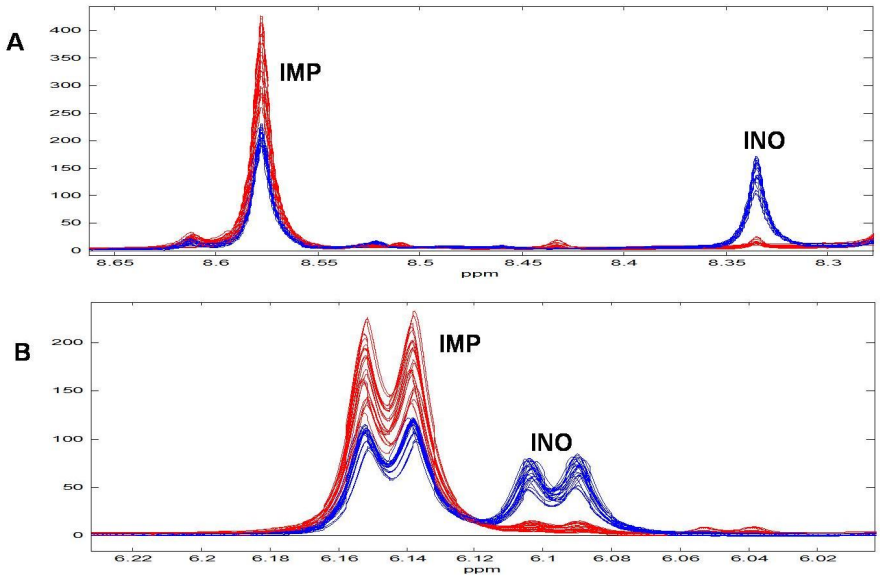


Figure 4.25: A) and **B**) signals belonging to the two metabolites responsible to the separation of samples according to time storage. Inosine increases during storage, whilst inosine monophosphate decreases for its completely metabolization into inosine

As above discussed in the paragraph describing the binned spectral data, the IMP/INO ratio becomes a good indicator of fish freshness. The same approach analysis has been used to verify the possibility for a classification of samples according to the aquaculture system.

Differently from the application of PCA on binned spectra, in the point data set PCA was applied separately on a first set of 27 samples at t_0 and on a second set, including the same number of samples at t_{16} (Figure 4.26 A and B).

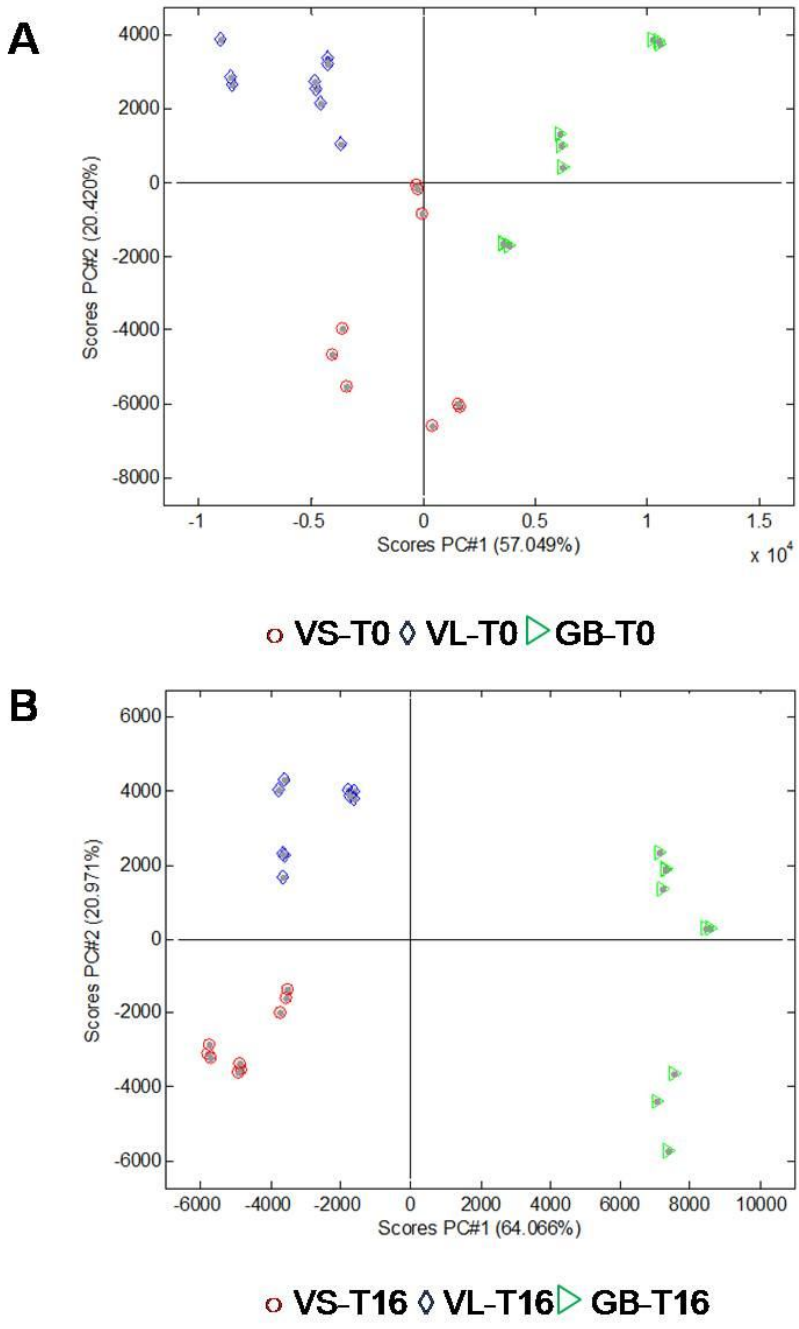


Figure 4.26: A) PCA scores plot of 27 ^{1}H-NMR fish spectra at t_0. PC1 shows a good separation only for samples bred in cages (GB) and **B**) PCA scores Plot of 27 ^{1}H-NMR fish spectra at t_{16}. Even in this plot, PC1 shows a good separation only for samples bred in cages (GB), whilst along PC2 lagoon (VL) and sink (VS) aquaculture systems are good defined

PCA is able to classify samples according to the different aquaculture system both at t_0, and, better, at t_{16} It is worth noting here that the first two principal components in PC plot, after 16 days, discriminate all categories with components that together explain 85% of the total variance. For this reason PC scores are able to condensate, with only two parameters, the whole different composition of fish due to their farming system.

To improve separation among aquaculture, and to verify the possibility of a better separation especially at t_0, a PLS2-DA, with a relative cross validation, has been applied on the spectral dataset.

The cross validation results (Table 4.2) confirms that only samples belonging to cages aquaculture system, at both storage time, can be predicted from the model.

Table 4.2: PLS2-DA Calibration model based on the cross validation

	t_0			t_{16}		
	VS	**VL**	**GB**	**VS**	**VL**	**GB**
RMSE	0,2486	0,2405	0,0846	0,361	0,3962	0,09396
R2	0,749	0.7837	0,971	0,4666	0,4151	0.9608

RMSE= Root Mean Squared Error of Cross Validation, R2= squared correlation coefficient

The best model representation (Figure 4.27 A and B) is obtained at t_{16} in which only two factors, or latent variables, are used to describe it, obtaining a squared correlation coefficient of predicted data close to one.

Conversely, the same correlation coefficient for the predictive model is obtained at t_0 by using 5 factors or latent variables.

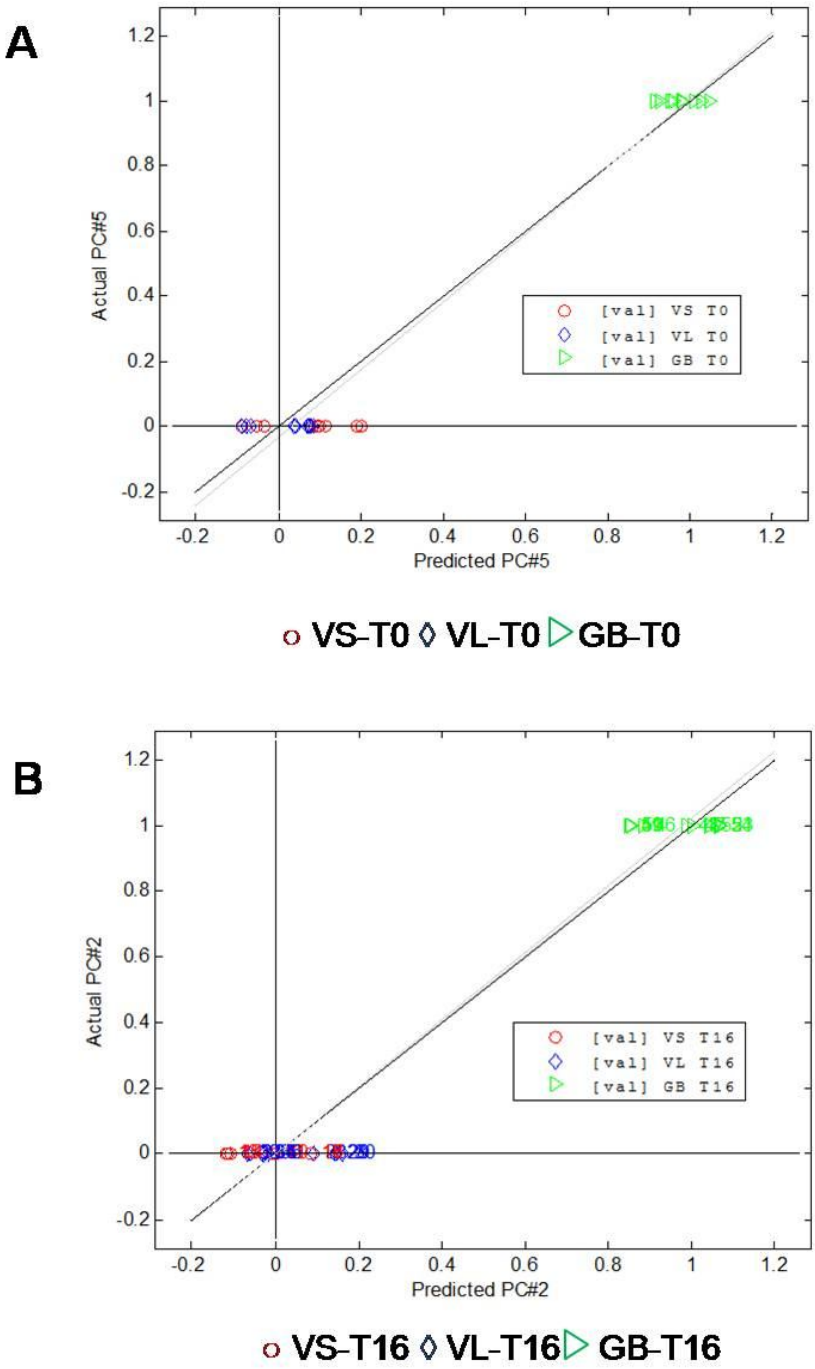


Figure 4.27: A) PLS2-DA prediction model at t_0. Cages samples are good predicted with 5 factors or latent variables and **B**) PLS2-DA prediction model at t_{16}. The result is the same but only with 2 factors or latent variables

Finally, an extended version of the canonical variates analysis (ECVA) (L. Nørgaard et *al.*, 2007) has been used, analyzing the three aquaculture systems.

In Figure 4.28 A the canonical variates of ECVA for samples,

calculated using a cross validation method, at t_0 are shown. At least two canonical directions have been used, i.e. one less than the number of the aquaculture systems.

The number of misclassified samples, according to the number of canonical variates, is shown by a barplot in Figure 4.28 B.

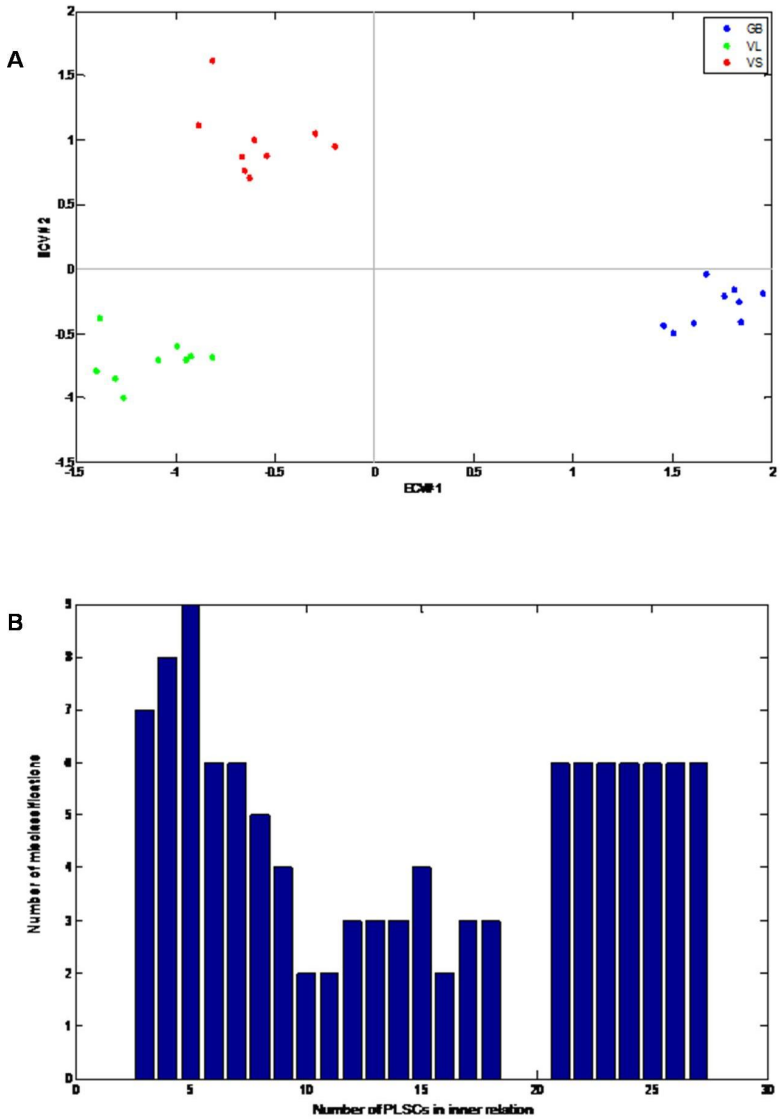


Figure 4.28: A) ECV#1 and ECV#2 scores plot for 27 ^{1}H-NMR fish spectra at t_0 and **B)** number of misclassifications according to the number of PLS components used. At t_0, ten PLScs are used in order to have 2 misclassified samples.

Compared to the PCA plot of Figure 4.26 A, the ECV analysis shows a deeper discriminant solution for samples at t_0.

The same representation can be done at t_{16}. Also in this case the

number of canonical directions that are necessary for separation is at least two and, compared to the corresponding PC plot, the separation is more pronounced, especially along ECV#1. Again, the number of misclassified samples decreases to 0 by using 4 PLS components (Figure 4.29).

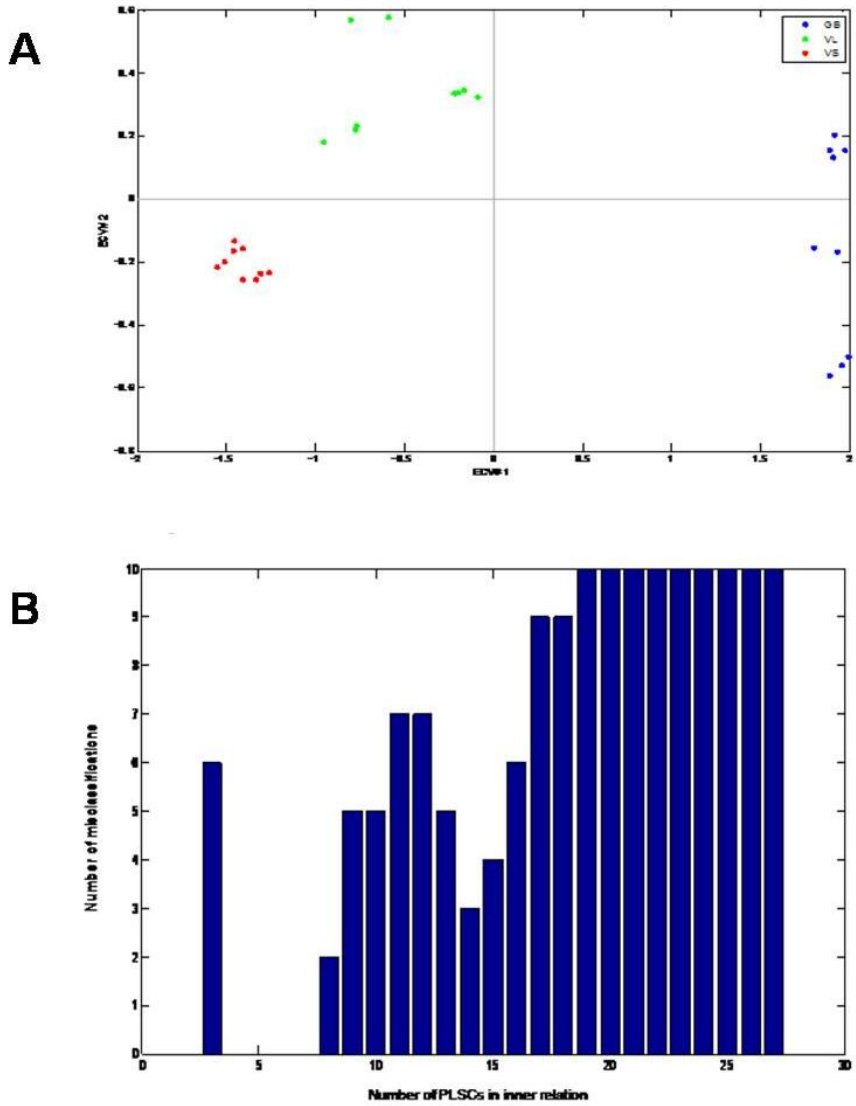


Figure 4.29: A) ECV#1 and ECV#2 scores plot for 27 ^{1}H-NMR fish spectra at t_{16} and **B)** number of misclassifications according to the number of PLS components used. At t_{16}, four PLScs are used in order to have 0 misclassified samples

Another way to graphically represent the ECV analysis is the iECVA. Similar to iPLS, the iECVA approach tries to find the regions in the ^{1}H-NMR spectrum that best describes the whole model.

In this way, it is possible to remove, from the spectra, that information (i.e. regions) not significant for the description of variance, thus reducing the number of data points able to describe the entire model.

In this work, iPLS (Figure 4.30) at t_{16} has been compared to iECVA (Figure 4.31) at the same time, because only after 16 days of ice storage the main differences among groups come out (L. Nørgaard et *al.*, 2000). The spectrum has been divided into intervals. From the iPLS plot, emerges that the most discriminative interval is the #62. It is possible to affirm that this interval possess the ability to better describe the entire dataset in the same manner as the whole spectrum is able to do. In fact, only two factors (latent variables) are used in the PLS model, whilst five of them are necessary if we consider all the intervals (entire spectrum dataset). On the contrary, the iECVA plot shows several intervals able to describe the model with same number of factors, misclassifying 0 samples. Four intervals are best discriminant: 25, 50, 62 and 82.

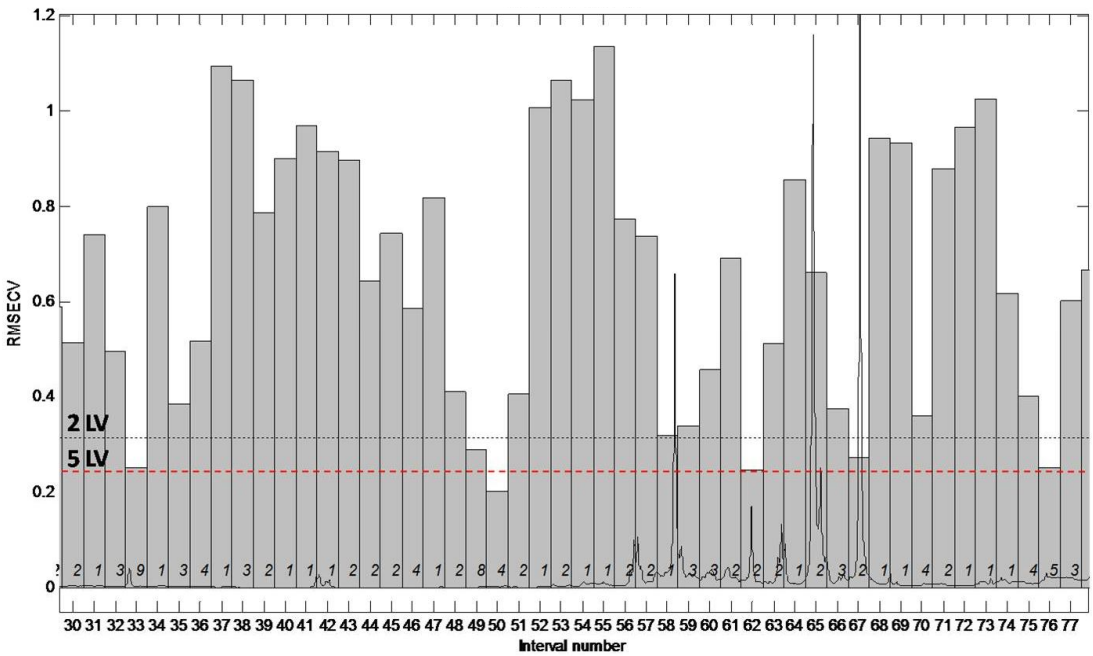


Figure 4.30: partially iPLS plot of 27 ^{1}H-NMR fish spectra at t_{16}. The intervals are described with a different number of LV in order to compare the prediction performance of the main interval (Black dot line) with the global one (Red dot Line)

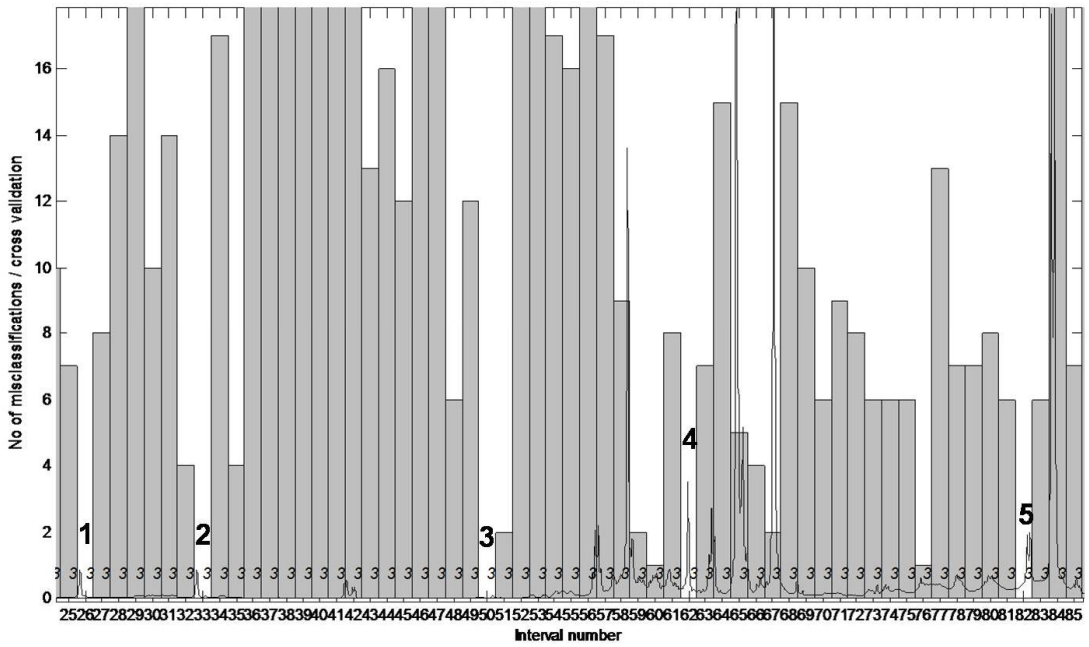


Figure 4.31: partially iECVA plot of 27 ^{1}H-NMR fish spectra at t_{16}. The interval number are plotted versus the number of misclassified. In this case all the intervals are described with the same number of LV. Four discriminant intervals have been found out

Plotting the relative intervals shown in Figure 4.31 (labelled from #1 to #5) of the ^{1}H-NMR fish spectrum at t_{16}, we obtain metabolites responsible of the separation among aquaculture. In Figure 4.32 all the spectra were superimposed in order to clearly see the differences in concentration of the discriminant metabolites. Alanine, glycine, histidine and an unknown metabolite become to be the representative discriminant factors.

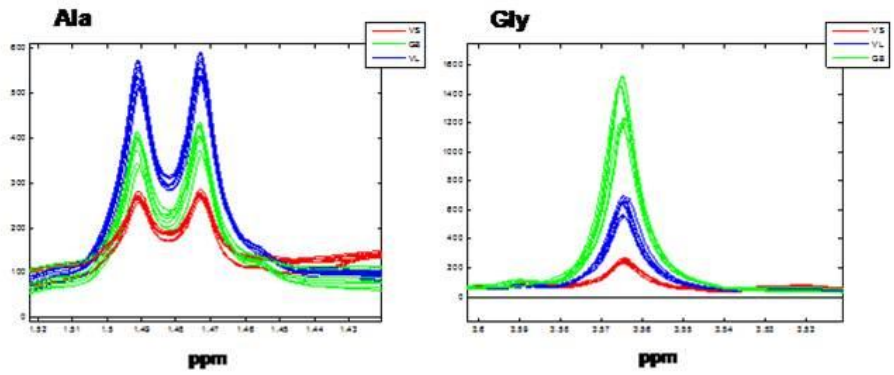


Figure 4.32 – part one: peaks plot of superimposed spectra divided according to aquaculture systems. Alanine, representing the interval #5 and glycine, interval #4 of Figure 4.31, are the most discriminant metabolites according to aquaculture systems

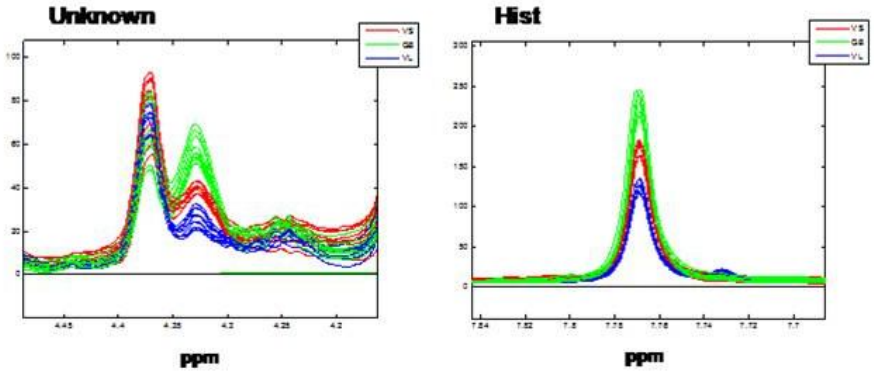


Figure 4.32 – part two: peaks plot of superimposed spectra divided according to aquaculture systems. Histidine, interval #1 and #2, and an unsigned compound, interval #3 of Figure 4.31, are the most discriminant metabolites according to aquaculture systems

4.4 CONCLUSION

Consumers today set restrictive demands for their foods; key words are **health**, **naturalness**, **quality**, and **safety** (Anon, 1998) and above all food should have a pleasing appearance, odor, taste and texture. Fish present a good starting point because it represents healthy and natural food. It is therefore a great challenge for the fish industry to develop delicious, convenient and high quality fish products in order to improve the competitiveness of this industry towards other commodities such as chicken, beef and pork in order to satisfy the consumer.

As it has been told in Chapter 4, par. 4.1, the term of quality included several aspects both from an objective and subjective point of view. On the base of our studies, ideally, if we want to have an unique indicator of the freshness, it should be able to integrate the effects of time and temperature at the same levels as the changes that occur in the fish. It should derive from a good correlation between its evolution and freshness decline or time of storage, should be non-subjective and independent of slaughter (sacrifice) conditions, and of physiological status of the fish. Thus, it is really important to have versatile and easy instruments able to detect those features characterizing the freshness in fish. From this point of view, the

NMR approach can give a lot of information in one step on the physiological status of fish, connected to the aquaculture systems, and taking in consideration both time and temperature effects.

The results obtained in this work affirm that NMR together with the multivariate statistics (ordination, classification and prediction) support the conclusion that from one side time of ice storage is a critical limiting variable for the Gilthead Sea bream but in the same time also the nutritional conditions and thus fish farming systems can affect the status of samples.

Even if all the information on the status of samples are condensed in few variables represented by metabolites like inosine, creatine, TMA and aminoacids, is really difficult to find a single indicator sufficient to evaluate sea food quality. On the contrary, is necessary to combining indicators and differentiating those that determine loss of freshness from those that detect the nutritional status of fish. In fact, the ratio between INO/IMP can give information on post-mortem time of storage, whilst the amount of aminoacids, in particularly glycine, on the status of samples ante mortem. Anyway, other post-mortem aging index like TMAO/TMA, depends on many ante mortem (the starting concentration of TMAO) and post-mortem factors (enzymatic conversion of TMAO in TMA) so that difficulty can be considered as a univocal quality markers. However, it remains necessary to establish useable criteria for fish freshness and spoilage that are practical both for the fish industry and the consumer, reflecting their demand for fish freshness (C. Delbarre-Ladrat et *al.*, 2006).

Moreover, the results of this work point on relevance the importance of the combination of several multivariate data analysis. Further elaboration, expect for PCA, PLS2-DA, ECVA and iPLS allowed to find other part of the spectrum and thus other metabolites that can better express the terms of differentiation. As it can be seen, other metabolites are involved in the discrimination among aquaculture

systems like Histidine and an unknown peak, but also alanine whose signals lay in the upfield region, the ones that from the previous PCA analysis was excluded.

REFERENCES

A. L., Stoll, W. E., Severus, M. P., Freeman, S., Rueter, H. A., Zboyan, E., Diamond, K. M., Cress and L. B., Marangell (1999). *Omega 3 fatty acids in bipolar disorder*. Arch. Gen. Psychiatr, 56, 407-412

A., Bhise, P. V., Krishnan, R., Aggarwal, M., Gaiha and J., Bhattacharjee (2005). *Effect of low dose omega-3 fatty acids substitution on blood pressure, hyperinsulinemia and dyslipidemia in Indians with essential hypertension: pilot study*. Indian Journal of Clinical Biochemistry, 20(2), 4-9

A., Craig, O., Cloarec, E., Holmes, J. K., Nicholson and J. C., Lindon (2006). *Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets*. Anal. Chem, 78, 2262-2267

A., Davidson (1972). *Mediterranean Seafood*. Penguin, 86-108

A., Huidobro, R., Mendes and M. L., Nunes (2001). *Slaughtering of gilthead sea bream (Sparus aurata) in liquid ice: influence on fish quality*. Eur Food Res Technol, 213, 267–272

ANON. (1998). *The Nordic fisheries in the new consumer era*. Report prepared for the Nordic Council of Ministers by the House of Mandag Morgen, Copenhagen, Denmark, 1-100

B. T., Lunestad, L., Nesse, J., Lassen, B., Svihus, T., Nesbakken, K., Fossum, J. T., Rosnes, H., Kruse and S., Yazdankhah (2007). *Salmonella in fish feed; occurrence and implications for fish andhuman health in Norway*. Aquaculture 265, 1–8

B., Sitter, J., Krane, I. S., Gribbestad, L., Jørgensen and M., Aursand (1999). *Quality evaluation of Atlantic halibut (Hippoglossus hippoglossus, L.) during ice storage using ^{1}H NMR spectroscopy.* Advances in Magnetic Resonance in Food Science, 226-237

C. H. S., Ruxton, S. C., Reed, M. J. A.., Simpson and K. J., Millington (2004). *The health benefits of omega-3 polyunsaturated fatty acids: a review of the evidence*. J Hum Nutr Dietet, 17, 449–459

C. M., Oomen, E. J. M., Feskens, L., Rasanen, F., Fidenza, A. M., Nissinen, A., Menotti, F. J., Kok and D., Kromholt (2000). *Fish consumption and coronary heart disease mortality in Finland, Italy and the Netherlands.* Am. J. Epidemiol, 151, 999–1006

C. M., Smagula and I. R. Adelman (1980). Uptake of ^{14}C-Glycine by Scales as an Index of Fish Growth: Effect of Fish Acclimation Temperature. Transactions of the American Fisheries Society, 109, 187–194

C. R., Santerre, P. B., Bush, D. H., Xu, G. W., Lewis, J. T., Davis, R. M., Grodner, R., Ingram, C. I., Wei and J. M,. Hinshaw (2001). *Metal Residues in Farm-Raised Channel Catfish, Rainbow Trout, and Red Swamp Crayfish from the Southern U.S.* Journal of Food Science 66 (2), 270–273

C., Alasalvar, K. D. A., Taylor and F., Shahidi (2005). *Comparison of Volatiles of Cultured and Wild Sea Bream (Sparus aurata) during Storage in Ice by Dynamic Headspace Analysis/Gas Chromatography-Mass Spectrometry*. J. Agric. Food Chem. 53, 2616-2622

C., Cahu, E., Salen and M., De Lorgeril (2004). *Farmed and wild fish in the prevention of cardiovascular diseases: Assessing possible differences in lipid nutritional values*. Nutr Metab Cardiovasc Dis, 14, 34–41

C., Delbarre-Ladrat, R., Chéret, R., Taylor and V., Verrez-Bagnis (2006). *Trends in Postmortem Aging in Fish: Understanding of Proteolysis and Disorganization of the Myofibrillar Structure*. Critical Reviews in Food Science and Nutrition, 46, 409–421

D. W., Cole, R., Cole, S. J., Gaydos, J., Gray, G., Hyland, M. L., Jacques, N., Powell-Dunford, C., Sawhney and W. W., Au (2008).

Aquaculture: Environmental, toxicological, and health issues. Int. J. Hyg. Environ. Health

E., Englehaupt (2007). *Farming the deep blue sea*. Environ. Sci. Technol, 41, 4188–4191

F. A., Vazquez-Ortiz, R.,Pacheco-Aguilar, M. E., Lugo-Sanchez, R. E., Villegas-Ozuna (1997). *Application of the Freshness Quality Index (K Value) for Fresh Fish to Canned Sardines from Northwestern Mexico*. Journal of Food Composition and Analysis, 10: 158–165

F. F., Brown, I. D., Campbell, P. W., Kuchel and D. L., Rabenstein (1977). *Human erythrocyte metabolism studies by ^{1}H spin echo NMR*. FEBS Lett. 82, 12-16

F. H., Larsen, F., van den Berg and S. B., Engelsen (2006). *An exploratory chemometric study of H-1 NMR spectra of table wines.* Journal of Chemometrics, 20, 198-208.

F., Özogul, K. D. A., Taylor, P., Quantick and Y., Özogul, (2000). *Chemical, microbiological and sensory evaluation of Atlantic herring (Clupea harengus) stored in ice, modified atmosphere and vacuum pack*. Food Chemistry, 71, 267–273

H., Hotelling (1993). *Analysis of a complex of statistical variables into principal components*. J. Educ. Psychol, 24, 417–441

H., Senzaki, A., Tsubara and H. Y., Takada (2001). *Effect of eicosapentanoic acid on the suppression of growth and metastasis of human breast cancer cells in vivo and in vitro*. World Rev Nutr Diet, 88, 117-125

H., Uchiyama, S., Ehira, T., Kawahara, W., Shimidu (1970). *Significance in measuring volatile base and trimethylamine nitrogen and nucleotides in fish muscle as indices of freshness*. Bull. Japan. Soc. Sci. Fish., 36: 177.

H., Winnin, F. H., Larsen, R., Bro and S. B., Engelsen (2008). *Quantitative analysis of NMR spectra with chemometrics*. Journal of Magnetic Resonance, 190, 26-32

H., Yokogoshi and H., Oda (2002). *Dietary taurine enhances cholesterol degradation and reduces serum and liver cholesterol concentrations in rats fed a high-cholesterol diet*. Amino Acids, 23, 433–439

H., Yokogoshi, H., Mochizuki, K., Nanami, Y., Hida, F., Miyachi and O., Hiroaki (1999). *Dietary Taurine Enhances Cholesterol Degradation and Reduces Serum and Liver Cholesterol Concentrations in Rats Fed a High-Cholesterol Diet1*. Journal of Nutrition, 129, 1705-1712

I. S., Gribbestad, S. B., Peterson, H., Fjøsne, S., Kvinnsland and J., Krane (1994). *^{1}H NMR spectroscopic characterization of perchloric acid extracts from breast carcinoma and non-involved breast tissue*. NMR Biomed. 7, 181-194

I. S., Park and N., Kim (1999). *Simultaneous determination of hypoxanthine, inosine and inosine 50-monophosphate with serially connected three enzyme reactors*. Analytica Chimica Acta., 394: 201–210

I., Martinez, T., Bathen, I. B., Standal, J., Halvorsen, M., Aursand, I. S., Gribbestad and D. E., Axelson (2005). *Bioactive compounds in cod (Gadus morhua) products and suitability of ^{1}H NMR metabolite profiling for classification of the products using multivariate data analysis*. J. Agric. Food Chem. 53, 6889-6895

J. K., Nicholson, J. L. Lindon and E., Holmes (1999). *'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data*. Xenobiotica, 29, 1181-1189

J. M., Kennish and D. E., Kramer (1986). *A review of highpressure chromatographic methods for measuring nucleotide degradation in fish muscle. In Seafood Quality Determination*. Procedings of an

International Symposium Coordinated by the University of Alaska. Sea Grant College Program, Anchorage, Alaska U.S.A. l&l4 November, ed. Kramer D. E. & Liston J. Elsevier Science Publishers V., Amsterdam, Netherlands, 209-219

J. M., Ryder, D. H., Buisson, D. N., Scott and G. C., Fletcher (1984). *Storage of New Zealand jack mackerel (Trachurusb novaezelandiae) in ice: chemical, microbiological and sensory assessment*. Journal of Food Science, 49: 1453-1456

J. N., Dingle and J. A., Hines (1971). Degradation of inosine 5-monophosphate in skeletal muscle of several north Atlantic fishes. J. Fish. Res. Bd. Canada, 28: 1125.

J., Arakaki and M., Suyama (1974). *Formation of Glycine from Purine Derivatives in the Fish Muscles Extract during Acid Hydrolysis*. Bulletin of the Japanese Society of Scientific Fisheries, 40, 731

K. H., Esbensen (2002). *Data analysis: In Practice. An introduction to multivariate data analysis and experimental design*. 5th edition, 490-492

K., Govindan (2002). *Marine benthos – A future perspective*. Proc. The National Seminar on Creeks, Estuaries and Mangroves – Pollution and Conservation, 28:30

K., Grigorakis (2007). *Compositional and organoleptic quality of farmed and wild gilthead sea bream (Sparus aurata) and sea bass (Dicentrarchus labrax) and factors affecting it: A review*. Aquaculture 272, 55–75

L., Nørgaard, A., Saudland, J., Wagner, P., Nielsen, L., Munk, and S. B., Engelsen (2000). *Interval partial least-squares regression (iPLS): a Comparative Chemometric study with an example from near infrared spectroscopy*. Applied Spectroscopy. 54, 3, 413 – 418

L., Nørgaard, R., Bro, F., Westad and S. B., Engelsen (2006). *A modification of canonical variates analysis to handle highly collinear multivariate data.* J. Chemometrics, 20, 425–435

L., Senso, M. D., Suárezb, T., Ruiz-Carab and M., García-Gallego (2007). *On the possible effects of harvesting season and chilled storage on the fatty acid profile of the fillet of farmed gilthead sea bream (Sparus aurata).* Food Chemistry, 101, 298–307

L., Sola, A, Moretti, D, Crosetti, N., Karaiskou, A., Magoulas, A. R., Rossi, M., Rye, A., Triantafyllidis and C. S., Tsigenopoulos (2007). *Gilthead sea bream - Sparus aurata*. Compendium of Genetic Impact of Aquaculture Activities on Native Population. 47-54

M. D., Huynh, R., Mackey and R., Gawley (1992). *Freshness assessment of Pacific fish species using K-value*. In: *Seafood Science and Technology* (edited by E. Graham Bligh), Oxford: Fishing News (Books) Limited, 258-268

M. T., Veciana-Nogues, M. S., Albala-Hurtado, M., Izquierdo-Pulido and M. C., Vidal-Carou (1996). *Validation of a gas-chromatographic method for volatile amine determination in fish samples*. Food Chemistry, 57, 569-573

M., Aursand and H., Grasdalen (1992). *Interpretation of the ^{13}C-NMR spectra of omega-3 fatty acids and lipid extracted from the white muscle of Atlantic salmon (Salmo salar)*. Chemistry and Physics of Lipids. 62, 39-251.

M., De Lorgeril and P., Salen (2002). *Fish and n-3 fatty acids for the prevention and treatment of coronary heart disease. Nutrition is not pharmacology*. Am J Med, 112, 316-319

M., Krajnovic-Ozretic, M., Najdek and B., Ozretic (1994). *Fatty acids in liver and muscle of farmed and wild sea bass (Dicentrarchus labrax L.).* Comparative Biochemistry and Physiology, 109A, 611–617

M., Manera, P., Visciano, P., Losito and A., Ianieri (2003). *Farmed Fish Pathology: Quality Aspects*. Veterinary Research Communications, 27 Suppl. 1, 695–698

M. M., Friedman (1997). *Nutritional Value of Proteins from Different Food Sources.* A Review J. Agric. Food Chem, 44, 6–29

M., Murata and M. Sakaguchi (1986). *Storage of yellowtail (Seriola quinqueradiata) white and dark muscles in ice: chances in content of adenine nucleotides and related compounds.* Journal of Food Science, 51, 321-326

M., Wyss and R., Kaddurah-Daouk (2000). *Creatine and Creatinine Metabolism*. Physiological Reviews, 80, 1108-1182

N., Church (1998). *MAP fish and crustaceans - sensory enhancement*. Food Sci. Technol. Today, 12, 73-83

N., Kalogeropoulos, M. N., Alexis and R. J., Henderson (1992). *Effects of dietary soybean and cod liver oil levels on growth and body composition of gilthead bream (Sparus aurata)*. Aquaculture, 104, 308

R. A., Snyder (2008). *Analysis of Dioxin-Furan, PCB and trace metal levels in farmed catfish, Escambia County, FL*. Center for Environmental Diagnostics and Bioremediation University of West Florida

R. S., Pardini, D., Wilson, S., Schiff, A. S., Bajo and R., Pierce (2005). *Nutritional Intervention With Omega-3 Fatty Acids in a Case of Malignant Fibrous Histiocytoma of the Lungs*. Nutrition and Cancer, 52(2), 121–129

R., Froese and D., Pauly (Editors) (2008). FishBase. World Wide Web electronic publication. www.fishbase.org, version (10/2008)

R., Jonsdottir, G., Olafsdottir, E., Martinsdottir and G., Stefansson (2004). *Flavor Characterization of Ripened Cod Roe by Gas Chromatography, Sensory Analysis, and Electronic Nose*. J. Agric. Food Chem., 52(20), 6250–6256

R., Martin (1988). *Contaminants in relation to the quality of seafoods*. Food Technol. 42, 104

R., Mendes, R., Quinta and M. L., Nunes (2001). *Changes in baseline levels of nucleotides during ice storage of fish and crustaceans from the Portuguese coast*. Eur Food Res Technol., 212, 141–146

S. D., Klaoudatos (1989). *Experimental results on reproduction and larval rearing of Sparus aurata*. In: *Aquaculture. A Biotechnology in Progress* (edited by N., DePauw, E., Jaspers, H., Ackefors and N., Wilkins), 197–201. Bredene, Belgium: European Aquaculture Society

S., Rezzi, I., Giani, K., Héberger, D. E., Axelson, V. M., Moretti, F., Reniero and C., Guillou (2007). *Classification of Gilthead Sea Bream (Sparus aurata) from ^{1}H NMR lipid profiling combined with principal component and linear discriminant analysis*. J. Agric. Food Chem. 55, 9963-9968

S., Watabe, M., Kamal and K., Hashimoto (1990). *Postmortem Changes in ATP, Creatine Phosphate, and Lactate in Sardine Muscle*. Journal of Food Science, 56, 151 – 153

T., Civera, R. M., Turi, E., Parisi and G., Fazio (1995). *Further investigations on total volatile basic nitrogen and trimethylamine in some Mediterranean teleosts during cold storage*. Science des Aliments, 15, 179–186.

T., Kind and O., Fiehn (2006). *Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm*. BMC Bioinformatics, 7-234

T., Saito, K., Arai and M., Matsuyoshi (1959). *A new method for estimating the freshness of fish*. Bull. Japan. Sot. Sci. Fish., 24, 749-750

V. R., Kyrana, and V. P., Lougovois (2002). *Sensory, chemical, microbiological assessment of farm-raised European sea bass (Dicentrarchus labrax) stored in melting ice.* Int. J. Food Sci. Technol. 37, 319–328

APPENDIX A AND B

Appendix A

```
############################################################

rm(list=ls(all=TRUE))

#####################################################################

#Sink fish Farming Samples

p1<-scan("1.txt", dec=".")
p2<-scan("2.txt", dec=".")
p3<-scan("3.txt", dec=".")
p4<-scan("4.txt", dec=".")
p5<-scan("5.txt", dec=".")
p6<-scan("6.txt", dec=".")
p7<-scan("7.txt", dec=".")
p8<-scan("8.txt", dec=".")
p9<-scan("9.txt", dec=".")
p10<-scan("10.txt", dec=".")
p11<-scan("11.txt", dec=".")
p12<-scan("12.txt", dec=".")
p13<-scan("13.txt", dec=".")
p14<-scan("14.txt", dec=".")
p15<-scan("15.txt", dec=".")
p16<-scan("16.txt", dec=".")
p17<-scan("17.txt", dec=".")
p18<-scan("18.txt", dec=".")

#Lagoon Fish Farming Samples

p19<-scan("1.txt", dec=".")
p20<-scan("2.txt", dec=".")
p21<-scan("3.txt", dec=".")
p22<-scan("4.txt", dec=".")
p23<-scan("5.txt", dec=".")
p24<-scan("6.txt", dec=".")
p25<-scan("7.txt", dec=".")
p26<-scan("8.txt", dec=".")
p27<-scan("9.txt", dec=".")
p28<-scan("10.txt", dec=".")
p29<-scan("11.txt", dec=".")
p30<-scan("12.txt", dec=".")
p31<-scan("13.txt", dec=".")
p32<-scan("14.txt", dec=".")
p33<-scan("15.txt", dec=".")
p34<-scan("16.txt", dec=".")
p35<-scan("17.txt", dec=".")
p36<-scan("18.txt", dec=".")
```

```
#Cages Fish Farming Samples

p37<-scan("1.txt", dec=".")
p38<-scan("2.txt", dec=".")
p39<-scan("3.txt", dec=".")
p40<-scan("4.txt", dec=".")
p41<-scan("5.txt", dec=".")
p42<-scan("6.txt", dec=".")
p43<-scan("7.txt", dec=".")
p44<-scan("8.txt", dec=".")
p45<-scan("9.txt", dec=".")
p46<-scan("10.txt", dec=".")
p47<-scan("11.txt", dec=".")
p48<-scan("12.txt", dec=".")
p49<-scan("13.txt", dec=".")
p50<-scan("14.txt", dec=".")
p51<-scan("15.txt", dec=".")
p52<-scan("16.txt", dec=".")
p53<-scan("17.txt", dec=".")
p54<-scan("18.txt", dec=".")

########################
#  Delete Water Signals #
########################

p1<-p1[c(0:15800,16800:32768)]
p2<-p2[c(0:15800,16800:32768)]
p3<-p3[c(0:15800,16800:32768)]
p4<-p4[c(0:15800,16800:32768)]
p5<-p5[c(0:15800,16800:32768)]
p6<-p6[c(0:15800,16800:32768)]
p7<-p7[c(0:15800,16800:32768)]
p8<-p8[c(0:15800,16800:32768)]
p9<-p9[c(0:15800,16800:32768)]
p10<-p10[c(0:15800,16800:32768)]
p11<-p11[c(0:15800,16800:32768)]
p12<-p12[c(0:15800,16800:32768)]
p13<-p13[c(0:15800,16800:32768)]
p14<-p14[c(0:15800,16800:32768)]
p15<-p15[c(0:15800,16800:32768)]
p16<-p16[c(0:15800,16800:32768)]
p17<-p17[c(0:15800,16800:32768)]
p18<-p18[c(0:15800,16800:32768)]
p19<-p19[c(0:15800,16800:32768)]
p20<-p20[c(0:15800,16800:32768)]
p21<-p21[c(0:15800,16800:32768)]
p22<-p22[c(0:15800,16800:32768)]
p23<-p23[c(0:15800,16800:32768)]
p24<-p24[c(0:15800,16800:32768)]
p25<-p25[c(0:15800,16800:32768)]
p26<-p26[c(0:15800,16800:32768)]
p27<-p27[c(0:15800,16800:32768)]
p28<-p28[c(0:15800,16800:32768)]
p29<-p29[c(0:15800,16800:32768)]
p30<-p30[c(0:15800,16800:32768)]
p31<-p31[c(0:15800,16800:32768)]
p32<-p32[c(0:15800,16800:32768)]
p33<-p33[c(0:15800,16800:32768)]
p34<-p34[c(0:15800,16800:32768)]
p35<-p35[c(0:15800,16800:32768)]
p36<-p36[c(0:15800,16800:32768)]
p37<-p37[c(0:15800,16800:32768)]
p38<-p38[c(0:15800,16800:32768)]
p39<-p39[c(0:15800,16800:32768)]
p40<-p40[c(0:15800,16800:32768)]
p41<-p41[c(0:15800,16800:32768)]
p42<-p42[c(0:15800,16800:32768)]
```

```
p43<-p43[c(0:15800,16800:32768)]
p44<-p44[c(0:15800,16800:32768)]
p45<-p45[c(0:15800,16800:32768)]
p46<-p46[c(0:15800,16800:32768)]
p47<-p47[c(0:15800,16800:32768)]
p48<-p48[c(0:15800,16800:32768)]
p49<-p49[c(0:15800,16800:32768)]
p50<-p50[c(0:15800,16800:32768)]
p51<-p51[c(0:15800,16800:32768)]
p52<-p52[c(0:15800,16800:32768)]
p53<-p53[c(0:15800,16800:32768)]
p54<-p54[c(0:15800,16800:32768)]

########################################
# Delete Points from the Spectrum Tails#
########################################

p1<-p1[c(8001:26000)]
p2<-p2[c(8001:26000)]
p3<-p3[c(8001:26000)]
p4<-p4[c(8001:26000)]
p5<-p5[c(8001:26000)]
p6<-p6[c(8001:26000)]
p7<-p7[c(8001:26000)]
p8<-p8[c(8001:26000)]
p9<-p9[c(8001:26000)]
p10<-p10[c(8001:26000)]
p11<-p11[c(8001:26000)]
p12<-p12[c(8001:26000)]
p13<-p13[c(8001:26000)]
p14<-p14[c(8001:26000)]
p15<-p15[c(8001:26000)]
p16<-p16[c(8001:26000)]
p17<-p17[c(8001:26000)]
p18<-p18[c(8001:26000)]
p19<-p19[c(8001:26000)]
p20<-p20[c(8001:26000)]
p21<-p21[c(8001:26000)]
p22<-p22[c(8001:26000)]
p23<-p23[c(8001:26000)]
p24<-p24[c(8001:26000)]
p25<-p25[c(8001:26000)]
p26<-p26[c(8001:26000)]
p27<-p27[c(8001:26000)]
p28<-p28[c(8001:26000)]
p29<-p29[c(8001:26000)]
p30<-p30[c(8001:26000)]
p31<-p31[c(8001:26000)]
p32<-p32[c(8001:26000)]
p33<-p33[c(8001:26000)]
p34<-p34[c(8001:26000)]
p35<-p35[c(8001:26000)]
p36<-p36[c(8001:26000)]
p37<-p37[c(8001:26000)]
p38<-p38[c(8001:26000)]
p39<-p39[c(8001:26000)]
p40<-p40[c(8001:26000)]
p41<-p41[c(8001:26000)]
p42<-p42[c(8001:26000)]
p43<-p43[c(8001:26000)]
p44<-p44[c(8001:26000)]
p45<-p45[c(8001:26000)]
p46<-p46[c(8001:26000)]
p47<-p47[c(8001:26000)]
p48<-p48[c(8001:26000)]
p49<-p49[c(8001:26000)]
p50<-p50[c(8001:26000)]
p51<-p51[c(8001:26000)]
```

```
p52<-p52[c(8001:26000)]
p53<-p53[c(8001:26000)]
p54<-p54[c(8001:26000)]



#########################################################
#    Collecting Spectra Data in a Matrix 54x18000   #
#########################################################

tabella<-matrix(0, nrow=54, ncol=18000)

tabella[1,]<-p1
tabella[2,]<-p2
tabella[3,]<-p3
tabella[4,]<-p4
tabella[5,]<-p5
tabella[6,]<-p6
tabella[7,]<-p7
tabella[8,]<-p8
tabella[9,]<-p9
tabella[10,]<-p10
tabella[11,]<-p11
tabella[12,]<-p12
tabella[13,]<-p13
tabella[14,]<-p14
tabella[15,]<-p15
tabella[16,]<-p16
tabella[17,]<-p17
tabella[18,]<-p18
tabella[19,]<-p19
tabella[20,]<-p20
tabella[21,]<-p21
tabella[22,]<-p22
tabella[23,]<-p23
tabella[24,]<-p24
tabella[25,]<-p25
tabella[26,]<-p26
tabella[27,]<-p27
tabella[28,]<-p28
tabella[29,]<-p29
tabella[30,]<-p30
tabella[31,]<-p31
tabella[32,]<-p32
tabella[33,]<-p33
tabella[34,]<-p34
tabella[35,]<-p35
tabella[36,]<-p36
tabella[37,]<-p37
tabella[38,]<-p38
tabella[39,]<-p39
tabella[40,]<-p40
tabella[41,]<-p41
tabella[42,]<-p42
tabella[43,]<-p43
tabella[44,]<-p44
tabella[45,]<-p45
tabella[46,]<-p46
tabella[47,]<-p47
tabella[48,]<-p48
tabella[49,]<-p49
tabella[50,]<-p50
tabella[51,]<-p51
tabella[52,]<-p52
tabella[53,]<-p53
tabella[54,]<-p54
```

```
#####################################################
# PCA auotoscaled on a Raw Spectral Data Matrix  #
#####################################################

sm<-edit(tabella)

analisi<-prcomp(tabella,center=T,scale=T)
y<-c(1,2)
plot(analisi$x[,y])
colorigruppi<-c(rep(2,18),rep(3,18),rep(4,18))
plot(analisi$x[,y], main="PCA", pch=" ")
text(analisi$x[,y],labels=c(1:54),col=colorigruppi)

######################
#    Mean Spectrum   #
######################

# for the next Step

medio<-apply(tabella,2,mean)
plot(medio, type="l")

which.max(medio[14400:14600])

# chiedo qual'è il valore massimo nell'intervallo scelto riferito
allo spettro medio

# il valore che ottengo lo aggiungo al primo valore dell'intervallo
sopra indicato

####################################
# Refer all Spectra on a Peack     #
####################################

ncamp<-54
punti<-18000
centro<-14496 #(rispetto al campione medio)
maxvet<-matrix(data=0, ncol=1, nrow=ncamp)
reganomer<-c(14400:14600)
for (i in 1:ncamp){
for (j in reganomer){
if (tabella[i,j]==max(tabella[i,reganomer])) maxvet[i,]<- centro-j
}
}
allineati<- matrix(data=0, ncol=punti, nrow=ncamp)
for (i in 1:ncamp) {
shift<-maxvet[i,1]
if (shift>0) allineati[i,(1+shift):punti]<-tabella[i,1:(punti-
shift)]
if (shift==0) allineati[i,]<-tabella[i,]
if (shift<0) allineati[i,1:(punti+shift)]<-tabella[i,(1-
shift):punti]
}
plot(maxvet)

########################################
#  Normalization on Total Specrtum Area  #
########################################

ncamp<-nrow(allineati)
npunti<-ncol(allineati)
riferiti<-matrix(data=0, ncol=npunti, nrow=ncamp)
area<-matrix(data=0, ncol=1, nrow=ncamp)
for (j in 1:ncamp){
area[j,]<-sum(allineati[j,])
for (i in 1:npunti){
riferiti[j,i]<-allineati[j,i]/area[j,]
}
```

```
}

#######################################################################
#Calcolo Coef. da moltiplicare a tutti gli spettri dopo normaliz   #
#######################################################################

ncamp<-nrow(allineati)
coefmat<-matrix(data=0, nrow=ncamp, ncol=1)
for(i in 1:ncamp){
coefmat[i,]<-sum(allineati[i,])
}
mediab<-sum(coefmat)/ncamp

riferiti<-riferiti*mediab

############################
#  Regional Coeficient    #
############################

# to obtain a coeficient toh ave a same maximum peack level

medio2<-apply(riferiti,2,mean)
plot(medio2, type="l")

#Max dowfield
max(medio2[0:7000])

#Max midfield
max(medio2[7001:11200])

#Max upfield
max(medio2[11200:18000])


max(medio2[7001:11200])/max(medio2[0:7000])

max(medio2[7001:11200])/max(medio2[11200:18000])

spettroamplificato<-matrix(data=0, ncol=18000, nrow=54)

for(i in 1:54){
spettroamplificato[i,0:7000]<-
riferiti[i,0:7000]*(max(medio2[7001:11200])/max(medio2[0:7000]))
spettroamplificato[i,7001:11200]<-riferiti[i,7001:11200]
spettroamplificato[i,11200:18000]<-
riferiti[i,11200:18000]*(max(medio2[7001:11200])/max(medio2[11200:18
000]))
}

#############################################################
#  PCA Meancentered Whole Spectrum with coeficient Region #
#############################################################

analisi<-prcomp(spettroamplificato,center=T,scale=F)
y<-c(1,2)
plot(analisi$x[,y], main="PCA", pch=" ")
colorigruppi<-c(rep(2,18),rep(3,18),rep(4,18))
plot(analisi$x[,y], xlab="PC1 (38%)",ylab="PC2
(19%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
legend(locator(2), c("VST0","VST16","VLT0","VLT16","GBT0","GBT16"),
pch=c(1,16,2,17,22,15), col=c(2,2,3,3,4,4))

#########################
# Cumulative Variance %  #
#########################
```

```
x<-0
varcum<-rep(0,54)
propvarcum<-rep(0,54)
for (i in 1:54) {
j<-i-1
varcum[i]<-analisi$sdev[i]^2+x
x<-varcum[i]
propvarcum[i]<-(varcum[i]/sum(analisi$sdev^2))*100
}
propvarcum

############################
#  To make noise value =0  #
############################

ncamp<-nrow(riferiti)
tabella2<-matrix(0, nrow=ncamp, ncol=ncol(riferiti))
for (i in 1:ncamp){
for (j in 1: ncol(riferiti)){
a<-(riferiti[i,j])
if (a<=0.8) tabella2[i,j]=0
if (a>0.8) tabella2[i,j]=riferiti[i,j]
}
}

################################################
#  PCA Meancentered Whole Spectrum with noise=0 #
################################################

analisi<-prcomp(tabella2,center=T,scale=F)
y<-c(1,2)
plot(analisi$x[,y], main="PCA", pch=" ")
colorigruppi<-c(rep(2,18),rep(3,18),rep(4,18))
plot(analisi$x[,y], xlab="PC1 (49%)",ylab="PC2
(31%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)

################################################
#  PCA Meancentered Downfield Region Riferiti #
################################################

analisi<-prcomp(riferiti[,0:7000],center=T,scale=F)
y<-c(1,2)
plot(analisi$x[,y], xlab="PC1 (43%)",ylab="PC2
(23%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)

legend(locator(2), c("VST0","VST16","VLT0","VLT16","GBT0","GBT16"),
pch=c(1,16,2,17,22,15), col=c(2,2,3,3,4,4))

################################################
#  PCA Meancentered Midfield Region Riferiti #
################################################

analisi<-prcomp(riferiti[,7001:11200],center=T,scale=F)
y<-c(1,2)
plot(analisi$x[,y])
plot(analisi$x[,y], xlab="PC1 (54%)",ylab="PC2
(31%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)

legend(locator(2), c("VST0","VST16","VLT0","VLT16","GBT0","GBT16"),
pch=c(1,16,2,17,22,15), col=c(2,2,3,3,4,4))

################################################
```

```
# PCA Meancentered Midfield Region Riferiti #
############################################

analisi<-prcomp(riferiti[,11200:18000],center=T,scale=F)
y<-c(1,2)

plot(analisi$x[,y])
plot(analisi$x[,y], xlab="PC1 (55%)",ylab="PC2
(18%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)

#########################
#  PCA Autoscaled Time 0 #
#########################

t0<-c(1:9,19:27,37:45)

coloT0<-c(rep(2,9),rep(3,9),rep(4,9))
analisi<-prcomp(riferiti[t0,],center=T,scale=T)
y<-c(1,2)
plot(analisi$x[,y], xlab="PC1 (46%)",ylab="PC2
(15%)",pch=c(rep(1,9), rep(2,9),rep(22,9)), cex=2.5, col=coloT0)

legend(locator(2), c("VST0","VLT0","GBT0"), pch=c(1,2,22),
col=c(2,3,4))

#########################
#  PCA Autoscaled Time 16 #
#########################

t16<-c(10:18,28:36,46:54)

coloT0<-c(rep(2,9),rep(3,9),rep(4,9))
analisi<-prcomp(riferiti[t16,],center=T,scale=T)
y<-c(1,2)
plot(analisi$x[,y], xlab="PC1 (36%)",ylab="PC2
(25%)",pch=c(rep(16,9), rep(17,9),rep(15,9)), cex=2.5, col=coloT0)

legend(locator(2), c("VST0","VLT0","GBT0"), pch=c(16,17,15),
col=c(2,3,4))

############################################
#  Binning Normalized and Refered Data Set #
############################################

ncamp<-nrow(riferiti)
intervallo<-matrix(data=0, nrow=ncamp, ncol=150)
for (z in 1:ncamp) {
i<-1
for (i in 1:150){
for (j in 1:120){
k<-((i-1)*120)+j
a<-riferiti[z,k] # to bin riferiti Matrix
intervallo[z,i]<-intervallo[z,i]+a
}
intervallo[z,i]<-intervallo[z,i]/120
}
}

########################################
#  Meancentered Binned PCA All Samples    #
########################################

analisi<-prcomp(intervallo,center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
```

```
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(51%)",cex=1.3, font=2))
title(ylab=list("PC2(29%)",cex=1.3, font=2))
title(main=list("PCA Binned Data all Samples", cex=2, font=1))
#########################
# Cumulative Variance %  #
#########################

x<-0
varcum<-rep(0,54)
propvarcum<-rep(0,54)
for (i in 1:54) {
j<-i-1
varcum[i]<-analisi$sdev[i]^2+x
x<-varcum[i]
propvarcum[i]<-(varcum[i]/sum(analisi$sdev^2))*100
}

propvarcum

#############################################################
#  Binning Normalized and Refered Data Set Without noise #
#############################################################

ncamp<-nrow(riferiti)
intsenzarum<-matrix(data=0, nrow=ncamp, ncol=150)
for (z in 1:ncamp) {
i<-1
for (i in 1:150){
for (j in 1:120){
k<-((i-1)*120)+j
a<-tabella2[z,k] # così binno la matrice riferiti (se voglio binnare
altre cambio il nome di a)
intsenzarum[z,i]<-intsenzarum[z,i]+a
}
intsenzarum[z,i]<-intsenzarum[z,i]/120
}
}

#########################################
#  PCA Meancentered Binned without noise  #
#########################################

analisi<-prcomp(intsenzarum,center=T,scale=F)
y<-c(1,2)
plot(analisi$x[,y], xlab="PC1 (51%)",ylab="PC2
(29%)",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)

#########################################################################
# Binning Normalized and Refered Data Set With Regional Coeficient #
#########################################################################

ncamp<-nrow(riferiti)
intconcoef<-matrix(data=0, nrow=ncamp, ncol=150)
for (z in 1:ncamp) {
i<-1
for (i in 1:150){
for (j in 1:120){
k<-((i-1)*120)+j
a<-spettroamplificato[z,k] # così binno la matrice
spettroamplificato (se voglio binnare altre cambio il nome di a)
intconcoef[z,i]<-intconcoef[z,i]+a
```

```
}
intconcoef[z,i]<-intconcoef[z,i]/120
}
}



#################################################
#  Meancentered Binned amplified PCA All Samples #
#################################################

analisi<-prcomp(intconcoef,center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(46%)",cex=1.3, font=2))
title(ylab=list("PC2(18%)",cex=1.3, font=2))
title(main=list("PCA Amplified Binned Data all Samples", cex=2,
font=1))

#############################################
# PCA Meancentered Binned Downfield Region #
#############################################

Downfield<-intervallo[,0:58]
analisi<-prcomp(Downfield,center=T,scale=F)
y<-c(1,2)

x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(54%)",cex=1.3, font=2))
title(ylab=list("PC2(20%)",cex=1.3, font=2))
title(main=list("PCA Binned Data Downfield Region", cex=2, font=1))

#####################################################
# BarPlot most representative Loadings Values  PC1  #
# Downfield Region                                  #
#####################################################

barplot(analisi$rotation[,1], axes=FALSE, col=c(rep("green",58)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Downfield Region PC1 Loadings", cex=2, font=1))
abline(h= 0.2)#to add a y=a line
abline(h=-0.2)#to add a second y=-a  line

#####################################################################
# Representation of the mean Area for Each bins All Samples     #
#####################################################################

z<-c(10920:11040)
riferitibis<-riferiti[1:54,]
integrale<-matrix(data=0, nrow=54, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

t0<-c(1:9,19:27,37:45)
t16<-c(10:18,28:36,46:54)
```

```
mean(integrale[t0,1])
mean(integrale[t16,1])

sd1<-sd(integrale[t0,1])
sd2<-sd(integrale[t16,1])


sd1
sd2

##########################################
# PCA Meancentered Binned Midfield Region #
##########################################

Midfield<-intervallo[,59:93]
analisi<-prcomp(Midfield,center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(60%)",cex=1.3, font=2))
title(ylab=list("PC2(29%)",cex=1.3, font=2))
title(main=list("PCA Binned Data Midfield Region", cex=2, font=1))

#######################################################
# BarPlot most representative Loadings Values  PC1  #
# Midfield Region                                     #
#######################################################

barplot(analisi$rotation[,1], axes=FALSE, col=c(rep("blue",35)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Midfield Region PC1 Loadings", cex=2, font=1))
abline(h= 0.15)#to add a y=a line
abline(h=-0.2)#to add a second y=-a  line

#######################################################
# BarPlot most representative Loadings Values  PC2  #
# Downfield Region                                    #
#######################################################

barplot(analisi$rotation[,2], axes=FALSE, col=c(rep("blue",35)))
axis(2, cex.axis=1.5)# to increas ylab values
title(ylab=list("Contribution of each Bin", cex=1.3, font=1))
title(main=list("Midfield Region PC2 Loadings", cex=2, font=1))
abline(h= 0.25)#to add a y=a line

#####################################################################
# Representation of the mean Area for Each bins All Samples     #
#####################################################################

z<-c(10560:10680)
riferitibis<-riferiti[1:54,]
integrale<-matrix(data=0, nrow=54, ncol=1)
for(i in 1:90){
integrale[i,]<-sum(riferitibis[i,z])
}

mean(integrale[1:9,1])
mean(integrale[10:18,1])
mean(integrale[19:27,1])
mean(integrale[28:36,1])
mean(integrale[37:45,1])
mean(integrale[46:54,1])
```

```
sd1<-sd(integrale[1:9,1])
sd2<-sd(integrale[10:18,1])
sd3<-sd(integrale[19:27,1])
sd4<-sd(integrale[28:36,1])
sd5<-sd(integrale[37:45,1])
sd6<-sd(integrale[46:54,1])


sd1
sd2
sd3
sd4
sd5
sd6

##########################################
# PCA Meancentered Binned Upfield Region  #
##########################################

Upfield<-intervallo[,94:150]
analisi<-prcomp(Upfield,center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(60%)",cex=1.3, font=2))
title(ylab=list("PC2(20%)",cex=1.3, font=2))
title(main=list("PCA Binned Data Upfield Region", cex=2, font=1))

##############################################
#  PCA Autoscaled All samples binned data t0 #
##############################################

t0<-c(1:9,19:27,37:45)

analisi<-prcomp(intervallo[t0,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab="
",pch=c(rep(1,9),rep(2,9),rep(22,9)), cex=2.5)
title(xlab=list("PC1(45%)",cex=1.3, font=2))
title(ylab=list("PC2(14%)",cex=1.3, font=2))

legend(locator(2), c("VST0","VLT0","GBT0"), pch=c(1,2,22))

##############################################
# PCA Autoscaled All samples binned data t16 #
##############################################

t16<-c(10:18,28:36,46:54)

analisi<-prcomp(intervallo[t0,],center=T,scale=T)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab="
",pch=c(rep(1,9),rep(2,9),rep(22,9)), cex=2.5)
title(xlab=list("PC1(36%)",cex=1.3, font=2))
title(ylab=list("PC2(25%)",cex=1.3, font=2))

legend(locator(2), c("VST16","VLT16","GBT16"), pch=c(16,17,15))
```

```
################################################
# Meancentered amplified Binned PCA All Samples #
# Down and Midfield Regions                     #
################################################

analisi<-prcomp(intconcoef[,0:93],center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(46%)",cex=1.3, font=2))
title(ylab=list("PC2(18%)",cex=1.3, font=2))
title(main=list("PCA Amplified Binned Data all Samples", cex=2,
font=1))

#######################################
# Meancentered Binned PCA All Samples #
# Down and Midfield Regions           #
#######################################

analisi<-prcomp(intervallo[,0:93],center=T,scale=F)
y<-c(1,2)
x<-c(analisi$x[,1])
y<-c(analisi$x[,2])
plot(x,y)
plot(x,y, cex.axis=1.3, xlab=" ",ylab=" ",pch=c(rep(1,9),
rep(16,9),rep(2,9),rep(17,9),rep(22,9),rep(15,9)), cex=2.5,
col=colorigruppi)
title(xlab=list("PC1(46%)",cex=1.3, font=2))
title(ylab=list("PC2(18%)",cex=1.3, font=2))
title(main=list("PCA Binned Data all Samples", cex=2, font=1))


#plot spettro down

x<-c(1:6960)
y<-c(riferiti[50,0:6960])
plot(x,y)
plot(x,y, type="l",cex.axis=1.3, xlab=" ",ylab=" ", col="green")

title(xlab=list("Specrtal Points",cex=1.3, font=2))
title(ylab=list("Intensity",cex=1.3, font=2))
title(main=list("Downfield Spectral Region", cex=2, font=1))

#plot spettro mid

x<-c(7080:11160)
y<-c(riferiti[50,7080:11160])
plot(x,y)
plot(x,y, type="l",cex.axis=1.3, xlab=" ",ylab=" ", col="blue")

title(xlab=list("Specrtal Points",cex=1.3, font=2))
title(ylab=list("Intensity",cex=1.3, font=2))
title(main=list("Midfield Spectral Region", cex=2, font=1))
```

Appendix B.1 Gilthead Sea bream – Sink – T_0 (VS-T0) spectrum

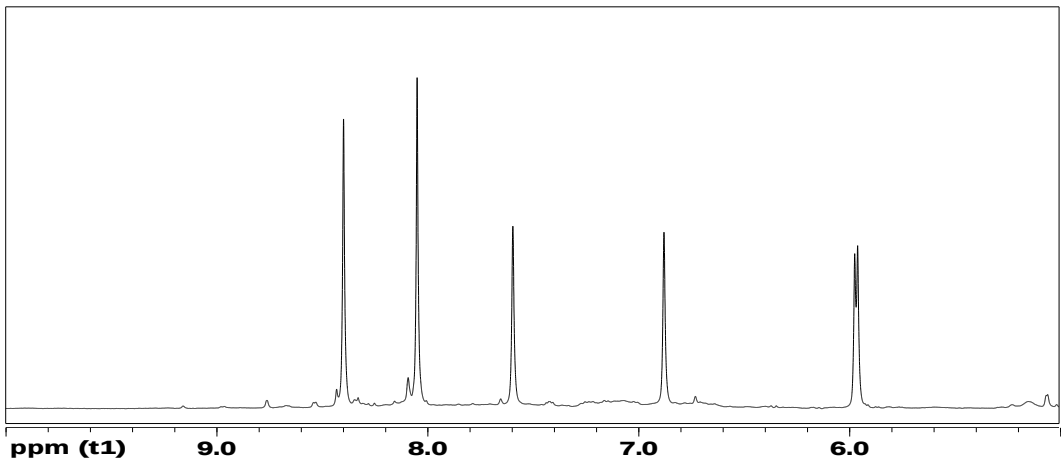


Figure Appendix B1: VS-T0: Downfield Region (> 5 ppm)

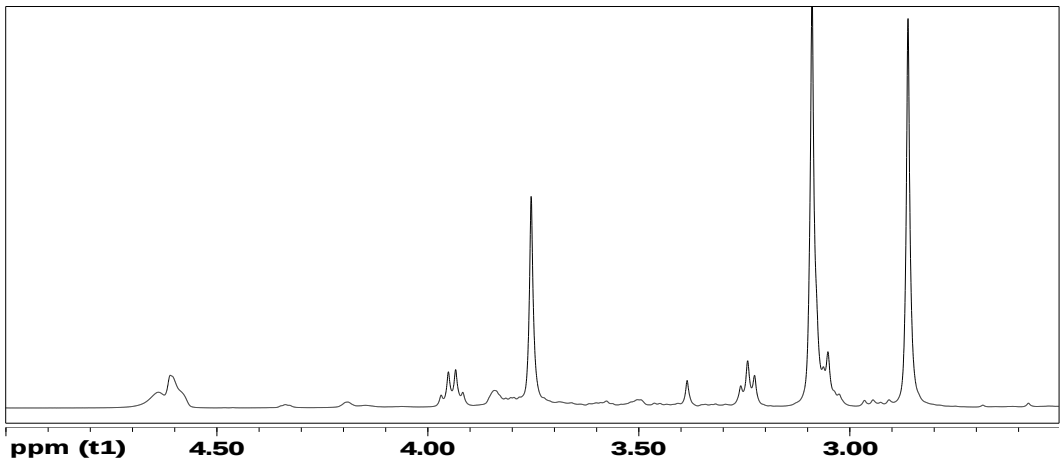


Figure Appendix B1: VS-T0: Midfield Region (5 < ppm < 2.5)

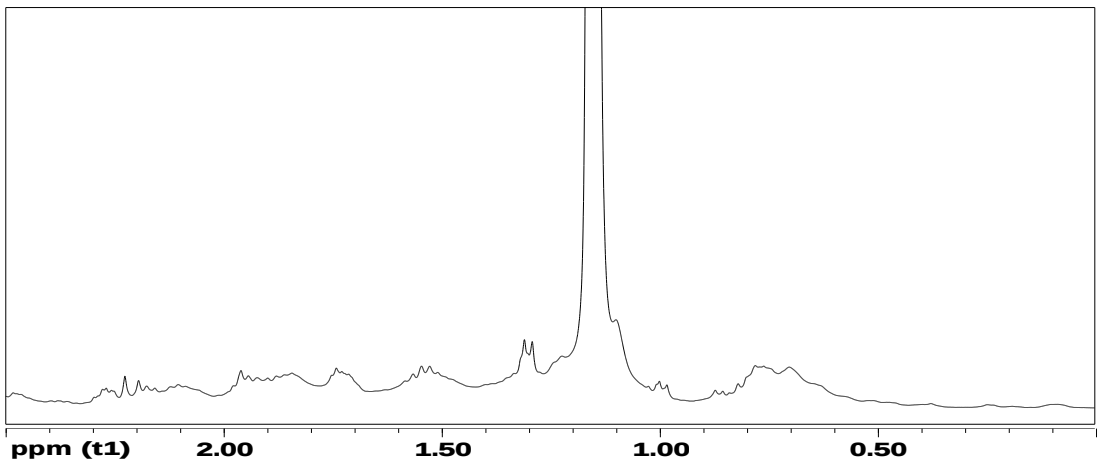


Figure Appendix B1: VS-T0: Upfield Region (< 2.5 ppm)

Appendix B.2 Gilthead Sea bream – Sink – T_{16} (VS-T16) spectrum

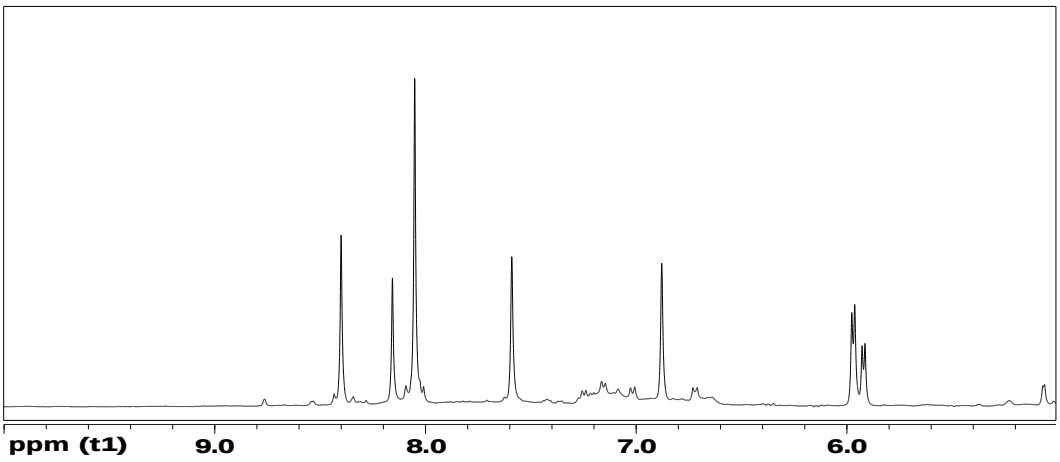


Figure Appendix B2: VS-T16: Downfield Region (> 5 ppm)

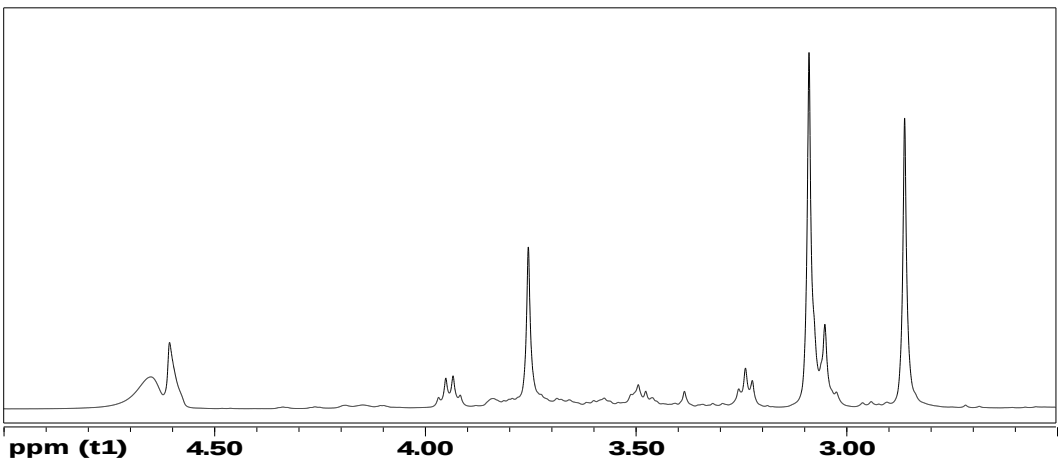


Figure Appendix B2: VS-T16: Midfield Region (5 < ppm < 2.5)

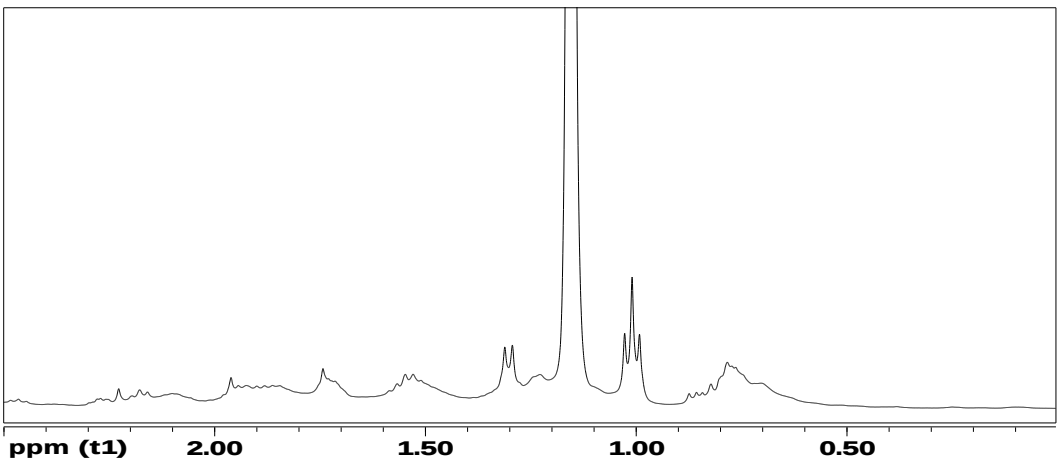


Figure Appendix B2: VS-T16: Upfield Region (< 2.5 ppm)

Appendix B.3 Gilthead Sea bream – Lagoon – T_0 (VL-T0) spectrum

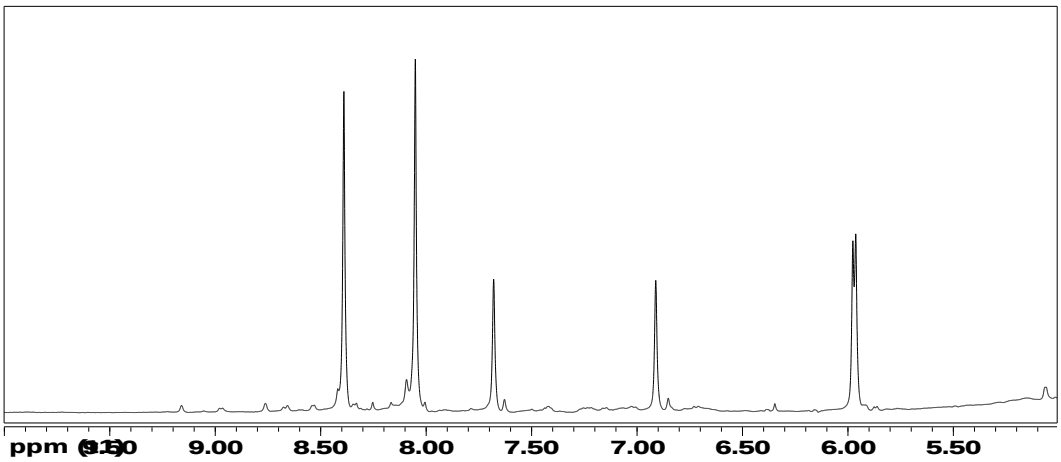


Figure Appendix B3: VL-T0: Downfield Region (> 5 ppm)

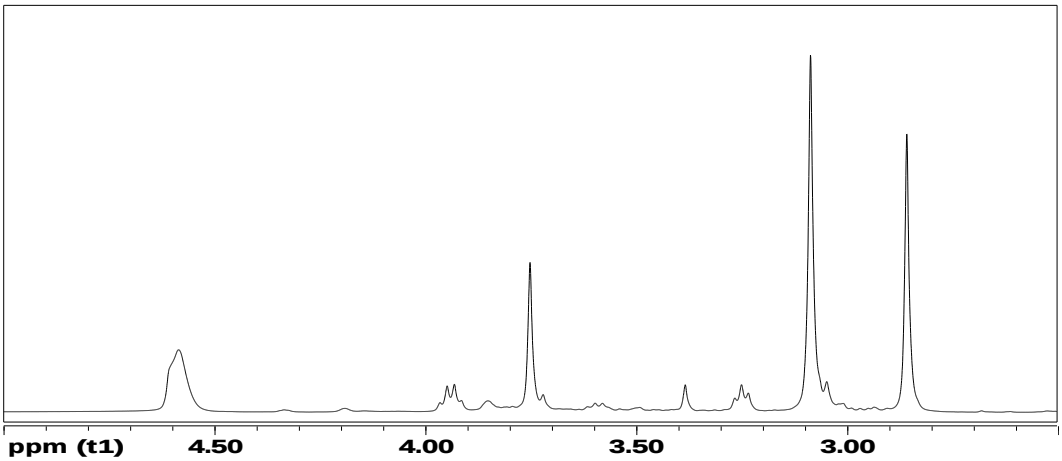


Figure Appendix B3: VL-T0: Midfield Region (5 < ppm < 2.5)

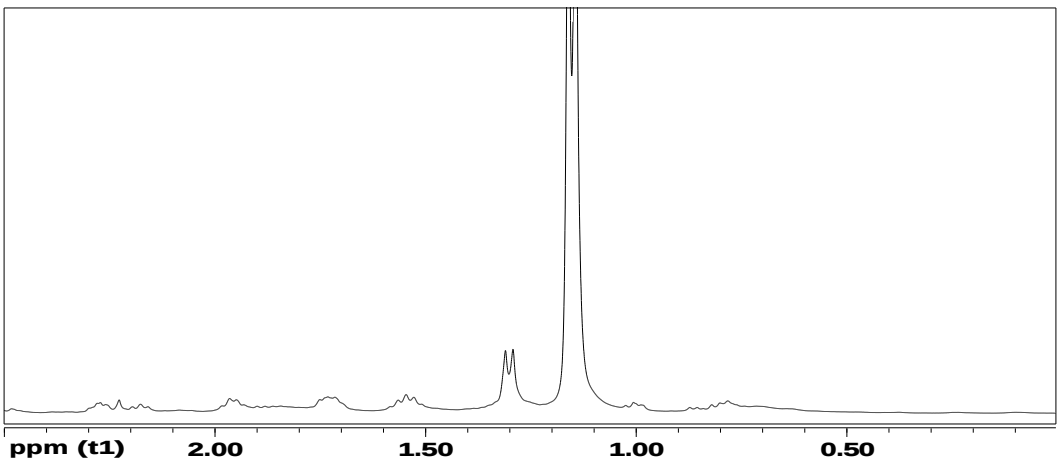


Figure Appendix B3: VL-T0: Upfield Region (< 2.5 ppm)

Appendix B.4 Gilthead Sea bream – Lagoon – T_{16} (VL-T16) spectrum

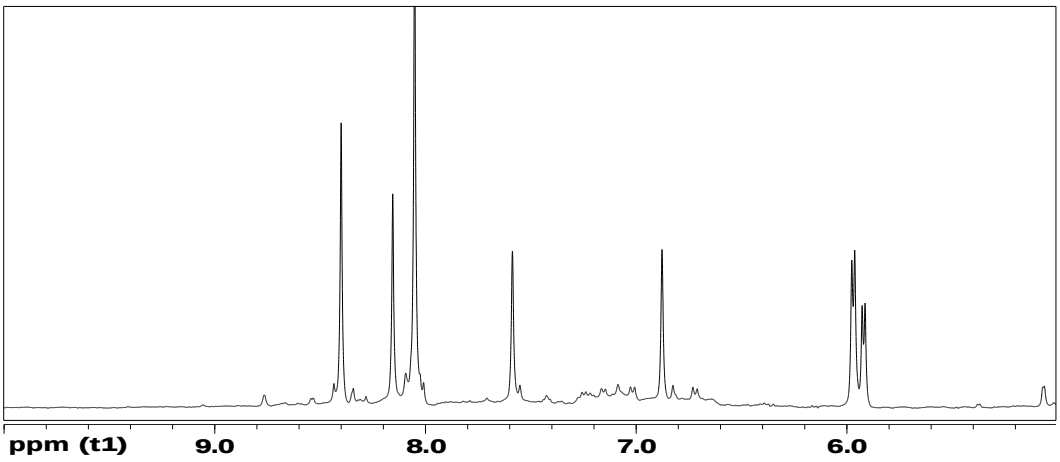


Figure Appendix B4: VL-T16: Downfield Region (> 5 ppm)

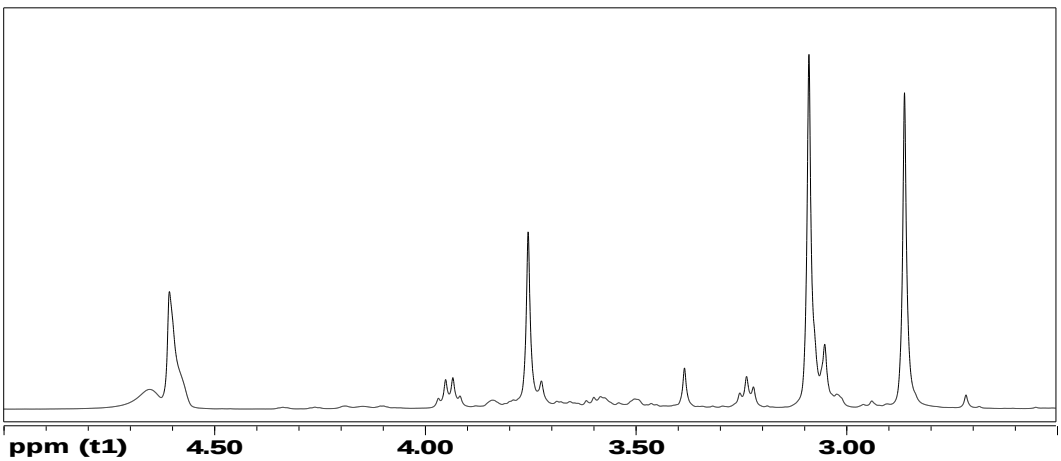


Figure Appendix B4: VL-T16: Midfield Region (5 < ppm < 2.5)

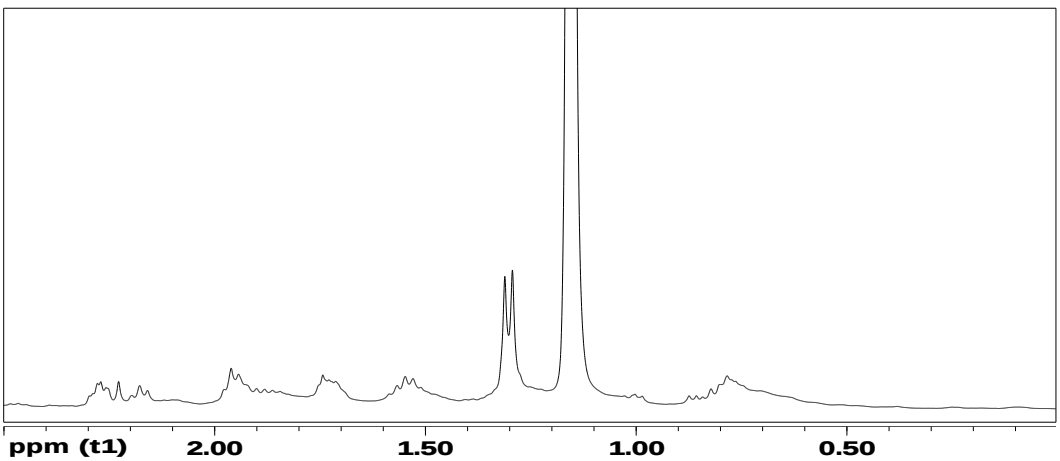


Figure Appendix B4: VL-T16: Upfield Region (> 2.5 ppm)

Appendix B.5 Gilthead Sea bream – Cage – T_0 (GB-T0) spectrum

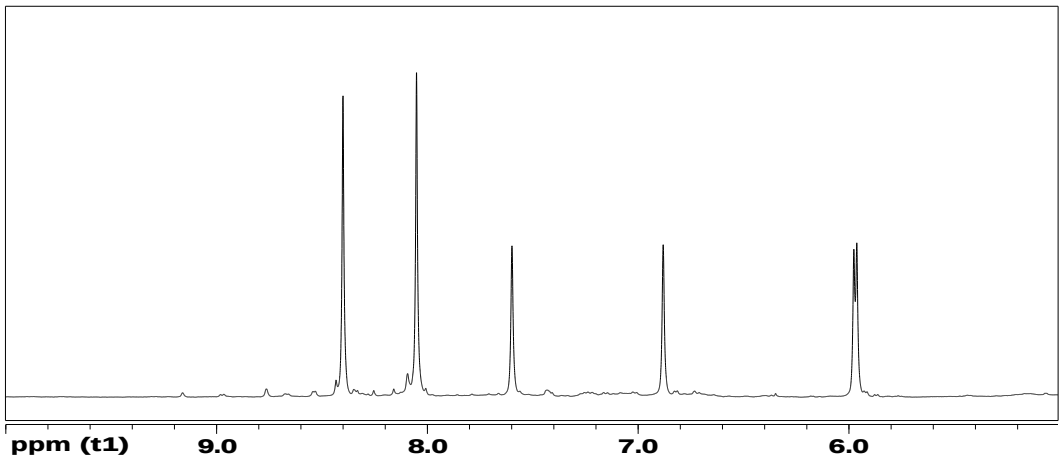


Figure Appendix B5: GB-T0: Downfield Region (> 5 ppm)

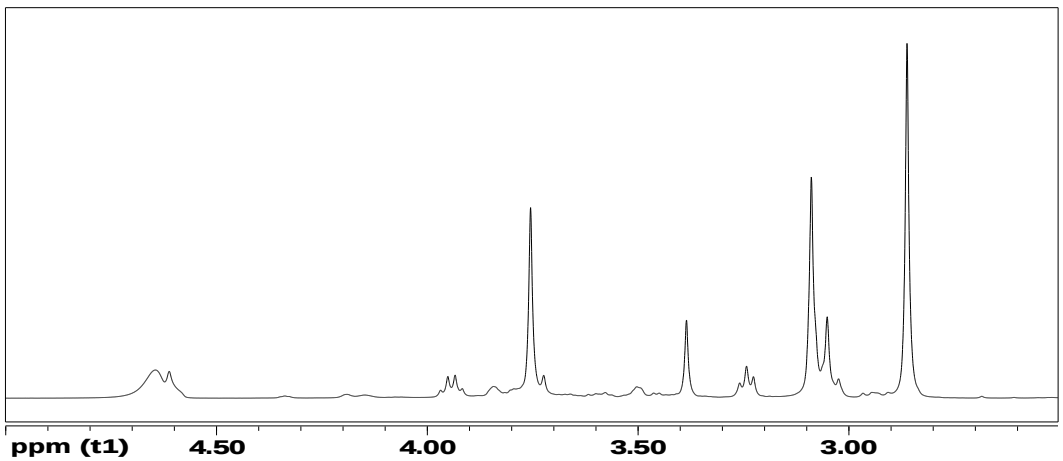


Figure Appendix B5: GB-T0: Midfield Region (5 < ppm < 2.5)

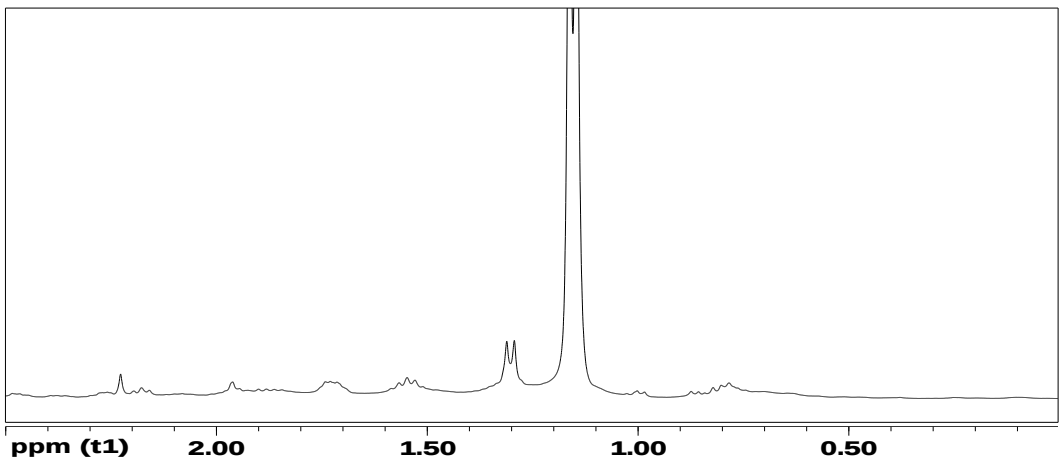


Figure Appendix B5: GB-T0: Upfield Region (< 2.5 ppm)

Appendix B.6 Gilthead Sea bream – Cage – T_{16} (GB-T16) spectrum

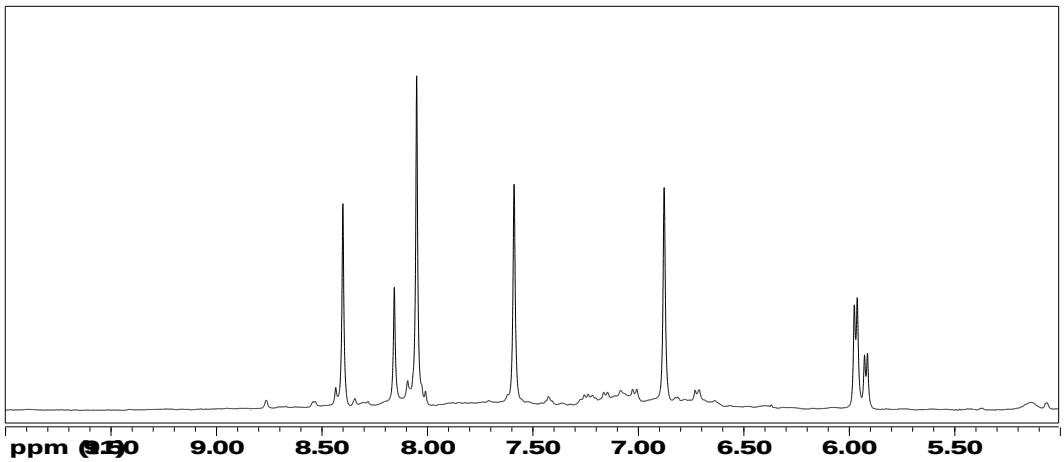


Figure Appendix B6: GB-T16: Downfield Region (> 5 ppm)

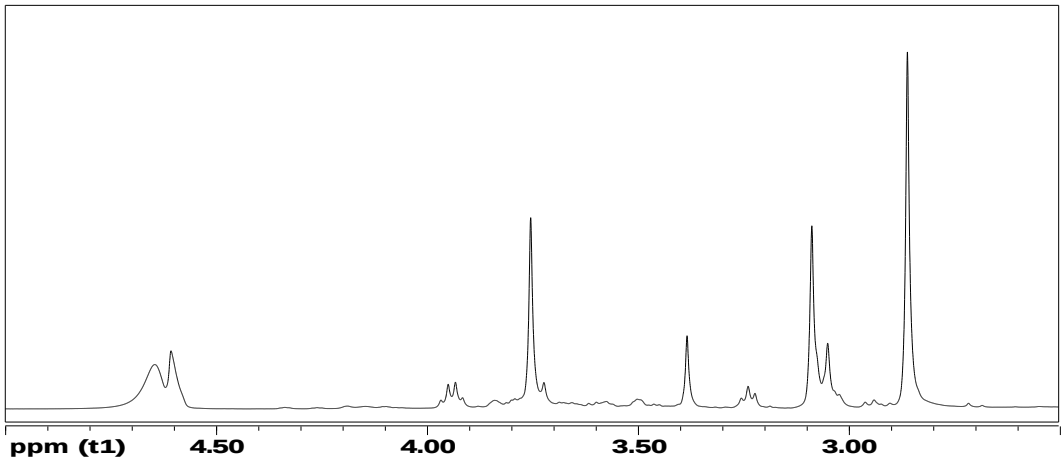


Figure Appendix B6: GB-T16: Midfield Region (5 < ppm < 2.5)

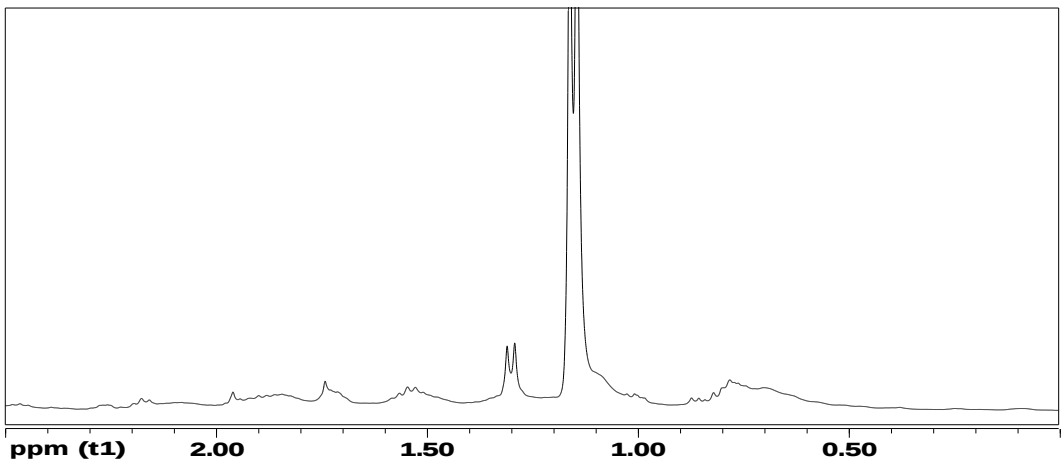


Figure Appendix B6: GB-T16: Upfield Region (> 2.5 ppm)