University of Florence

International Doctorate in Structural Biology

Cycle XXII (2007-2009)

Metabolomics studies by NMR

Ph.D. thesis of

Stefano Nepi

Tutor	Coordinator
Prof. Claudio Luchinat	Prof. Claudio Luchinat

S.S.D. CHIM/03

This thesis has been approved by the University of Florence,

the University of Frankfurt and the Utrecht University

A Cristina

Ai miei genitori

At last gleams of light have come,

and I am almost convinced (quite contrary to opinion I started with)

that species are not (it is like confessing a murder) immutable.

Charles A. Darwin

1 INTRODUCTION

Metabolomics in the system biology

The decoding of human and other mammalian genomes during the 1990s caused a strong evolution in the field of molecular biology, giving birth to new fields of science, called "omics" sciences. All these new "omics" sciences provided a great number of biological features about complex systems, giving the scientific community a new point of view on this life form. Mammalian animals and men can presently be considered as "superorganisms" (1), in which environmental and lifestyle effects completely influence biomolecular organization. Such factors can also modify gene and protein expression and metabolites levels, leading to differences between various individuals *(inter-individual)*, but also inside the same subject *(intra-individual)*. Moreover, also the symbiotic gut microflora interacts with the host with a specific metabolism and a genome that usually is not well-known. This great complexity leads to the definition of "Global System Biology" (2) to highlight the necessity to integrate multivariate biological information to better understand the behaviour of the so called superorganism. The first born of these new sciences is transcriptomics, which studies the determination of gene expression changes between subjects (3), and proteomics, which deals with the determination of all protein expression changes in a cell or tissue (4). In this context both **metabolomics** and **metabonomics** are defined (see **Figure 1.1**). The term metabonomics is defined as "*the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification"* (5), while the term metabolomics is introduced later as the *"comprehensive and quantitative analysis of all metabolites"* in a system (6). Even if there is a difference between the two terms, they are almost considered as equivalent and used interchangeably by the scientific community (1) as well as in this text. Though the term metabonomics was coined many years later, the concept of metabonomics was born with the first simultaneous analysis of metabolites present in biological fluids through ^{1}H NMR spectroscopy in the 1980s (7). The data obtained from this kind of analysis are very complex and they were interpreted by using a multivariate statistics approach to classify the samples according to their biological status (8-9).

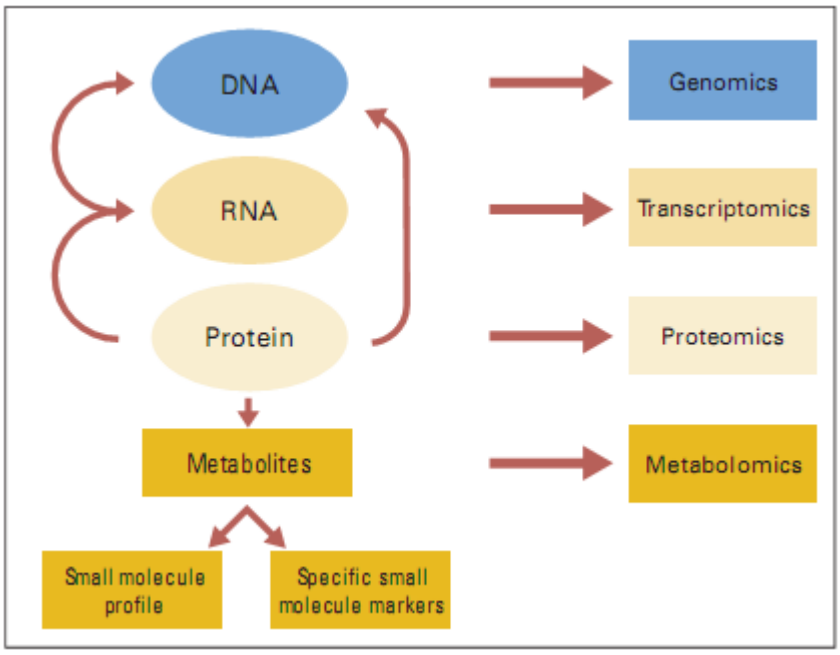


Figure 1.1 The "omics" sciences.

Most metabolomics studies involved common biofluids as urine and serum/plasma that are easily obtainable from mammalians, especially from humans. Moreover, these two types of biofluids are obtained in a non- or poorly-invasively way and they are easy to find and collect because they are commonly used for many other biological analyses. Hence they can be used for disease diagnosis and in a clinical trials setting for monitoring drug therapy. Besides, all biological fluids are useful for metabolomics analysis, e.g. cerebrospinal fluid, synovial fluid, exhaled breath condensate, saliva and so on (10). There are some problems associated to the use of these peculiar fluids for which they are not normally analyzed: poor collectable quantity, highly-invasive extraction techniques, poor metabolites information carried on. In addition, some metabolomics studies have used tissue samples and their aqueous or lipid extraction (11) or *in vitro* cell systems (12). The number of different metabolites in these human fluids is unknown; estimates range from a minimum of 2,000 to 3,000 to a maximum of around 20,000 metabolites, compared with an estimated 23,000 genes and 60,000 proteins (13). Small metabolites, that are low-molecular weight compounds that serve as substrates and products in various metabolic pathways, are of particular interest to metabolomics researchers. These small molecules include compounds such as lipids,

sugars, and amino acids, as well as bioactive products acting at very low concentrations in tissue signalling functions (14).

Analytical technologies in metabolomics

The main analytical techniques employed in metabolomics are nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS). NMR spectroscopy is a non-destructive technique that provides detailed information on molecular structure of both pure compounds and complex mixtures (15), and for these reasons can be applied to metabolomics studies. High-throughput NMR spectroscopy consists of an automatic sample preparation step and a fast ^{1}H NMR analysis, about eight minutes for each sample. The automatic sample preparation involves addition of buffer and D_2O, as magnetic field lock signal for spectrometer; it enables to prepare a significant number of samples in a short time. The obtained spectra profiles are essentially the superposition of ^{1}H NMR spectra of thousands different small molecules (up to 2500 for urine, up to 200 for serum/plasma) present in sample at concentration >1 μM (16). A typical ^{1}H NMR spectrum is obtained using water suppression techniques and adding TSP (sodium trimethylsilyl [2,2,3,3-$^{2}H_4$]propionate) as internal reference, it contains for urine predominantly sharp lines due to small molecules, while for serum and plasma both broad and sharp signals are present due respectively to macromolecules, as lipoproteins, and low molecular metabolites. These broad signals of macromolecules can be suppressed by applying a Carr-Purcell-Meiboom-Gill (CPMG) filter to a standard 1D sequence. One of the principal disadvantages of the NMR approach is the difficult identification of all metabolites in the samples, a process that involves a large number of techniques like two-dimensional NMR experiments. Indeed the ^{1}H NMR of biological fluids is very complex and even though many resonances can be directly assigned basing on chemical shifts, multiplicity and by addition test, various two-dimensional NMR experiments are necessary to increase, but not to complete, the identification of biomarkers in biofluids. These 2D experiments include: i) ^{1}H-^{1}H J-resolved (J-res), to attenuate macromolecules signals and to give more information about multiplicity and coupling patterns ii) ^{1}H-^{1}H correlation spectroscopy (COSY) and ^{1}H-^{1}H total correlation spectroscopy (TOCSY), to provide ^{1}H-^{1}H spin-spin coupling connectivities iii) various

heteronuclear experiments that use information coming from other types of nuclei as ^{13}C, ^{15}N and, if it is present, ^{31}P, as for example ^{13}C-^{1}H HSQC, to obtain information on the direct coupling ^{13}C-^{1}H.

Mass spectrometry is a destructive technique that requires a very poor quantity of samples. Over the last few years its application to mammalian study has been increasing, especially for its great sensitivity, higher than NMR, and because it is a major technique for molecular identification, through the use of tandem methods for fragmention studies or of Fourier transform MS for a very accurate mass determination (1). Mass is usually coupled with chromatographic techniques as GC (Gas Chromatography) and LC (Liquid Chromatography) to separate different classes of substances (17). The process of ionization, for biofluids as urine, is usually generated by the electron spray (ESI) method, and then both positive and negative chromatogram are measured. A three-dimensional chromatogram (retention time, mass, intensity) is generated by applying the HPLC (or GC)-MS approach. The great advantage is the possibility to cut off any mass peak from interfering substance (as for example drug metabolites or contaminant) without altering the dataset (1).

It is possible to affirm that the two techniques are complementary for fully characterization: even if MS is better for identification of metabolites and for sensitivity, NMR is particularly useful for distinguishing isomers, informing on molecular conformation and studying molecular dynamics (1). Moreover, both NMR and MS techniques are being improved and developed.

The sensitivity problems of NMR are partially solved by using cryogenic probes, where the detector coil and the pre-amplifier are cooled around 20 K, increasing the signal-noise ratio (18); moreover, the recent development of a technique called high resolution ^{1}H magic angle spinning (MAS) made feasible the acquisition of data on little pieces of tissue without any treatment; indeed, with the rapid spinning of the sample at an angle of 54.7° relative to magnetic field applied, it is reduced the line broadening effect and the associated loss of information (19-21). As far as MS metabolomics is concerned, the introduction of UPLC (Ultra Performance Liquid Chromatography) enabled better peak resolution and further increase in sensitivity and speed analysis

and it is still successfully applied to metabolomics studies (as UPLC-MS) (22). Finally, some "hyphenated" approaches like HPLC-NMR-MS (23), in which each eluting HPLC peak is split to enable a parallel analysis by NMR and MS, are still used for a complete metabolic identification.

Metabolomics and disease

However, the strongest point of NMR metabolomics is associated to the capability to analyze, at the same time, all the metabolites present in biofluids with this approach. It is possible to obtain information about the status of a sample (and therefore of a subject) through the analysis of the so- called **fingerprint**. In NMR metabolomics the term fingerprint usually indicates the complete NMR profile of each analyzed sample. Applying statistical techniques, various samples are comparable through the exam of their fingerprint, without requiring the knowledge of all metabolites. This constitutes a true revolution to identify metabolic differences among various samples. Indeed metabolomics is still used for many purposes as for example to relate toxic or therapeutic effects of xenobiotics to normality (24), but one of the principal application of metabolomics is to aid human disease diagnosis. Several inborn errors of metabolism in children, for instance, are diagnosed with the use of NMR spectroscopy of urine and serum (25). Moreover, in the last few years the use of metabolomics approach has increased to provide significant information on a wide range of pathologies, as cancer (26), ischemia-riperfusion (27), meningitis (28), diabetes (29), neurological disorders (30), liver fibrosis and cirrhosis (31), inflammatory bowel disease (32) and so on. Exploring pathologies through the holistic metabolomics approach can be very important to get new lights inside their mechanism and their impact on humans and animals. Nevertheless, it is possible to obtain a fast and non invasive diagnosis *(pre-diagnosis)* of some of them, as it is observable by reading a work about coronary artery disease (33), in which the sensitivity and the specificity of the test is either 92% or 93%, and therefore this allows to distinguish between normal coronary artery subjects and triple coronary vessel disease subjects with an high value

of accuracy. Sensitivity, specificity and accuracy are biostatistical parameters well explained in **Table 1.1**.

While for angiogenesis the complete fingerprint of the illness is the responsible of the discrimination between healthy and sick subjects, it is also possible to highlight specific biomarkers for a selected pathology, or specific altered pathways, with metabolomics. This can be achieved without *a priori* knowledge about molecular characteristics. Therefore, metabolomics can be also used as a screening for markers and, once these are identified, can lead to the successive development of specific kits for a cheaper and extensive diagnosis of diseases.

		Condition (as determined by **GOLD STANDARD**)		
		Positive	Negative	
Test Outcome	Positive	True Positive (TP)	False Positive (FP) (Type I error)	Positive Predicted Value
	Negative	False Negative (FN) (Type II error)	True Negative (TN)	Negative Predicted Value
		Sensitivity (TP/FN+TP)	**Specificity (TN/FP+TN)**	**Accuracy (TP+TN/TP+TN+FP+FN)**

Table 1.1 Biostatistical parameters.

Future Applications

One of the long-term goals of metabolomics is clearly the understanding of the existing relationships between genetic polymorphisms of different individuals and their metabolic fingerprint, in order to completely understand the response of different organisms to external stimuli. The achievement of this aim could be very important in the field of *pharmacometabolomics,* leading to personalized healthcare (34). Thus, an individual's drug treatment will be tailored so as to achieve maximal efficacy and avoid adverse drug reactions. In order to solve this purpose it is necessary to exactly determine the genetic and environmental influences on the basal metabolic fingerprint of an individual, since these will also influence the outcome of a chemical intervention. In this direction a great improvement is obtained with the demonstration of the existence of unique individual metabolic phenotypes, called *metabotypes* (35-36).

Moreover, there is a wide range of current and emerging applications of metabolomics: i) investigation of invertebrates tissue extracts and biofluids as monitors of environmental toxicity (37-38) ii) investigation of plant biofluids (39) iii) studying of relationship between specific changes in gene expression and alteration of biochemical process (functional genomics) (40) iv) epidemiological study on large cohorts of biofluids samples to highlight differences in metabolism of various populations across the globe (41).

Chemometrics methods

Once the spectra of each biofluid are collected, it is necessary to analyze these data. Both NMR and mass spectrum can be thought of as a multidimensional set of metabolic coordinates, whose values are the spectral intensity at each data point (ppm for NMR and retention time for mass). Starting from these data, the principal aims of metabolomics can be summarized in four goals: i) visualization of overall differences, trends and relationships between samples and variables, ii) determination of whether there is a significant difference between groups, for instance between healthy subjects and sick subjects for a pathology, iii) highlighting all metabolites that are responsible for these differences and, maybe a little less important, iv) construct a predictive model for new samples. Multivariate statistical analyses are the key to achieve these goals. At present with the NMR spectra the variables are obtained by “bucketing”. Bucketing is a procedure used to reduce the total number of variables. One bucket (or bin) is a little slice of spectrum. The corresponding intensity of this slice is calculated in order to obtain our primary variables for each bucket. Obviously, the size of the buckets is one of the parameter to choose, even if 0.02 ppm is the commonest size.

One of the simplest multivariate techniques largely used in metabolomics is PCA (Principal Component Analysis). PCA is a linear technique that expresses the maximum of variance in a data set through a small number of factors called Principal Components (PCs). Each PC is obtained as a linear combination of the primary variables (buckets). The first PC (PC1) expresses the maximum percentage of variability, then the value of variance quickly decreases in a way that the first PCs are the responsible for the great part of variability, while the last PCs are significantly less

important and express noise variability. Moreover, each PC is orthogonal and therefore independent with respect to the others. The conversion of data matrix into PCs gives two new matrices: score matrix and loading matrix. The scores express the coordinates for the samples in the model, indeed every dot in a score plot represents a single spectrum, and may be considered the new variable. The loadings represent the way in which the buckets are linearly combined. Hence, in the loading plot, each point represents a different spectral intensity and how the old variables weight to discriminate between samples, practically which buckets are responsible for the maximum variance.

Moreover, in metabolomics, it is possible to apply also supervised methods of analysis. The most used of these methods is PLS (Partial Least Square) (42). PLS relates a matrix containing data that are independent from the samples, such as spectral intensity, called X matrix, with a matrix containing dependent variables (Y matrix), such as information about the nature of the samples (healthy or sick in case of studies of pathologies). In order to obtain valid data from PLS and, generally, from all supervised methods, it is necessary to split the samples into two sets, called “training set” and “validation set”. The training set is used to build the mathematical model which is used to analyze the validation set. Naturally many random and variable training sets and validation sets can be built to correctly and fully investigate the data through a robust model. One recent modification of the PLS method is constituted by the OPLS (Orthogonal Projection to Latent Structures method) (43), that is used to remove irrelevant and confusing parameters. The basic idea of OPLS is to separate the systematic variation in X matrix in two parts, one linearly related to Y matrix, and the other unrelated (orthogonal) to Y. This partition facilitates model interpretation and model execution on a new set of samples. Both OPLS and PLS can be used combined with DA (Discriminant Analysis) to establish the optimum of discrimination between samples.

Only linear methods are described above. These kinds of methods are the most used in metabolomics, on the other hand the distribution of metabolic data is often not well approximated by a multivariate normal, due to non-normality in the distribution of the

single metabolites or, more commonly, to the combination of several groups of normally distributed metabolites. Therefore the application of PCA, or corresponding supervised methods like PLS, can bring to the lack of information. For these reasons the application of various non-linear methods of multivariate analysis can be very useful in metabolomics. However, it is necessary to consider that non-linear methods employ more tunable parameters, which can lead to a difficult interpretation of data and to a reduction of model robustness. Moreover, they are usually more susceptible to overfitting and to the effect of noise. Therefore, even if their application is theoretically very useful, they have to be applied with extreme caution. Naturally also these methods are divided into unsupervised (as PCA) and supervised (as PLS); a brief description of some of them is presented below.

HCA (Hierarchical Cluster Analysis) It is widely used in all areas of science and it is recently applied to a metabolomic study to explore a set of 20 toxicology studies (44). These methods cluster the data forming a tree diagram (also called dendrogram), in which the relationships between samples are expressed. The algorithm starts calculating the distance between all pairs of data points and, then, proceeds finding the closest pairs of cluster at each iteration (initially each cluster consists of a single data point). The main problem in the application of these methods to metabolomic studies is that the reproducibility is not good,; the inclusion of new data,for instance, requires a complete re-computation of the dendrogram that can lead to a new structure not necessarily similar to that generated from the previous training set. Moreover, HCA does not generate diagnostic information about what features are responsible for the classification in sub-clusters.

SOMs (Self-Organising Maps) The SOMs is an unsupervised method of classification that reduces the dimensionality of the data through the creation of an array of nodes, each one described by a “codebook” vector. During the training, each sample is presented to the map and the node with the closest reference vector is selected as the final node for the sample (also called “winning node”). Clearly the number of chosen nodes is a fundamental variable; with few nodes, for instance, the map does not represent faithfully the data distribution, while with a large number of nodes

phenomena as overfitting and noise susceptibility influence the analysis. Moreover, another disadvantage is associated to the final position of similar clusters, that may end up in distant parts of the SOM despite their similarity. For all these reasons they are not widely used in metabolomics, even if in a study on breast cancer SOMs are applied in combination with kNN (k-Nearest Neighbour) classification (45). In R software some scripts to apply supervised SOMs have been developed.

K-means clustering This unsupervised algorithm is widely used in some fields as transcriptomics. It starts to work selecting the desired number of clusters and randomly assigning their center. For each iteration data points are classified through the assignment to one of the clusters, basing on the closest cluster center. After this addition the new cluster centres are recomputed. Despite their popularity, k-means are not often applied to metabolomics. The main reasons are the absence of diagnostic tools and associated visualization. In order to avoid these disadvantages this technique can be applied in combination with other statistical procedures, as in a metabolomic study of plant and marine invertebrate extracts with HCA (46).

kNN (k-Nearest Neighbour) classification This is the simplest of all supervised classification approaches. Every sample is classified to the class most frequently expressed among the *k* nearest neighbour. Therefore the *k* is the fundamental parameter to set. Small values of *k* can lead to the construction of a model subject to significant statistical fluctuations, while large values of *k* reduce statistical errors but can smooth out many details of the class distribution. As *k*-means, this method gives a classification of the sample without associated visualization and interpretation of data and, therefore, it is not applied only in metabolomics (46,47).

ANNs (Artificial Neural Networks) ANNs is one of the most popular supervised method for pattern recognition in biomedical area. It consists of a network of nodes. Each of these nodes performs an operation to give a single output. Theoretically the nodes can be divided into three layers: i) input nodes ii) hidden nodes and iii) output nodes. Generally in metabolomic studies each input node is formed by each single spectrum, while each output node represents each class of samples (even if it is better to have one more output node corresponding to an unknown class). Each hidden node

receives some information from various input nodes and re-combines this information using a non-linear activation function to give a signal to the output nodes. For their great versatility the ANNs are still applied to some metabolomic studies (48-51), even if this technique is not so easy to use. Indeed, some experience is required to select the optimal architecture (for instance the number of hidden nodes) in order to avoid error in spectra classification. Moreover, it is often difficult to clearly understand the connection weights to completely understand the parameters (buckets in our case) responsible for identification.

Bibliography

(1) Lindon J. C., Nicholson J. K., Holmes E.; *The Handbook of Metabonomics and Metabolomics* **2007** p.ix Elsevier.

(2) Nicholson J. K., Wilson I. D.; *Understanding "global" system biology: Metabonomics and the continuum of metabolism* Nat. Rev. Drug Disc. **2003** 2 668:76.

(3) Baldi P., Hatfield G. W.*; DNA microarrays and gene expression* **2002** p.230 Cambridge University Press.

(4) Wasinger V. C., Cordwell S. J., Cerpa-Poljak A., Yan J. X., Gooley A. A., et al.; *Progress with gene-product mapping of the Mollicutes: Mycoplasma Genitalium* Electrophoresis **1995** 16(7) 1090:4.

(5) Nicholson J. K., Lindon J. C., Holmes E.; *"Metabonomics" : understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data* Xenobiotica **1999** 29 1181:9.

(6) Fiehn O.; *Metabolomics – the link between genotypes and phenotypes* Plant Mol. Biol. **2002** 155:71.

(7) Nicholson J. K., Wilson I. D.; *High-resolution proton magnetic resonance spectroscopy of biological fluids* Prog. NMR Spectrosc. **1989** 21 449:501.

(8) Gartland K. P. R., Sanins S. M., Nicholson J. K., Sweatman B. C., Beddell C. R., et al.; *Pattern recognition analysis of high resolution ^{1}H NMR spectra of urine. A nonlinear mapping approach to the classification of toxicological data* NMR Biomed. **1990** 3 166:72.

(9) Gartland K. P. R., Beddell C. R., Lindon J. C., Nicholson J. K.; *The application of pattern recognition methods to the analysis and classification of toxicological data derived from proton NMR spectroscopy of urine* Mol. Pharmacol. **1991** 629:42.

(10) Lindon J. C., Nicholson J. K., Everett J. R.; *NMR spectroscopy of biofluids* Ann. Reports on NMR Spectrosc. **1999** 38 1:88 Webb G. A.

(11) Mayr M., Chung Y. L., Mayr U., Yin X. K., Ly L.; et al.; *Proteomics and metabolomic analyses of atherosclerotic vessels from apolipoprotein E-deficient mice reveal alterations in inflammation, oxidative stress, and energy metabolism* Arterioscl. Thromb. Vasc. Biol. **2005** 25 2135:42.

(12) Waters N. J., Holmes E., Waterfield C. J., Duncan Farrant R., Nicholson J. K.; *NMR and pattern recognition studies on liver extracts and intact livers from rats treated with α–naphthylisothiocyanate* Biochem. Pharmacol. **2002** 64 67:77.

(13) Claudino W. M., Quattrone A., Biganzoli L., Pestrin M., Bertini I., et al.; *Metabolomics: available results, current research projects in breast cancer, and future applications* J. Clin. Onc. **2007** 25 2840:6.

(14) German J. B., Hammock B. D., Watkins S. M.; *Metabolomics: building on a century of biochemistry to guide human health* Metabolomics **2005** 1 3:9.

(15) Claridge T. D. W.; *High-Resolution NMR techniques in organic chemistry* **2004** p. 384 Elsevier.

(16) Wishart D. S., Tzur D., Knox C., Eisner R., Guo A. C., et al.; *HMDB: the Human Metabolome Database* Nucleic Acids Res. **2007** 35 D521:6.

(17) Yang J., Xu G., Zheng Y., Kong H., Pang T., et al.; *Diagnosis of liver cancer using HPLC-based metabonomics avoiding false-positive result from hepatitis and hepatocirrhosis diseases* J. Chromatogr. B **2004** 813 59:65.

(18) Tomlins A., Foxall P. J. D., Lindon J. C., Lynch M. J., Spraul M., et al.; *High resolution magic angle spinning ^{1}H nuclear magnetic resonance analysis of intact prostatic hyperplastic and tumour tissues* Anal. Commun. **1998** 35 113:5.

(19) Keun H. C., Beckonert O., Griffin J. L., Richter C., Moskau D., et al.; *Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies* Anal. Chem. **2002** 74 4588:93.

(20) Garrod S. L., Humpfer E., Spraul M., Connor S. C., Polley S., et al.; *High-resolution magic angle spinning ^{1}H-NMR spectroscopic studies on intact rat renal cortex and medulla* Magn. Res. Med. **1999** 41 1108:18.

(21) Cheng L. L., Chang I. W., Louis D. N., Gonzalez R. G.; *Correlation of high-resolution magic angle spinning proton magnetic resonance spectroscopy with histopathology of intact human brain tumor specimens* Cancer Res. **1998** 58 1825:32.

(22) Wilson I. D., Nicholoson J. K., Castro-Perez J., Granger J. H., Johnson K. A., et al.; *High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies* J. Proteome Res. **2005** 4 591:8.

(23) Lindon J. C., Nicholson J. K., Wilson I. D.; *Directly-coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development* J. Chromatog. B. **2000** 748 233:58.

(24) Lindon J. C., Holmes E., Nicholson J. K.; *Toxicological applications of magnetic resonance* Prog. NMR Spectrosc. **2004** 45 109:43.

(25) Moolenaar S. H., Engelke U. F. H., Wevers R. A.; *Proton nuclear magnetic resonance spectroscopy of body fluids in the field of inborn errors of metabolism* Ann. Clin. Biochem. **2003** 40 16:24.

(26) Teichert F., Verschoyle R. D., Greaves P., Edwards R. E., Teahan O., et al.; *Metabolic profiling of transgenic adenocarcinoma of mouse prostate (TRAMP) tissue by 1H-NMR analysis: evidence for unusual phospholipid metabolism* Prostate **2008** 68 1035:47.

(27) Coolen S. A., Daykin C. A., van Duynhoven J. P., van Dorsten F. A., Wulfert F., et al.; *Measurement of ischaemia-reperfusion in patients with intermittent claudication using NMR-based metabonomics* NMR Biomed. **2008** 21 686:95.

(28) Coen M., O'Sullivan M., Bubb W. A., Kuchel P. W., Sorrell T.; *Proton nuclear magnetic resonance-based metabonomics for rapid diagnosis of meningitis and ventriculitis* Clin. Infect. Dis. **2005** 41 1582:90.

(29) Fearnside J. F., Dumas M. E., Rothwell A. R., Wilder S. P., Cloarec O., et al.; *Phylometabonomic patterns of adaptation to high fat diet feeding in inbred mice* PLoS ONE **2008** 3 e1668.

(30) Bartsch T., Alfke K., Wolff S., Rohr A., Jansen O., et al.; *Focal MR spectroscopy of hippocampal CA-1 lesions in transient global amnesia* Neurology **2008** 70 1030:5.

(31) Constantinou M. A., Theocharis S. E., Mikros E.; *Application of metabonomics on an experimental model of fibrosis and cirrhosis induced by thioacetamide in rats* Toxicol. Appl. Pharmacol. **2007** 218 11:9.

(32) Marchesi J. R., Holmes E., Khan F., Kochhar S., Scanlan P., et al.; *Rapid and noninvasive metabonomic characterization of inflammatory bowel disease* J. Proteome Res. **2007** 6 546:51.

(33) Brindle J. T., Antti H., Holmes E., Tranter G., Nicholson J. K., et al.; *Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using 1H-NMR-based metabonomics* Nature Med. **2002** 8 1439:45.

(34) Clayton T. A., Lindon J. C., Antti H., Charuel C., Hanton G., et al.; *Pharmaco-metabonomic phenotyping and personalised drug treatment* Nature **2006** 440 1535:42.

(35) Gavaghan C. L., Holmes E., Lenz E., Wilson I. D., Nicholson J. K.; *An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57BL10J and Alpk:ApfCD mouse* FEBS Lett. **2000** 484 169:74.

(36) Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H., et al.; *Evidence of different metabolic phenotypes in humans* Proc. Natl. Acad. Sci. U.S.A. **2008** 105 1420:4.

(37) Griffin J. L., Walker L. A., Troke J., Osborn D., Shone R. F., et al.; *The initial pathogenesis of cadmium induced renal toxicity* FEBS Lett. **2000** 478 147:50.

(38) Griffin J. L., Walker L. A., Shore R. F., Nicholson J. K.; *High-resolution magic angle spinning 1H-NMR spectroscopy studies on the renal biochemistry in the bank vole*

(Clethrionomys glareolus) and the effects of arsenic (As^{3+}) toxicity Xenobiotica **2001** 3 377:85.

(39) Bino R. J., Hall R. D., Fiehn O., Kopka J., Saito K., et al.; *Potential of metabolomics as a functional genomics tool* Trends Plant Sci. **2004** 9 418:25.

(40) Fiehn O.; *Combining genomic, metabolome analysis, and biochemical modeling to understand metabolic networks* Comp. Funct. Genom. **2001** 2 155:68.

(41) Barton R. H., Nicholson J. K., Elliott P., Holmes E.; *High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study* Int. J. Epidem. **2008** 37 1 37:40.

(42) Wold S., Eriksson L., Sjöström M.; *PLS in chemistry, encyclopedia of computational chemistry* **1998** 2006:16 John Wiley and Sons.

(43) Trygg J., Wold S.; *Orthogonal projections to latent structures (O-PLS)* J. Chemometr. **2002** 16 119:28.

(44) Beckonert O., Bollard E., Ebbels T. M. D., Keun H. C., Antti H., et al.; *NMR-based ,etabonomic toxicity classification: Hierarchical cluster analysis and k-nearest-neighbor approaches* Anal. Chim. Acta **2003** 490 3.

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

(46) Pierens G. K., Palframan M. E., Tranter C. J., Carroll A. R., Quinn J. R.; *A robust clustering approach for NMR spectra of natural products extracts* Magn. Reson. Chem. **2005** 43 359:65.

(47) Oust A., Moretro T., Kirschner C., Narvhus J. A., Kohler A.; *FT-IR spectroscopy for identification of closely related lactobacilli* J. Microbiol. Methods **2004** 59 149:62.

(48) Howells S. L., Maxwell R. J., Peet A. C., Griffiths J. R.; *An investigation of tumor 1H nuclear magnetic resonance spectra by the application of chemometric techniques* Magn. Reson. Med. **1992** 28 214:36.

(49) Maxwell R. J., Martinez-Perez I., Cerdan S., Cabanas M. E., Arus C., et al.; *Pattern recognition analysis of 1H NMR spectra from perchloric acid extracts of human brain tumor biopsies* Magn. Reson. Med. **1998** 39 869:77.

(50) El-Deredy W., Ashmore S. M., Branston N. M., Darling J. L., Williams S. R., et al.; *Pretreatment prediction of the chemotherapeutic response of human glioma cell cultures using nuclear magnetic resonance spectroscopy and artificial neural networks* Cancer Res. **1997** 57 4196:9.

(51) Lisboa P. J., Kirby S. P., Vellido A., Lee Y. Y., El-Deredy W.; *Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection* NMR Biomed. **1998** 11 225:34.

2 METABOLOMICS STUDIES BY NMR

Metabolomics has established itself as an useful complement to the characterization of pathologies. The metabolome, that is considered the downstream of genome, transcriptome and proteome, is the best representation of a healthy or diseased phenotype of an organism. Indeed, metabolome amplifies changes caused by a biological perturbation. As opposed to metabolomics, which places a greater emphasis on comprehensive metabolic profiling, *metabonomics* is more often used to describe multiple (but not necessarily comprehensive) metabolic changes caused by a biological perturbation. Nuclear Magnetic Resonance (NMR)-based metabonomics offers evident advantages in contrast with knowledge-guided search of metabolites in pathological samples. NMR-based metabonomics makes no assumptions on the identity of the metabolites that are relevant for the selected pathology. Information on the metabolite pattern alterations that can be significantly associated to the pathology is directly obtained through statistical analysis of the NMR profiles. Usually, metabonomics does not rely on the measurement of a single metabolite-associated peak(s) but analyze spectra as whole: metabonomic profiles are essentially the superposition of the ^{1}H NMR (in the most popular version of NMR-based metabonomics) spectra of tens to thousands different small molecules (up to 2500 in the case of urine) present in the sample at > 1 μM concentration.[5] In principle a NMR profile contains qualitative and quantitative information on all of them. Small changes in enzymes concentrations can reflect in considerable alterations in intermediate products and because the fact that metabolic networks are connected by few high concentrated nodes, that can be investigated by NMR-based metabonomic analysis of biological fluids such as serum, plasma, and urine. MetNoMet (comparison between samples coming from the previous three breast cancer projects after starting therapy to highlight possible difference between metastatic and non-metastatic subject) serum.

Practically metabolomics seems to have a staggering diagnostic potential, explaining what actually happens to an organism and not what might happens as genomics or proteomics. Moreover its new approach, that is the contemporary analysis of all metabolites in a biofluid, indipendently of their assignment and their classes (fingerprint analysis) can be very useful to get new hints on various classes of pathologies.

Aim of the work

In this context my research projects have been developed during these three years. The general aim of the research can be splitted in two. First of all I tested the great potential of metabolomics in diagnosis and prognosis of pathologies or as new tools to discover biomarkers

and to understand complex biological mechanism of a pathology. In order to solve this goal, two diseases are selected: celiac disease (CD) and breast ductal carcinoma. CD is chosen because can be considered as a “metabolic” pathology or, at least, a pathology that seriously affects the metabolism of ill subjects. Indeed the mal-absorption associated to the pathology causes significant alteration in the metabolism of sick subjects. On the other hand, breast ductal carcinoma is a pathology with a great impact in the human population, especially in women. Therefore it is studied for many points of view and any contribution is important and fundamental to the definition of its onset and progression. Furthermore it is necessary to remember that almost all kinds of tumors can evolve in the metastatic form. The metastatic cells develop mechanisms in order to pass basal membrane of tissues and, then, the blood stream transfer them from one organ to others. Thus, the metabolomic studies of serum/plasma of subjects affected by a carcinoma is clearly very useful to define a general metastatic risk and to single out a specific fingerprint of the metastasis. Finally the application of metabolomics approach to these (but also to other) pathologies results to be very useful in order to highlight an eventual correlation between metabolic fingerprints and specific pathologies, try to diagnose and evaluate the advancement of a pathology and to evaluate for each subject the risk factors to contract a pathology, single out new biomarkers and study in detail several metabolic pathways.

Nevertheless the metabolomic analysis of healthy subjects is clearly interesting and, therefore, constitutes the second principal aim of my research activity. Essentially the study of healthy subjects is important in metabolomics to have a clear definition of what is the variability. Indeed, metabolomics analyzes all metabolites that are present in a biofluid. Thus, when samples of subjects that have a specific pathology are analyzed, their metabolism and, consequently, their fingerprint is heavily affected by diet, physical activity and, in general, other environmental factors, besides than pathology. In this way the repeated study of urine samples (see chapter 5) coming from different subjects is very important in order to define better inter- and intra-variability and what are the effects on the metabolome of both environmental and genetic factor. Furthermore it is important to consider that the definition of healthy state is not completely clear and unambiguous. Therefore the metabolomics study of healthy individuals can be very useful to characterized their healthy states and to understand if it is possible to divide the healthy subjects basing on their metabolism. This can lead to define some classes of healthy subjects that present different “responses” to various stimuli such as drug intake, diet, physical activity and so on.

Finally each single metabolomic project developed during these three years, presents peculiar goals and aims that can be also significantly different from each other. These heterogeneous goals are separately treated in the each proper project chapter.

3 CELIAC DISEASE AND METABOLOMICS

Celiac Disease

Definition and Etiology

Celiac disease (CD) (also called coeliac disease or sprue) is an autoimmune disorder of the small intestine caused by intolerance to gluten that occurs in people of all ages. It is characterized by immune-mediated enteropathy resulting in mal-digestion and mal-absorption (1). Although no pharmaceutical treatment is actually known to deal with it, a lifelong free-gluten diet is sufficient to reverse the symptoms and to allow the patient to spend a normal life (2). In particular, the pathology is caused by an immune reaction to gliadin, a gluten protein normally present in wheat, barley and rye. Upon exposure to gliadin, the enzyme tissue trans-glutaminase (tTG) modifies the protein, and the immune system cross-reacts with the small-bowel tissue, causing an inflammatory reaction, that leads to a truncating of the villi lining the small intestine (called villous atrophy). This interferes with the absorption of nutrients, because the intestinal villi are responsible for absorption (3). CD is defined as a multi-factorial disorder in which both genetic and environmental factors play a crucial role in pathogenesis (4). Genetically it is associated with specific alleles: HLA-DQ2 and HLA-DQ8. HLA-DQ2 is expressed in more than 90% of people with CD (5). However, the expression of these two alleles is not sufficient to develop CD and, moreover, results from studies on siblings and on homozygous twins suggest that they are not the main causes of the disease onset (6,7).

Prevalence and Incidence

In epidemiology, **prevalence** is a statistical parameter associated to a disease that represents the total number of cases in a population at a given time divided by the number of individuals of the same population. Historically for celiac disease the prevalence was considered of about 0.02% (8), but the introduction of new diagnostic practices caused an increase in this value. At present the prevalence is considered to be between 0.05% and 0.27%, even if for some populations of Southern Europe (as Italian), India and U.S.A. the prevalence is indicated to be between 0.33% and 1.06% in children (for Sahrawi people is 5,66%) and between 0.18% and 1.20% in adults (9-10).

Incidence is a measure of the risk of developing some new conditions within a specified period of time. Some studies report a general decline of incidence during the 1970s and a successive rising during the late 1980s and 1990s contemporary to the development of new methods of analysis and diagnosis (9, 11-13). In general incidence conveys information about the risk of contracting the disease, whereas prevalence indicates how widespread the disease is.

Clinical Manifestation of Celiac Disease

The symptoms of CD are various. Abdominal distension and pain, vomiting, steatorrhea, diarrhea, weight of loss and fatigue are relatively common symptoms in CD patients. Children between nine and twenty-four months tend to present - together with bowel symptoms - also growth problems and pyloric stenosis. The described symptoms are relatively common for gastrointestinal mal-absorption pathologies. However, in case of CD, most patients have a constellation of other clinical manifestations, such as anemia; moreover, several other conditions are described associated to CD as T-cell lymphoma (14-15), osteoporosis (16), neurologic disease (17-18) and some autoimmune disorders (19), like type 1 diabetes (20), autoimmune thyroiditis (21) and Addison's disease (22). Nevertheless, there are two types of CD that do not manifest themselves with significant symptoms or associated pathologies. These two forms of CD are called: **silent CD** and **latent CD**. While in the silent form the patients do not show any symptoms but present alteration of intestinal mucosa (villous atrophy), in the latent form the patients do not show both symptoms and intestinal damage but only have a clear predisposition to the pathology (positivity to anti-gliadin (AGA) and anti-endomysium (EMA) antibodies testing).

Diagnosis and Therapeutics

Several tests could be used to diagnose CD, including serological and endoscopic tests. The biopsy by fiber-optic endoscopy is the test which carries the higher value of sensitivity (about 100%) and lower frequencies of error, due to the possibility to have false positive (specificity is about 61%). In each case endoscopy represents a very invasive test; nowadays it is used to definitively confirm the presence of the CD in

patients with positive serology and/or high-risk symptoms, as weight loss, anemia (more than 120 g/l in females and 130 g/l in males) and diarrhea (23). Serological tests are the first approach used to determine CD. Most used tests are the indirect immune-fluorescence measures of four antibodies : IgA-anti-reticulin (ARA) , IgA-anti-gliadin (AGA) and, above all, IgA-anti-endomysium (EMA) and IgA-anti-transglutaminase (tTG) (24). In parallel total levels of all IgA are checked to avoid the possibility to have false negative results associated to celiac patients with IgA deficiency; in this case IgG antibodies could be useful to diagnose CD (25). These tests have a very high value of sensitivity (over 90%) and specificity (about 99%) (26). Similar values of specificity and sensitivity are associated with anti-tTG test, even if it is quicker and easier than anti-EMA test (27). Although serology appears as the better first approach to diagnose CD it is interesting to note that some cases of apparent seronegative CD occur, even if in presence of normal serum IgA (28).

Historically the only one therapeutical approach in CD is a complete gluten-free diet (2). After a period of diet, varying from a few weeks to some months, all the symptoms, included villi atrophy, are totally reversed. All other tried strategies, for example the use of bacterial prolyl endopeptidase as dietary supplement to degrade anti-gliadin peptides (29), failed *in vivo*.

Metabolomic study of Celiac Disease

Aim of the work

Although CD is a widely studied pathology, many questions still remain unsolved . In particular, some symptoms and effects associated to the pathology are not well explained, for example up to 87% (30) of CD patients present chronic fatigue, that sometimes is the only symptom of undiagnosed CD (31-32); the origin of this syndrome is still unravelled and its attribution to mal-absorption is not sufficient. Clearly CD is a pathology with a direct impact on the metabolism, therefore the metabolomic approach, giving a holistic point of view of the pathologies, could be helpful to better understand some of these unclear CD mechanisms. To solve these purposes of investigation, thirty-four both urine and serum samples of different celiac patients (7

males, 27 females, mean age 38.7 +/- 13.7 years) and thirty-four both urine and serum samples of different healthy subjects (13 males, 21 females, mean age 37.6 +/- 14.8 years) are collected and analyzed. Moreover, it is also interesting the metabolomic analysis of follow-up patients. Normally the reversibility of the CD after the start of the gluten-free diet is only determined by the disappearing of the common symptoms. The analysis of the gut mucosa is not usually done by the patient when the pathology becomes asymptomatic, because is very invasive and tiresome. Thus metabolomic analysis of the follow-up samples should be useful: i) to validate the novel approach ii) to suggest a non invasive method of diagnosis of the CD reversibility. For this purpose urine and serum samples are also collected after three (n=17, 3 males, 14 females, mean age 40.5 +/- 14.1 years), six (n=10, 1 male, 9 females, mean age 38.7 +/- 13.8 years) and twelve (n=13, 2 males, 11 females, mean age 41.3 +/- 11.5 years) months of treatment with a strict gluten-free diet.

Results and Discussion

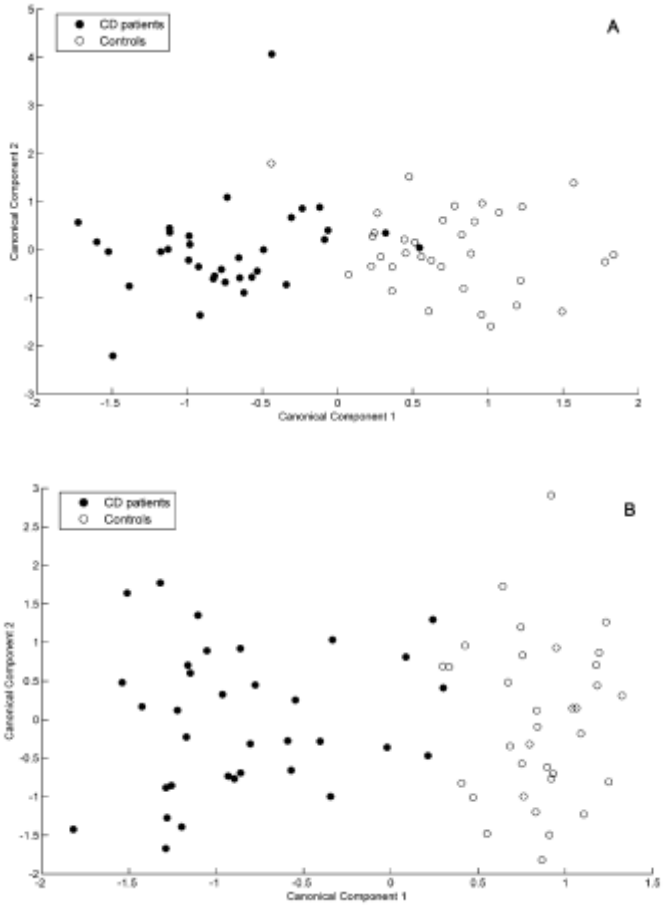
Four different statistical analyses are used to discriminate between celiac patients and healthy subjects (HS). The best discrimination was obtained using six-fold PLS-RCC-SVM method on the test sets with value of accuracy of 94.1% for serum CPMG spectra, 92.6% for serum NOESY spectra and 83.3% for urine NOESY spectra as reported in **Table 2.1**.

	CD patients	HS	sensitivity	specificity	accuracy
Serum CPMG Spectra					
CD patients	32	2	94.1%		94.1%
HS	2	32		94.1%	
Serum NOESY Spectra					
CD patients	30	4	88.2%		92.6%
HS	1	33		97.2%	
Urine NOESY Spectra					
CD patients	24	3	88.9%		83.3%
HS	5	16		76.2%	

Table 2.1 Classification results for sensitivity, specificity and accuracy obtained for serum CPMG and NOESY spectra and urine NOESY spectra.

These high values of discrimination are obtained by a simple statistical analysis and they demonstrate the existence of a metabolic signature for celiac disease. In details,

the lowest values of discrimination for the urine are probably related to the higher day-per-day variability expressed in the metabolic profile of urine samples (33); more surprisingly the values of CPMG serum spectra are higher than Noesy serum spectra. Not according to the relative frequent hypercholesterolemia (34-36) in CD patients, it seems that lipid variations do not contribute to the metabolomic signature of celiac disease. Separations between the two groups are also well visible in the **Figure 2.1**. Focusing the analysis on the CPMG (**Figure 2.1 A**) cluster and discrimination values, it is noticed that only three subjects are mis-clustered and only four subjects are, then, mis-assigned.



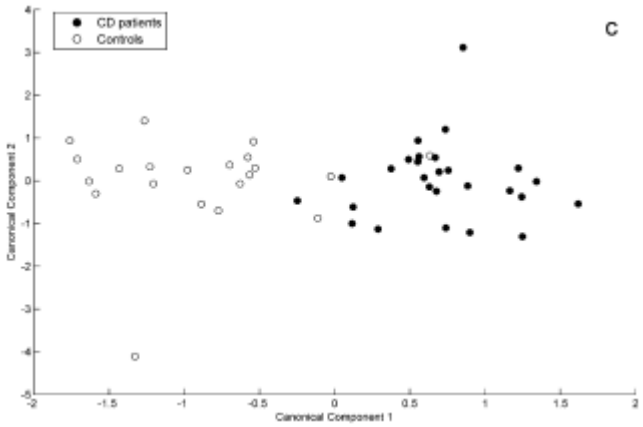


Figure 2.1 Clustering obtained by use of PLS-RCC method on serum CPMG (A) and NOESY (B) spectra, and urine NOESY (C) spectra.

Practically the SVM method mis-classified only one subject that is correctly clustered by PLS-RCC method. Interestingly this subject, which belongs to the group of controls (HS), has a familial history of lymphoma and myeloid leukemia. Regarding the other three subjects, two of them are celiac classified as HS and one is a control classified as celiac disease affected. Both CD subjects are completely asymptomatic, they are detected as celiac during familial screening, while the last mis-classified HS has a history of thyroid carcinoma and presents low levels of folate and ferritin like a great part of CD subjects; moreover, during collection he assumed levothyroxine. To assess which resonance peaks are significantly discriminating between CD subjects and HS, each bucket is analyzed using both ANOVA and non-parametric analogue Kruskal-Wallis test. Discriminating buckets are chosen on the basis of the *P* value, it must be lower than *0.05* (for urine) and *0.01* (for serum), with applying Bonferroni correction. Using this kind of variance analysis many metabolites show significant differences in concentration between two groups (see **Table 2.2** and **Table 2.3**).

	CD patients		HS	
	mean	CI 95%	mean	CI 95%
3-OH-Butyrate	0.7570	0.5926−0.9215	0.4512	0.3348−0.5675
Asparagine	0.0628	0.0563−0.0693	0.0782	0.0709−0.0855
Choline	1.3822	1.2231−1.5413	1.5543	1.3724−1.7361
Creatinine	0.1815	0.1658−0.1973	0.2164	0.1976−0.2352
Glucose	2.0659	1.6415−2.4903	1.3127	1.0249−1.6006
Glycoproteins	0.0081	0.0079−0.0082	0.0082	0.0082−0.0083
Isoleucine	0.4563	0.4159−0.4966	0.5029	0.4579−0.5479
Lactate	3.7368	3.3743−4.0992	4.1618	3.8688−4.4548
Leucine	0.7779	0.7223−0.8335	0.8474	0.7945−0.9003
Lipids	0.025	0.025−0.026	0.027	0.026−0.027
Methionine	0.0699	0.0645−0.0754	0.0863	0.0782−0.0945
Methylamine	0.0660	0.0601−0.0719	0.0823	0.0750−0.0895
Methylglutarate	0.0754	0.0660−0.0848	0.0897	0.0795−0.0999
Pyruvate	0.1203	0.1065−0.1341	0.1477	0.1282−0.1671
Proline	0.1300	0.1188−0.1412	0.1389	0.1309−0.1469
Valine	0.1241	0.0979−0.1504	0.1613	0.1370−0.1856

Table 2.2 Metabolites that are statistically different in sera of CD patients with respect to HS (*P<0.01,* Bonferroni correction applied).

	CD patients		HS	
	mean	CI 95%	mean	CI 95%
Acetoacetate	0.5734	0.5100−0.6369	0.4543	0.4027−0.5059
Choline	0.2682	0.2468−0.2895	0.2474	0.2284−0.2664
Glutamate/ Glutamine	0.3109	0.2764−0.3454	0.3235	0.2852−0.3618
Gylicine	1.2308	1.0778−1.3839	1.0321	0.8970−1.1672
Indoxyl Sulfate	0.2438	0.1990−0.2887	0.1897	0.1565−0.2228
Mannitol	1.2370	1.1230−1.3510	1.7713	1.3372−2.2053
mHHPA	0.1753	0.1455−0.2052	0.1438	0.1318−0.1558
PAG	0.4764	0.4082−0.5445	0.3544	0.3076−0.4012
Pyrimidines	0.0294	0.0207−0.0381	0.0544	0.0271−0.0837
Uracile	0.1261	0.1097−0.1426	0.0929	0.0725−0.1132

Table 2.3 Metabolites that are statistically different in urine of CD patients with respect to HS (*P<0.05*, Bonferroni correction applied).

Interestingly, it is noticed that some metabolites present in the above tables are involved in the same metabolic pathways. In particular the **Table 2.2** data suggest an alteration of the glycolysis process, the metabolism of glucose. Indeed higher levels of glucose and lower level of pyruvate, the last product of glycolysis, are present in sera of CD patients than in HS. If in literature the presence of high levels of glucose in CD patients is reported (37), decreased pyruvate is never signalled in CD patients. Moreover, if the alteration of glucose is linkable to various conditions, as for example the up-regulation of glucose intake at the level of microvillus membrane caused by an

alteration of the lipid-protein ratio of the same membrane, the decreasing of pyruvate in sera is consistent with the reduction of glycolysis. Furthermore, this hypothesis is confirmed by higher levels of 3-hydroxybutyrate in sera and acetoacetate in urine of CD patients (**Table 2.3**). These two metabolites are the product of the ketonic bodies catabolism, a supply pathway to convert energy as in case of reduction of glycolytic activity. Besides, ketonic bodies catabolism also lipid β-oxidation seems to be activated in CD patients, reading **Table 2.2** lower level of lipid (determined through their R-**CH_2**-CH_2-CO and –C=C-**CH_2**-C=C- signals) are present in sera. This is probably caused by i) activation of lipid β-oxidation to replace the lack of energy due to less efficient glycolysis ii) alteration in gut intake of lipids due to mal-absorption associated to villi atrophy. The conversion of energy through these two ways is quite less efficient than glycolysis and it should explain the high frequency of chronic fatigue in CD patient. Finally these suggestions should be confirmed by follow-up metabolites data expressed in **Table 2.4** (see below), after 12 months of gluten-free diet all the previous considered metabolites return to normal levels with the exception of acetoacetate. Other interesting altered metabolites reported in **Table 2.2** and **Table 2.3** are choline and some aminoacids (such as asparagine, leucine, methionine, proline, valine) and, above all, IS (indoxyl sulfate), mHPPA (meta-HydroxyPhenylPropionic Acid) and PAG (PhenylAcetylGlycine). Choline and aminoacids are found to be lower: this is probably a direct consequence of mal-absorption by villi atrophy. The lower levels of IS, mHPPA and PAG are due to an alteration of gut microflora. While mHPPA is one of the several products of the microbiologically mediated breakdown of poliphenols (as caffeic acid) and the conjugate chlorogenic acid (38-39), IS is a harmful uremic toxin produced by the liver starting from indole. Indole is one of the products of tryptophan metabolism by intestinal bacteria (40). PAG is only recently attributed to gut microflora (41) and its contribution to the production of these metabolites has not been fully characterized yet (42). However, these three metabolites suggest a significant alteration of the gut microflora activity, that is recently indicated as one of the probable major environmental factor involved in the pathogenesis of CD (44). Some types of statistical analysis are also applied by follow-up samples. Substantially the number of these samples is not quite sufficient and uniform to give a clear description of what is

happening during the gluten-free diet period. However, as reported in **Figure 2.2**, after 12 months of gluten-free diet (GFD) all samples but one are classified as belonging to the HS group, while after 3 and 6 months a great large number of samples remain classified as CD (**Figure 2.3**).

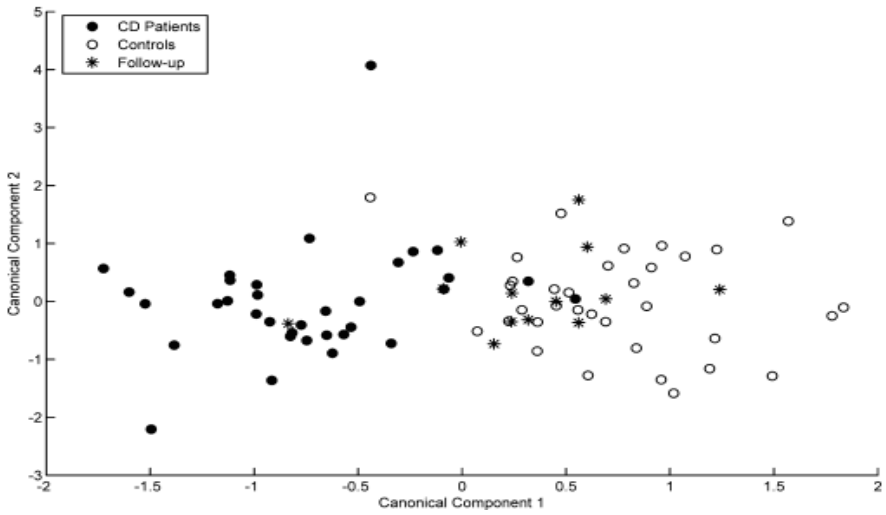


Figure 2.2 Predictive clustering of CPMG serum spectra of patients after 12 months of GFD

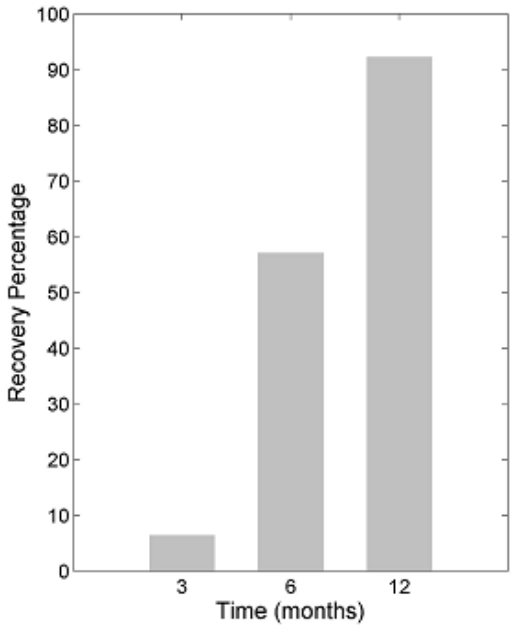


Figure 2.3 Recovery percentage of CD patients under gluten-free diet in function of the months of diet.

Furthermore, metabolites that show a statistically significant variation between untreated CD patients and corresponding follow-up after 12 months of GFD, are found using a *P* value discrimination of *0.05*. Obtained data are reported in **Table 2.4**.

metabolite	12 month follow-up
3-Hydroxybutyric acid	↓
Asparagine	↑
Choline	↑
Glucose	↓
Isoleucine	↑
Lactate	↑
Lipids	↑
Lipoproteins	↑
Leucine	↑
Methionine	↑
Valine	↑

[a] The arrows (↑ and ↓) indicate increase and decrease of levels after diet, respectively.

Table 2.4 Metabolites that significantly varies after 12 months of GFD.

They confirm the previous suggestion about metabolic processes involved in celiac disease, indicating a return to "normality" after a few months of diet. In particular it is noticed a strong correlation between glucose and 3-hydroxybutyric acid levels (with *r* value of 0.93), as to indicate the restore of the glycolysis as principal pathway of conversion of energy.

Conclusions and Perspectives

This work, published in the Journal of Proteome Research and attached here in the proper chapter, fully shows the existence of a typical metabolomic signature for celiac disease, based on three components: i) mal-absorption, ii) alteration of energy metabolism and iii) alteration of gut microflora. In this way it also suggests to explore new features and to investigate some unsolved aspects of CD, such as chronic fatigue. Nevertheless, it is necessary to extend them to distinguish CD from other causes of mal-absorption, such as small intestinal bacteria overgrowth, Crohn's disease, short bowel syndrome and so on; moreover, it is required to completely understand the capability of metabolomic to clearly highlight the presence of CD, even for peculiar forms of the pathology, as silent form and latent form. For the last purpose another study has already started with more than 100 samples of urine and serum analyzed. The study plainly provides a blind analysis of a significant number of samples coming

from subjects with various characteristics (HS, classic CD subjects, CD subjects on diet, latent CD subjects, silent CD subjects, subjects that suffers of other intestinal pathologies). These samples will be tested using as classification trial the test sets developed during the first study.

Bibliography

(1) Holtmeier W., Caspary W.; *Celiac disease* Orphanet J Rare Dis **2006** 1 1:3.

(2) Kupper C.; *Dietary guidelines and implementation for celiac disease* Gastroenterology **2005** 128 S121:S127.

(3) Green P. H., Jabri B.; *Coeliac disease* Lancet **2003** 362 383:91.

(4) Schuppan D.; *Current concepts of celiac disease pathogenesis* Gastroenterology **2000** 119 234:42.

(5) van Heel D. A., Hunt K., Greco L., Wijmenga C.; *Genetics in coeliac disease* Best Pract Res Clin Gastroenterol **2005** 19 323:39.

(6) Bevan S., Popat S., Braegger C.P., Busch A., O'Donoghue O., et al*.; Contribution of the MHC region to the familial risk of coeliac disease* J Med Genet **1999** 36 687:90.

(7) Greco L., Romino R., Coto I., Di Cosmo N., Percopo S.,et al.; *The first large population based twin study of coeliac disease* Gut **2002** 50 624:8.

(8) Barker J. M., Liu E.; *Celiac disease: pathophysiology, clinical manifestations, and associated autoimmune* conditions Adv Pediatr **2008** 55 349:65.

(9) Fasano A., Berti I, Gerarduzzi T, Not T., Colletti R. B., et al.; *Prevalence of celiac disease in at-risk and not-at-risk groups in the United-States: a large multicenter study* Arch. Intern. Med. **2003** 163 286:92.

(10) Catassi C., Ratsch I. M., Gandolfi L.; Pratesi R.; Fabiani E., et al.; *Why is coeliac disease endemic in the people of the Sahara?* Lancet **1999** 354 647:8.

(11) Stevens F. M., Egan-Mitchell B., Cryan E., McCarthy C. F., McNicholl B.; *Decreasing incidence of coeliac disease* Arch Dis Child **1987** 62 465:8.

(12) Maki M., Kallonen K., Lahdeaho M. L., Visakorpi J. K.; *Changing pattern of childhood coeliac disease in Finland* Acta Paediatr Scand **1988** 77 408:12.

(13) George E. K., Mearin M. L., Franken H. C., Houwen R. H., Hirasing R. A., et al.; *Twenty years of childhood coeliac disease in The Netherlands: a rapidly increasing incidence?* Gut **1997** 40 61:6.

(14) Catassi C., Fabiani E., Corrao G., Barbato M., De Renzo A., et al.; *Risk of non-Hodgkin lymphoma in celiac disease* JAMA, J. Am. Med. Assoc. **2002** 287 1413:9.

(15) Mearin M. L., Catassi C., Brousse N., Brand R., Collin P., et al.; *European multi-centre study on coeliac disease and non-Hodgkin lymphoma* Eur. J. Gastroenterol. Hepatol. **2006** 18 187:94.

(16) Vasquez H., Mazure R., Gonzalez D., Flores D., Pedreira S., et al.; *Risk of fractures in celiac disease patients: a cross-sectional, case-control study* Am. J. Gastroenterol. **2000** 95 183:9.

(17) Dickey W.; *Epilepsy, cerebral calcifications, and coeliac disease* Lancet **1994** 344 1585:6.

(18) Luostarinen L., Himanen S. L., Luostarinen M., Collin P., Pirttila T.; *Neuromuscular and sensory disturbances in patients with well treated coeliac disease* J. Neurol. Neurosurg. Psychiatry **2003** 74 490:4.

(19) Kaukinen K., Collin P., Mykkanen A. H., Partanen J., Maki M., et al.; *Celiac disease and autoimmune endocrinologic disorders* J. Dig. Dis. Sci. **1999** 44 1428:33.

(20) Cronin C. C., Feighery A., Ferriss J. B., Liddy C., Shanahan F., et al.; *High prevalence of celiac disease among patients with insulin-dependent (type I) diabetes mellitus* Am. J. Gastroenterol. **1997** 92 2210:2.

(21) Ch'ng C. L., Jones M. K., Kingham J. G. *Celiac disease and autoimmune thyroid disease* Clin. Med. Res. **2007** 5 184:92.

(22) Myhre A. G., Aarsetoy H., Undlien D. E., Hovdenak N., Aksnes L., et al.; *High frequency of coeliac disease among patients with autoimmune adrenocortical failure* Scand. J. Gastroenterol. **2003** 38 511:5.

(23) Hopper A., Cross S., Hurlstone D., McAlindon M., Lobo A., et al.; *Pre-endoscopy serological testing for coeliac disease: evaluation of a clinical decision tool* BMJ **2007** 334(7596) 729.

(24) Chorzelski T. P., Beutner E. H., Sulej J., Tchorzewska H., Jablonska S., et al.; *IgA anti-endomysium antibody. A new immunological marker of dermatitis herpetiformis and coeliac disease* Br. J. Dermatol. **1984** 111 395:402.

(25) Korponay-Szabó I. R., Dahlbom I., Laurila K., Koskinen S., Woolley N., et al.; *Elevation of IgG antibodies against tissue transglutaminase as a diagnostic tool for coeliac disease in selective IgA deficiency* Gut **2003** 52(11) 1567:71.

(26) James M. W., Scott B. B.; *Endomysial antibody in the diagnosis and management of coeliac disease* Postgrad. Med. J. **2000** 76 466:8.

(27) Feighery C.; *Fornightly review: coeliac disease* BMJ **1999** 319 236:9.

(28) van Heel D. A., West J.; *Recent advance in coeliac disease* Gut **2006** 55 1037:46.

(29) Matysiak-Budnik T., Candalh C., Cellier C., Dugave C., Namane A, et al.; *Limited efficiency of prolyl endpeptidases in the detoxification of gliadin petides in celiac disease* Gut **2005** 129 786:96.

(30) Pare P., Douville P., Caron D., Lagace R.; *Adult celiac sprue: changes in the pattern of clinical recognition* J. Clin. Gastroenterol. **1988** 10 395:400.

(31) Ciacci C., Peluso G., Iannoni E., Siniscalchi M., Iovino P., et al.; *L-Carnitine in the treatment of fatigue in adult celiac disease patients: a pilot study* Aliment. Pharmacol. Ther. **2005** 22 489:94.

(32) Empson M.; *Celiac disease or chronic fatigue syndrome--can the current CDC working case definition discriminate?* Am. J. Med. **1998** 105 79:80.

(33) Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H., et al.; *Evidence of different metabolic phenotypes in humans* Proc. Natl. Acad. Sci. U.S.A. **2008** 105 1420:4.

(34) Vuoristo M, Tarpila S., Miettinen T. A.; *Serum lipids and fecal steroids in patients with celiac disease: effects of gluten-free diet and cholestyramine* Gastroenterology **1980** 78 1518:25.

(35) Ciacci C., Cirillo M., Giorgetti G., Alfinito F., Franchi A., et al.; *Low plasma cholesterol: a correlate of nondiagnosed celiac disease in adults with hypochromic anemia* Am. J. Gastroenterol. **1999** 94 1888:91.

(36) Capristo E., Addolorato G., Mingrone G., Scarfone A., Greco A. V., et al.; *Low-serum high-density lipoprotein-cholesterol concentration as a sign of celiac disease* Am. J. Gastroenterol. **2000** 95 3331:2.

(37) West J., Logan R. F., Hill P. G., Khaw K. T.; *The iceberg of celiac disease: what is below the waterline?* Clin. Gastroenterol. Hepatol. **2007** 39 922:8.

(38) Phipps A. N., Stewart J., Wright B., Wilson I. D.*; Effect of diet on the urinary excretion of hippuric acid and other dietary-derived aromatics in rat. A complex*

interaction between diet, gut microflora and substrate specificity Xenobiotica **1998** 28 527:37.

(39) Williams R. E., Eyton-Jones H. W., Farnworth M. J., Gallagher R., Provan W. M.; *Effect of intestinal microflora on the urinary metabolic profile of rats: a (1)H-nuclear magnetic resonance spectroscopy study* Xenobiotica **2002** 32 783:94.

(40) Gao X. X., Ge H., Zheng W. F., Tan R. X.; *NMR-based metabonomic for detection of Helicobacter pylori infection in gerbils: which is more descriptive* Helicobacter **2008** 13 103:11.

(41) Nicholson J. K., Holmes E., Wilson I. D.; *Gut microorganisms, mammalian metabolism and personalized health care* Nat. Rev. Microbiol. **2005** 3 431:38.

(42) Forsberg G., Fahlgren A., Horstedt P., Hammarstrom S., Hernell O., et al.; *Presence of bacteria and innate immunity of intestinal epithelium in childhood celiac disease* Am. J. Gastroenterol. **2004** 99 894:904.

4 BREAST CANCER AND METABOLOMICS

Breast Cancer

Introduction and classification

Breast cancer is a type of cancer that usually starts in the inner lining of the milk ducts or lobules. It is the second most common cancer after lung cancer worldwide, with an incidence of 10.4 % (calculated on both genders), and the fifth most common cause of cancer death (519.000 ca. deaths in the 2004, equal to 7% of all cancer deaths) (1). Breast cancer is about 100 times more frequent in women than in men, even if the rates of survival are practically equivalent. Currently four different types of classification exist for breast cancer. All these classifications are done on the basis of different criteria and serve different purposes. These different schemes consider i) the type of pathology, ii) the grade of the tumor, iii) genetic and proteic expression and iv) the stage of tumor.

Type of pathology : It is a classification based on histological appearance and on some other pathological criteria. From this point of view, the most common types of breast cancer are ductal carcinoma (malignant cancer in breast ducts) and invasive lobular carcinoma (malignant cancer in breast lobules).

Grade of tumor : It is determined by the pathologist using a microscopy and the Bloom-Richardson-Elston staging system (2-3). With the microscopy it is assigned a score ranging from 1 to 3 to the three followed features: i) percentage of tumor with normal ducts, ii) number of observable mitotic figures and iii) characteristic of cell nuclei. Therefore, the final score will be included between 3 (well differentiated, best prognosis) and 9 (poorly differentiated, worst prognosis). In details, tumor with scores between 3 and 5 are grade 1, between 6 and 7 are grade 2 and between 8 and 9 are grade 3.

Genetic and Proteic Expression : This test is usually done by immunohistochemistry. The breast cancer cells are tested for expression of some genes, as estrogen receptor (ER) and progesterone receptor (PR), and of some proteins, as human epidermial growth factor receptor 2 (HER2).

Stage of tumor : It is used the so-called classification of malignant tumors (TNM). TNM classification gives a code of a tumor constituted by a letter, that can be “T”, “N” or “M”, and respectively means **T**umor, lymph **N**ode and **M**etastases. The letter is followed by a numeric or alphanumeric code to substantially indicate some parameters of tumor as invasiveness, dimension, presence of tumor cells outside the breast, number, size and location of breast cells deposits in lymph node.

Finally, these are not all possible classifications, for instance it is possible to do a classification based on the presence of inflammatory states (4); moreover, some of these parameters can be significantly modified over time.

Etiology and epidemiology

Although the first work on the epidemiology and etiology of breast cancer was published in 1926 by Janet E. Lane-Claypon for British Ministry of Health, at present it is not possible to establish correctly the epidemiological risk factor and etiology for every type of breast cancer. Indeed epidemiological research allows us to identify factor risks for a population, as incidence and prevalence, but does not give information about the single individual. At the same time about 5% of new breast cancer are attributable to hereditary factors, while the etiology of the other 95% is unknown (5). Breast cancer, like other forms of cancer, is considered to result from multiple environmental and hereditary factors. Moreover, a series of primary risk factors are identified as gender (6), age, hormones (7), a high-fat diet (8), alcohol intake (9), obesity, but only a small increase in breast cancer frequency is attributed in this study to these factors, and these studies are often not well randomized. Furthermore, the expression of two genes, called BRCA1 and BRCA2, is associated with an increase of about 30-40% of breast and ovarian cancer risk (10). Finally, personal and familial history of breast cancer significantly increases the risk of this pathology, while, for instance, some races, as Latina, Asian or Afroamerican are less subject to this pathology than Caucasian.

Clinical manifestations and signs of breast cancer

It is considered that, in about 80% of cases (1), breast cancer is discovered by the patients themselves when they find a lump that feels different from the surrounding breast tissue. Sometimes the lumps, especially if they are very small are discovered through a mammography. Moreover, the presence of a lump in the armpits (lymph nodes) is also a possible symptom of breast cancer. Obviously many other signs may be included, as changes in size or shape, skin dimpling, nipple inversion. Pain (called mastodynia) is another possible symptom but it is not characteristic. Indeed, it is possible to have it in many other breast pathologies such as mastitis and fibro-adenoma. When breast cancer cells invade the lymphatic dermals, that are small vessels in the skin of breast, its presentation can resemble skin inflammation and thus is known as inflammatory breast cancer. Symptoms of inflammatory breast cancer include pain, swelling, warmth and redness throughout the breast. Another reported symptom complex of breast cancer is Paget's disease. This is a syndrome that presents eczematoid skin changes such as redness and mild flaking of the nipple skin. As Paget's advances, symptoms may include itching, increased sensitivity, burning, and pain. There may also be discharge from the nipple. Approximately half of women diagnosed with Paget's also have a lump in the breast (11). Furthermore, when breast cancer is manifest in the metastatic form a great number of new symptoms also appear, which depend on the location of metastasis. Common sites of metastasis include bone, liver, lung and brain (12). Unexplained weight loss can occasionally herald an occult breast cancer, often associated to very frequent symptoms as fever and chills. Bone or joint pains can sometimes be manifestations of metastatic breast cancer, as some neurological symptoms. These symptoms are "non-specific", meaning they can also be manifestations of many other illnesses. However, the presence of one or more of these symptoms have to be seriously considered by the person, because of the possibility of an underlying breast cancer at almost any age.

Diagnosis and Therapeutics

The most common used screening method for diagnosis of breast cancer is a combination of X-ray mammography and clinical breast exam. In women at higher risk

than normal, such as those with a strong family history of cancer, additional tools may include genetic testing or breast Magnetic Resonance Imaging (MRI). Breast self-examination was a form of screening that was heavily advocated in the past, but several large studies have shown that it does not have a survival benefit for women and often causes considerably anxiety. This is due to the fact that breast cancer can be detected at a relatively advanced stage operating in this way, whereas other methods push to identify the cancer at an earlier stage where curative treatment is more possible. X-ray mammography uses x-rays to examine the breast for any uncharacteristic masses or lumps. Regular mammograms are recommended in several countries in women over a certain age as a screening tool. Genetic testing for breast cancer typically involves testing for mutations in the BRCA genes. This is not generally a recommended technique except for those at elevated risk of breast cancer. While previously discussed screening techniques are useful in determining the possibility of cancer, a further testing is necessary to confirm whether a lump detected on screening is cancer or other, like a simple cyst. The common diagnosis is elaborated after a "triple test" of clinical breast examination led by a trained specialist. This triple test comprises mammography, fine needle aspiration and cytology (FNAC). Both mammography and clinical breast exams, also used for screening, can indicate an approximate likelihood that a lump is cancer, and may also identify any other lesions, while FNAC extracts a small portion of fluid from the lump. If the fluid is clear the cancer is highly probably absent, while if there is presence of blood in the fluid, it is necessary a microscope inspection to check the presence of cancer cells. These three tools can be used to diagnose breast cancer with a good degree of accuracy. Another useful option is biopsy, which consists in the removal of either part or entire lump. As for other tumors, the mainstay of breast cancer treatment is surgery when it is localized, with possible adjuvant hormonal therapy (with tamoxifen or an aromatase inhibitor), chemotherapy, and/or radiotherapy. Depending on clinical criteria (age, type of cancer, size, metastasis) patients are roughly divided into high risk and low risk cases, with each risk category following different rules for therapy.

Metabolomic study of Breast Cancer

Aim of the work

The natural aim of every work which investigates breast or other types of cancer, is the improvement of the knowledge of both pathologies and exploitable clinical therapies. Metabonomics, being a science that provides a dynamic portrait of metabolic status, can be very useful in these directions. Particularly the improvement of the prediction of clinical outcomes is necessary to have a better approach to breast cancer, individualize the therapy and reduce the high side effects associated to it. For these purposes, the attention is focused on the possibility to discriminate i) pre-operative not-metastatic breast cancer subjects ii) post-operative not-metastatic breast cancer subjects and iii) post-operative metastatic breast cancer subjects, by relying on the metabonomic analysis of the serum. Moreover, an important goal is to check the capability of metabonomics to single out the presence of micrometastasis in an organism. Indeed, in the treatment of early breast cancer, a critical issue is the identification of which individuals can benefit from adjuvant intervention. As mentioned above, the patients are divided into two groups (high and low risk) and their therapy is decided starting from this classification. However, the risk is often under- or over-estimate and, as a consequence, the applied therapy is not completely corrected and appropriated. One, and probably the most, important cause of this mis-assignment is the existence of unpredicted micrometastasis in the host. The study of micrometastasis is a step not easily solvable. In this work it is hypothesized that a metabonomic fingerprint of the micrometastasis exists and to check this metabonomic risk, data are compared with 10-year mortality rates data obtained from the use of Adjuvant!online (13) software.

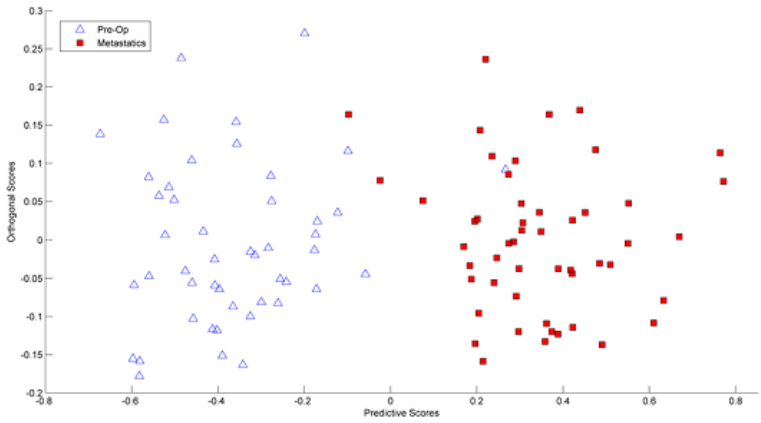
Collected serum samples are divided into three classes: i) 44 pre-operative early breast cancer ii) 98 post-operative early breast cancer (44 as the same of pre-operative, 45 new recruited) iii) 51 metastatic breast cancer. For early breast cancer patients the mean time between pre-operative collection and surgery is 16 days (range 2:40), while for post-operative is 33 days (range 16-55).

Results and Discussion

A clusterization between early pre-operative and metastatic subjects was obtained for both cpmg and serum spectra using O-PLS (**Figure 3.1 A** e **3.1 B**). Looking at the figure it is noticeable the clearly separation between the two groups revealing the existence of a metabonomic fingerprint for metastatic pathology with respect to not-metastatic. After this process, a double cross validation scheme is applied to single out the correct value of prediction of defined statistical approach. A great number of individuals are correctly classified, whilst some cases of misclassification exist. Right percentage of recognition for CPMG are 75% of sensitivity, 69% of specificity and a global predictive accuracy of 72%. Similar values are obtained using NOESY1D, sensitivity of 77%, specificity of 68% and predictive accuracy of 73%. Through the application of Wilcoxon test with Bonferroni correction, some metabolites that are significantly discriminating between the two classes (p value < 0.05) are identified. Indeed metastatic subjects are characterized by higher value of glucose and of some aminoacids as phenylalanine, proline, lysine and N-acetyl cysteine and by lower value of many lipidic signals that are present in unfiltered noesy spectra.

Clusterization showed in **Figure 3.1** suggests that some pre-operative samples present very similar characteristic to metastatic samples. In order to better understand the cause of these misclassification, a factor of risk, called “Metabolomic risk” is assigned to every not-metastatic sample. The value of each Metabolomic risk is assigned measuring the Euclidean distance of each dot from the centre of metastatic cluster. Obviously, it is supposed that samples that have a higher distance from metastatic center have less metastatic characteristic and, therefore, they have a lower value of risk. Practically, the shorter is the distance the higher is the risk, and vice versa.

1A



1B

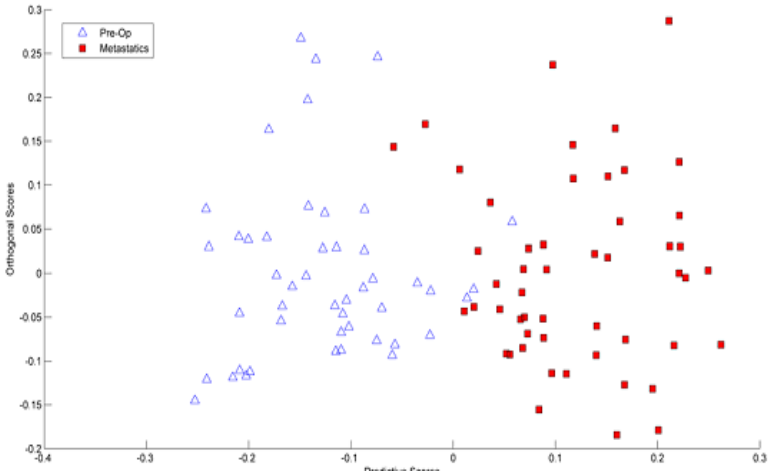


Figure 3.1 O-PLS of pre-operative (N=44) and metastatic (N=51) patients showing near complete separation of patient groups; CPMG (A) and NOESY1D (B) techniques.

The next step is to compare the Metabolomic risk with the established risk by using Adjuvant!online software. Adjuvant!online is a free software that is very useful to predict the risk of relapses and mortality associated to a breast tumor in not-operated subjects, if they are not treated; moreover, the software also values the reduction of the risk in case of specific therapies. Parameters used by software are related to form and type of tumor, as size, hormonal receptor (ER) status, lymph node involvement

(14), but also to some patient characteristics, as age, diet, used drugs In **Figure 3.2** the comparison between metabolomic risk and 10-year mortality by Adjuvant!online is reported.

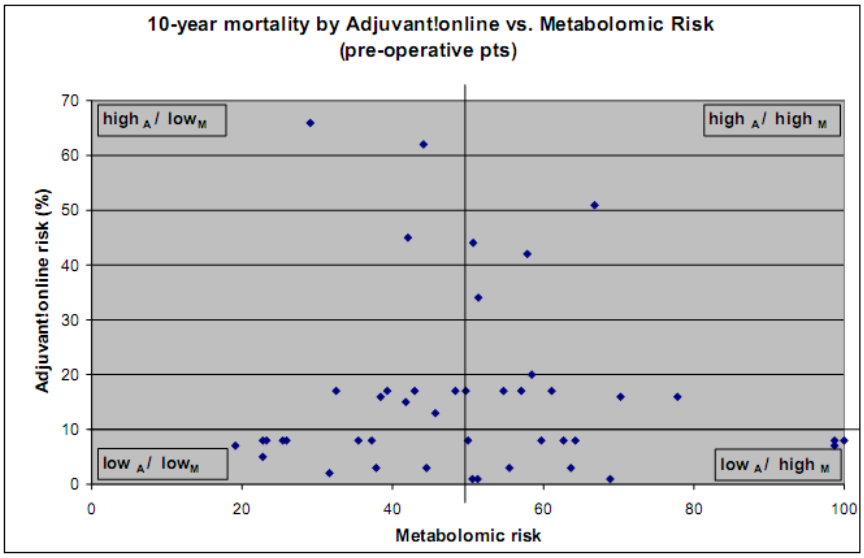


Figure 3.2 Correlation between 10-year breast cancer mortality estimated by Adjuvant!online (A) and Metabolomic risk (M) for pre-operative patients (pts).

A clear relationship between the two established risks does not seem to be present at a first exam of the previous figure. Particularly, it is noticeable that there is a significant discordance for three samples (present in the right extreme limit) that are classified as having a very high metabolomic risk whilst they do not seem to be a risk for adjuvant. The percentage value of concordance is about 48% for both metabolomic high risk and metabolomic low risk. To explain these data, the best possible assumption is related to the presence of micrometastasis in some breast cancer hosts. Indeed micrometastasis are not considered in calculation of risk by Adjuvant!. Therefore, pre-operative patients that have a metabonomic fingerprint similar to metastatic subjects and, as a consequence, are the nearest to metastatic cluster, could have residual micrometastasis in their own bodies. To confirm these hypotheses, it is decided to use post-operative serum samples coming from same subjects as validation test set, in order to minimize inter-individual variation. The same methods of

clusterization and classification for pre-operative specimen are also used in this case leading to following results (see **Figure 3.3**).

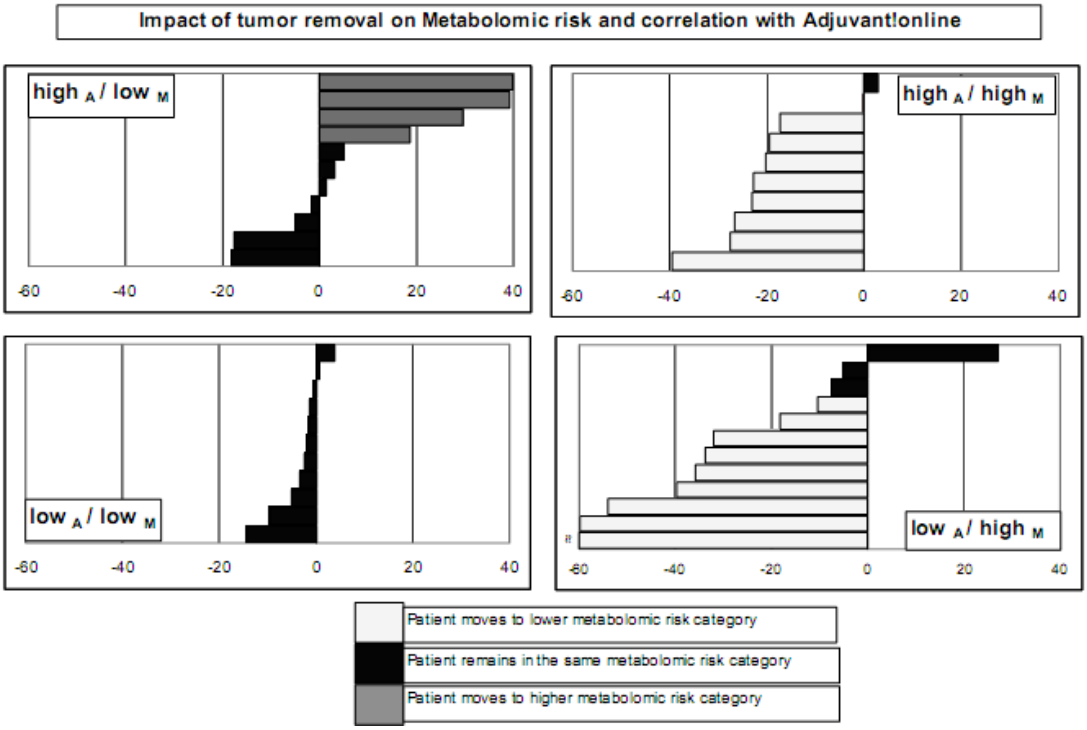


Figure 3.3 Impact of surgical removal of primary tumor on metabolomic risk

In the above figures, each bar represents the change of the metabonomic risk in pre- and post-operative samples; the longer the bar, the higher the variation of risk. In details i) 100% (11 of 11) of previously classified as low_a/low_m subjects remain in the same class of risk, ii) 92% (9 of 12) $low_a/high_m$ show a risk reduction, iii) 80% (8 of 10) $high_a/high_m$ have a reduction of metabonomic risk after surgery, iv) 36% ca (4 of 11) of $high_a/low_m$ show interestingly a sizeable increase of metabonomic risk, whilst the others remain in the same class of risk. This last trend is the most interesting to analyze, because it seems to apparently be contrasting logical supposition. Indeed if a not-metastatic tumor is removed from the hosts, the risk of recidive and/or death should decrease. Actually, it is still suggested in literature that the surgical removal of a primary tumor could increase the risk of a future disease (15-17). The stress that is associated to a surgery, and that can cause both immunosuppression and proliferation of growth factors, can significantly trigger the possibility of tumor growth (18-19).

Conclusions and Perspectives

In this work it is showed that a metabonomic fingerprint seems to exist for metastatic patients and it is observable analyzing serum samples. The differences between metastatic and early breast cancer patients are noticeable in both pre-operative and post-operative samples (i.e. coming from the same subject). Actually the collected biostatistical data are not very high in term of accuracy (about 70%). This is also probably due to great inter-variability. To overcome this problem it is necessary to collect more than one sample from the same subject in a short period of time. However, a definition of metabonomic risk, as percentage probability of 10-years death risk after surgery, is defined and compared with Adjuvant!online data. Metabonomic risks are partially in agreement with Adjuvant!online risk, but there are also significant differences that are attributed to the possible existence of micrometastatis (not identifiable by Adjuvant!online). The only and better way to confirm these suggestions is to repeat the study in a class of subjects which is also donating follow- up samples, in order to highlight the effects of a possible progression of the pathology. Another important approach can be the analysis of tissues of breast cancer. The analyses of tissue is still used in metabonomics and it has led to the significant results of suggesting sarcosine (20) as a possible biomarker of prostate cancer progression. Similarly, an identical approach is useful in order to completely understand the metabolism of breast cancer and also to study in details both the role of phenylalanine in tumor progression and the importance of its pathways in the onset of breast cancer and, moreover, in progression to metastasis.

Bibliography

(1) World Health Organization (web site www.who.int) *Fact sheets N. 297, Cancer* **2006.**

(2) Bloom H. J., Richardson W. W.; *Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years* Br. J. Cancer **1957** 11 359:77.

(3) Genestie C., Zafrani B., Asselain B., Fourquet A., Rozan S., et al.; *Comparison of the prognostic value of Scarff-Bloom-Richardson and Nottingham histological grades in a series of 825 cases of breast cancer: major importance of the mitotic count as a component of both grading systems* Anticancer Res. **1998** 18 571:6.

(4) Giordano S. H., Hortobagyi G. N.; *Inflammatory breast cancer: clinical progress and the main problems that must be addressed* Breast Cancer Res. **2003** 5284:8.

(5) Madigan M. P., Ziegler R. G., Benichou J., Byrne C., Hoover R. N.; *Proportion of breast cancer cases in the United States explained by well-established risk factors* J. Natl. Cancer Inst. **1995** 87 1681:5.

(6) Giordano S. H., Cohen D. S., Buzdar A. U., Perkins G., Hortobagyi G. N.; *Breast carcinoma in men: a population-based study* Cancer **2004** 10151:7.

(7) Yager J. D., Davidson N. E.; *Estrogen carcinogenesis in breast cancer* N. Engl. J. Med. **2006** 354 270:82.

(8) Chlebowski R. T., Blackburn G. L., Thomson C. A., Nixon D. W., Shapiro A., et al.; *Dietary fat reduction and breast cancer outcome: interim efficacy results from the Women's Intervention Nutrition Study* J. Natl. Cancer Inst. **2006** 98 1767:76.

(9) Boffetta P., Hashibe M., La Vecchia C., Zatonski W., Rehm J.; *The burden of cancer attributable to alcohol drinking* Int. J. Cancer. **2006** 119 884:7.

(10) Venkitaraman A. R.; *Cancer susceptibility and the functions of BRCA1 and BRCA2* Cell **2002** 108 171:82.

(11) Marcus E.; *The management of Paget's disease of the breast* Curr. Treat. Options Oncol. **2004** 5 153:60.

(12) Lacroix M.; *Significance, detection and markers of disseminated breast cancer cells* Endocr. Relat. Cancer **2006** 13 1033:67.

(13) Goldstein L., Gray R., Badve S., Childs B. H., Yoshizawa C., et al.; *Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features* J. Clin. Oncol. **2008** 26 4063:71.

(14) Ravdin P. M., Siminoff L. A., Davis G. J., Mercer M. B., Hewlett J., et al.; *Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer* J. Clin. Oncol. **2001** 19 980:91.

(15) Baker D. G., Masterson T. M., Pace R., Constable W. C., Wanebo H. J.; *The influence of the surgical wound on local tumor recurrence* Surgery **1989** 106 525:32.

(16) Hofer S. O., Shrayer D., Reichner J. S., Hoekstra H. J., Wanebo H. J.; *Wound-induced tumor progression: a probable role in recurrence after tumor resection* Arch. Surg. **1998** 133 383:9.

(17) Taglibue E., Agresti R., Carcangiu M. L., Ghirelli C., Morelli D., et al.; *Role of HER2 in wound-induced breast carcinoma proliferation* Lancet **2003** 362 527:33.

(18) Cole W. H.; *The increase in immunosuppression and its role in the development of malignant lesions* J. Surg. Oncol. **1985** 30 139:44.

(19) Fisher B., Gunduz N., Coyle J., Rudock C., Saffer E.; *Presence of a growth-stimulating factor in serum following primary tumor removal in mice* Cancer Res. **1989** 49 1996:2001.

(20) Sreekumar A., Poisson L. M., Rajendiran T. M., Khan A. p., Cao Q., et al. ; *Metabolomics profile delineate potential role for sarcosine in prostate cancer progression* Nature **2009** 457 910:4.

5 METABOLIC PHENOTYPES (METABOTYPES) IN HUMAN URINE

Background

Since their birth, metabolomics has showed a great potential in various research areas. In particular the relevance of metabolomics can be greatly enhanced by the determination of the existence of a metabolic phenotype, typical for each subject. Differences in experimental metabolic profiles due to genetic strain differences in animal models have been observed, leading to the suggestion that each individual or group of individuals may be characterized by a different metabotype, defined as "the multiparametric description of an organism in a given physiological state based on metabolomic data" (1). The availability of a characteristic metabotype of an individual could be fundamental in many fields as nutrigenomics (2,3), in evaluation of drug efficacy, in pharmacometabolomics (4), and in studies of personalized nutrition aimed at maintaining metabolic health and avoiding loss of homeostasis or correcting homeostasis dysregulations. Of course, a fundamental condition of metabotypes is their stability over time. A major problem is that the experimental metabolic profile is influenced not only by the genotype but also by age, lifestyle, environmental factors, nutritional status, assumption of drugs, and by other metabolites from symbiotic organisms, as gut microflora (5-7). Consequently, changes in the metabolic profile of biologically complex organisms (as humans) in response to pathological stimuli may be difficult to distinguish from normal physiological variations. Despite these factors, the experimental evidence of the existence of a metabolic phenotype is recently collected (8), assessing the influence of possible perturbing factors on the metabolic profiles and minimizing them in order to eliminate noise due to random daily variation. Indeed the approach used for the determination of the metabotypes counts on the NMR analysis of multiple urine samples (40 for each individual) taken in quite consecutive days (about 2-3 months) from 22 healthy subjects. Each of these subjects avoids alcohol and drug intake the day before the collection, and, moreover, fills a complete dietary sheet in order to better determine changes due to food behaviour. In this way, it is obtained for the first time a natural, stable, and invariant metabolic profile that is typical for a given subject, even if not necessarily unique. As hypothesized, the identification of a characteristic individual fingerprint is very useful because it may allow the researchers to i) better plan personalized therapy and nutrition, ii) perform

studies of pharmacometabolomics to better predict and assess drug efficacy and toxicity, iii) follow phenotype changes as a function of disease progression, possibly leading to earlier diagnosis and prognosis, iv) perform cost-effective screenings on large human populations and v) address how possible long-term changes may be related to aging. The fingerprint is assessed to be stable only in a short period.

Aim of the work

The discovery of the existence of a stable individual human phenotype in a short period of time (about 2-3 months) is clearly an important improvement for metabolomics (8). At the same time it is necessary to unravel the behaviour of the metabotypes in a more extensive time period, in order to completely help researchers in all above mentioned fields and, especially, for medical application. To solve this goal 11 subjects of the project called MetRef1 (8) are recruited again after 2 years, entering in the MetRef2 project, and 4 of them are recruited one more time 1 year later, entering in the MetRef3 project. Nine healthy individuals, not present in MetRef1, are recruited after 2 years of MetRef1 together with the other 11 subjects. These new recruited subjects result to be useful to further extend the analyzed metabotypes and try to define the possible saturation of the metabolic space, in order to highlight the eventual presence of shared metabotypes in two or more subjects. In this way the project MetRef2 is constituted by 20 individuals (9 males, 11 females), aging 25-55, donating 40 urine samples (first in the morning, pre-prandial) collected over a period of about 3 months. While MetRef3 is constituted only by 40 samples donated by 4 subjects recruited for the third time (see **Figure 4.1**).

Moreover, there are some singularities in the new collections that, although not statistically relevant, can contribute to better understand environmental and genetic contributions to the definition of the individual metabolic phenotypes. As it is noticeable in **Figure 4.1**, two new recruited subjects are homozygous twins (indicated with TA and TB codes), while other two subjects (BU and BV) are father and son. Moreover, some subjects that are donating in more than one project quite drastically change their lifestyle, as BC, who, during the two collections, moved from Italy (MetRef1) to Spain (MetRef2), varying the diet, and AR, who quitted smoking.

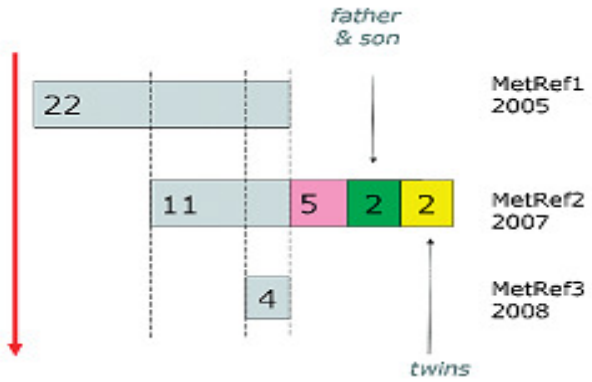


Figure 4.1 Collection scheme of the three MetRef projects..

Results and Discussion

To correctly define the metabolic space, it is necessary to start from the data published about the first MetRef work (8). These results are represented in **Figure 4.2**.

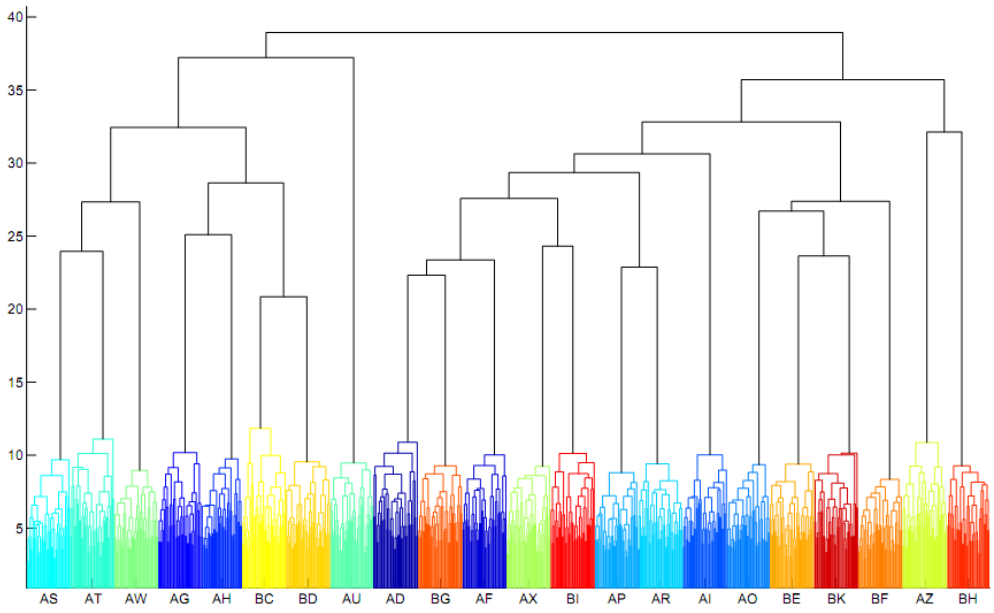


Figure 4.2 Dendrogram relative to cluster analysis on the 21-dimensional PCA/CA subspace for the 22 subjects of the MetRef1 collection.

In the dendrogram, each vertical bar is a single spectrum while each horizontal bar represents the inter-sampling distances for all spectra in a statistical space defined through a PCA/CA analysis. Moreover, the colours obviously highlight the spectra belonging to an individual (as reported in x axis), and therefore the black coloured bars

result to be a representation of the interindividual distances; practically every fork gives a measure of the metabolic affinity between the individuals. These representation is very useful in this case since it is totally impossible to perform the 21 dimensional space of the system. However, it is clearly visible that all subjects present a typical metabotype, there are not overlaps among all coloured lines by each individual. Furthermore, these trends are confirmed through a predictive analysis using the single vote classification: the mean value is 99.7% for MetRef1 collection (see **Table 4.1**). The same identical analysis is done on the 20 individuals of the MetRef2 projects giving very similar results both as dendrogram visualization (**Figure 4.3**) and as percentage of recognition that results to have a mean of about 99.6% (**Table 4.1**).

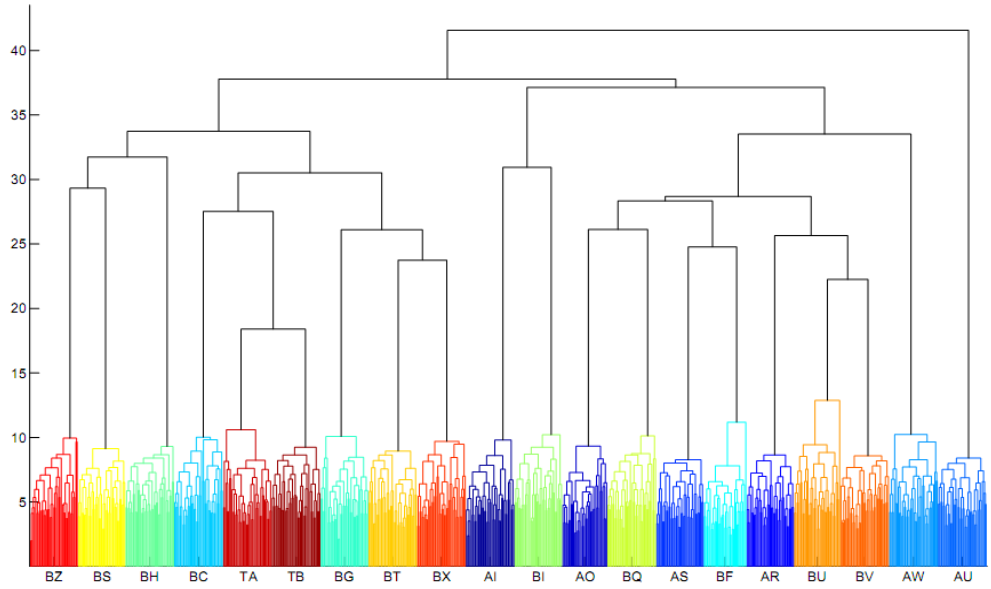


Figure 4.3 Dendrograms relative to cluster analysis on the 19-dimensional PCA/CA subspace for the 20 subjects of MetRef2 collection.

These results confirm the robustness of the method still used in the first work. Moreover, analogous results, as dendrograms (**Figure 4.4**) and recognition values (**Table 4.1**), are obtained pooling together all 31 individuals of the MetRef1 and MetRef2 collection (clearly for the subjects that are participating to both collections the analyzed spectra are about 80 for each one). All these data confirm the existence of various individual metabolic phenotypes and suggest that the number of these are higher than 20 and not yet defined. Thus the metabolic space is probably distant to be satured and this opens significant perspective in fields as biology and medicine.

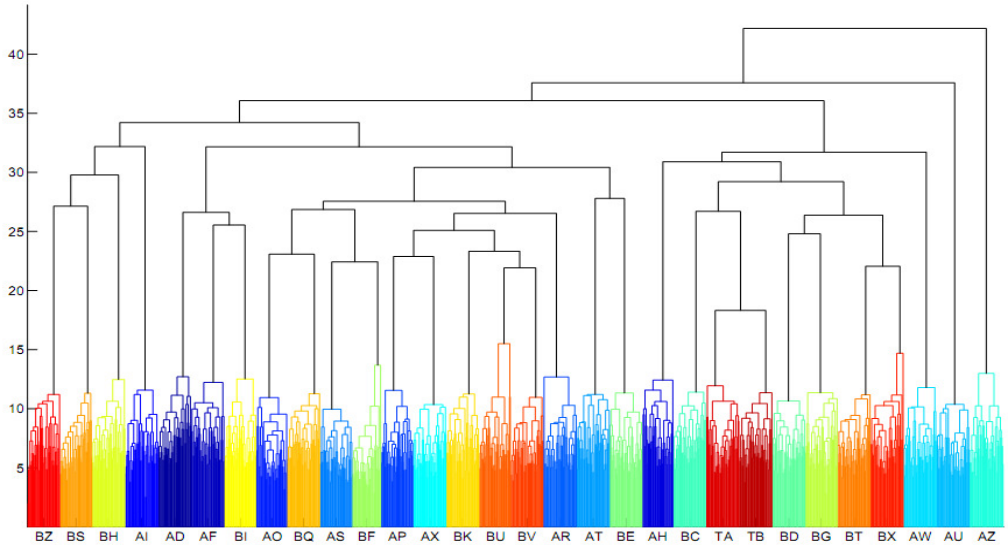


Figure 4.4 Dendrogram relative to cluster analysis on the 30-dimensional PCA/CA subspace for the 31 different subjects of both MetRef 1 and MetRef2 collections.

(a) 2005		(b) 2007		(c) 31-donors		(d) 46-pseudodonors	
AD	99.905%	AI	100.000%	AD	99.895%	AD	99.990%
AF	100.000%	AO	100.000%	AF	100.000%	AF	100.000%
AG	100.000%	AR	99.995%	AG	99.997%	AG	100.000%
AH	99.922%	AS	99.941%	AH	99.919%	AH	99.989%
AI	99.760%	AU	100.000%	AI	100.000%	AI	99.775%
AO	97.500%	AW	99.998%	AO	100.000%	AO	97.500%
AP	97.500%	BC	99.705%	AP	97.500%	AP	97.500%
AR	100.000%	BF	99.629%	AR	100.000%	AR	100.000%
AS	100.000%	BG	100.000%	AS	99.992%	AS	99.792%
AT	99.995%	BH	100.000%	AT	100.000%	AT	100.000%
AU	100.000%	BI	98.462%	AU	100.000%	AU	99.379%
AW	100.000%	BQ	99.998%	AW	100.000%	AW	100.000%
AX	100.000%	BS	100.000%	AX	100.000%	AX	100.000%
AZ	100.000%	BT	99.983%	AZ	100.000%	AZ	100.000%
BC	99.998%	BU	96.647%	BC	99.965%	BC	98.630%
BD	100.000%	BV	99.998%	BD	100.000%	BD	100.000%
BE	100.000%	BX	99.998%	BE	100.000%	BE	100.000%
BF	100.000%	BZ	100.000%	BF	99.497%	BF	100.000%
BG	99.933%	TA	98.450%	BG	100.000%	BG	98.838%
BH	100.000%	TB	98.235%	BH	100.000%	BH	98.845%
BI	100.000%	**MEAN =**	**99.552%**	BI	97.846%	BI	99.323%
BK	99.470%			BK	99.193%	BK	99.855%
MEAN =	**99.726%**			BQ	100.000%	AI	99.845%
				BS	100.000%	AO	99.716%
				BT	100.000%	AR	97.649%
				BU	95.750%	AS	86.669%
				BV	100.000%	AU	99.998%
				BX	100.000%	AW	99.848%
				BZ	100.000%	BC	99.988%
				TA	97.668%	BF	97.411%
				TB	99.688%	BG	94.456%
				MEAN =	**99.578%**	BH	99.378%
						BI	99.049%
						BQ	100.000%
						BS	100.000%
						BT	100.000%
						BU	94.982%
						BV	100.000%
						BX	100.000%
						BZ	100.000%
						TA	97.648%
						TB	99.900%
						AR	99.233%
						AS	70.103%
						AU	99.525%
						AW	91.617%
						MEAN =	**98.183%**

Table 4.1 The individual single vote scores, respectively for a) MetRef1 collection, b) MetRef2 collection, c) the 31 Different Donors from both collections, d) All 46 Pseudodonors.

One fundamental step in the development of this work is the identification of time stability of the metabotypes. In order to satisfy this request the spectra of the same individuals that are present in different collections (11 for MetRef1 and MetRef2, 4 for MetRef3) are used in a single vote classification as both training and test set (see **Table 4.2**)

MetRef1 → MetRef2		MetRef1 → MetRef3		MetRef2 → MetRef3	
AI	99.971%	AR	99.522%	AR	100.000%
AO	100.000%	AS	94.439%	AS	99.915%
AR	97.242%	AU	99.944%	AU	97.377%
AS	98.754%	AW	98.746%	AW	92.699%
AU	100.000%	**MEAN =**	**98.163%**	**MEAN =**	**97.498%**
AW	99.712%				
BC	99.434%				
BF	97.132%				
BG	98.645%				
BH	100.000%				
BI	98.723%				
MEAN =	**99.056%**				

Table 4.2 Single vote classification.
Different collections are used as training
and test sets.

As it is noticeable, the means are high and substantially similar, even if, using MetRef3 as training set and MetRef2 as test set, the value is slightly lower. These high percentages reported in the previous table demonstrate that the metabotypes substantially remain stable in a time scale of about two-three years and this is a significant basis to the medical implication of the metabotypes. Indeed this justifies that drugs are metabolized in different ways by different subjects and, therefore, they may have different both positive and adverse effects. Moreover, these findings suggest the necessity to develop all medical therapies in a personal way or, even better, in different ways according to various existing metabotypes. In details it is possible to note in the previous table that some individuals, as, for instance AS, present value of recognition slightly lower than others. To better understand what the cause of this result is, it is possible to see the distance colour-coded matrix for all individuals that participate in at least two collections reported in **Figure 4.5**. The dark blue diagonal represents the spectra of all individuals in the same collection, that are clearly associated to the lowest distances (dark blue colour code), while the other four light blue square diagonals are associated to the spectrum of the same individual in

different collections to demonstrate that there is a partial confusion between various collections due to a relative stability of the metabotypes over time. For the individual AS the light blue squares in these four diagonals are relative darker than for the other subjects, to highlight shortest distances between collections that cause a major confusion (see also **Table 4.1** column d). In conclusion AS has the most stable metabotypes in a short time scale.

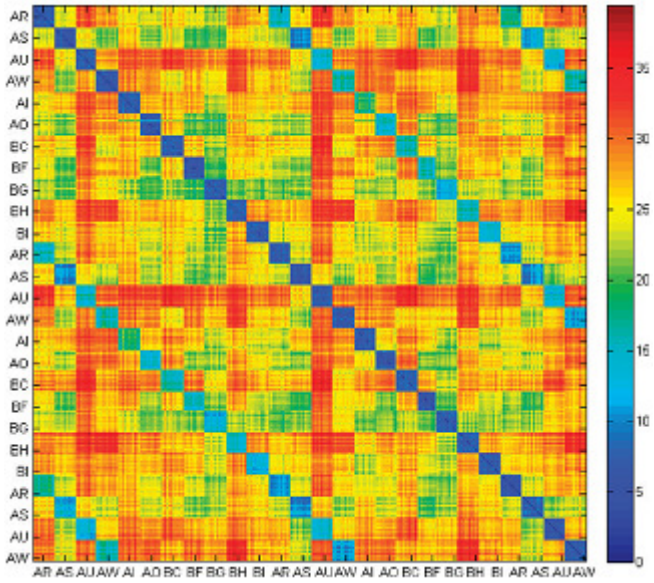


Figure 4.5 Distances in metabolic space

for all individuals that participated

in at least two collections.

Same results are also showed by the dendrogram in **Figure 4.6**, which is obtained considering each set of 40 spectra coming from the same individuals that participated to all three collections as different and not correlated pseudoindividuals.

This approach confirms the stability of the metabotypes. Indeed the same pseudoindividuals are clustered together and in some cases it is presented also a superimposition as, for instance, it has been highlighted for AS with **Figure 4.5**. Some considerations about the environmental and genetic contributions to metabotypes can be done by examining the previous figure. In particular TA and TB (homozygous twins) present the shortest fork in dendrogram **4.4**, followed by BU and BV that are father and son, indicating the presence of a genomic component. Furthermore, in the new

dendrogram the fork between TA and TB is one of the shortest and it is comparable with the forks among same pseudoindividuals. Practically TA and TB show a behaviour analogous to the same individuals in different collections in the dendrogram. These findings are also confirmed by the use of a single vote classification; the few misdiagnosed samples for TA and TB are assigned to one another, not casually assigned to other subjects as it happens for all individuals.

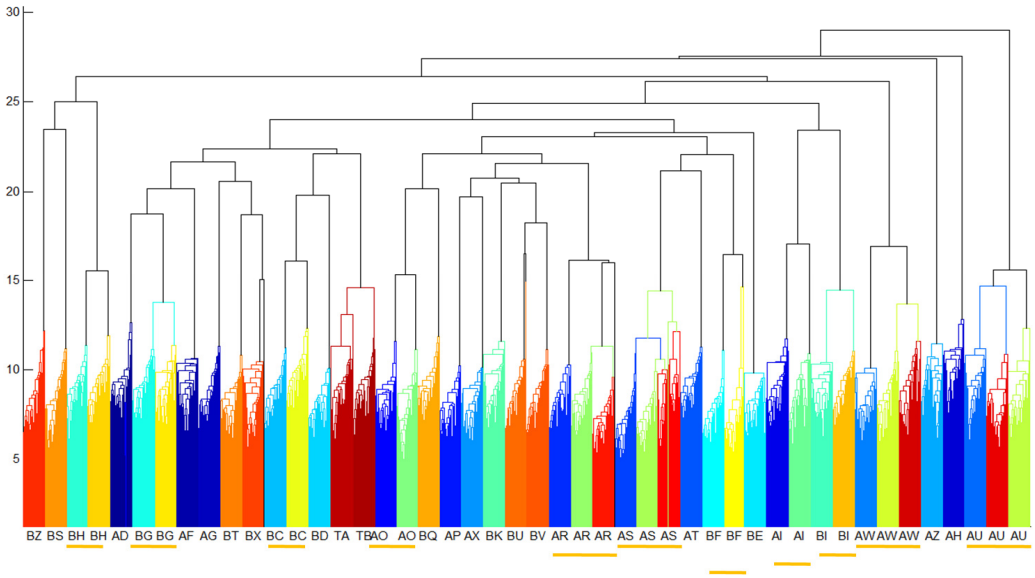


Figure 4.6 Dendrogram relative to cluster analysis on the 45-dimensional PCA/CA subspace for all the 46 different pseudoindividuals of all three collections.

Finally it is necessary to make a consideration on the variability of the spectra. Even in this work it is clearly shown the time stability of the metabotypes, the detailed visual analysis of the spectra reveals some differences between various collections. These differences are classified as “spikes”, “waves” and “jumps” and they seem principally due to diet effects. In particular, i) spikes are signals that appear in a single spectrum having marked differences in intensity between preceding and following days, ii) waves are signals that have a more gradual variation in intensity but persist for more days, iii) jumps are signals that markedly appear in a collection as spikes but they remain practically unchanged for all the collection and, sometimes, they are carried to next collection and are reversible. Spikes are principally due to particular food intake,

for instance the peak of TMAO (TriMethylAmine-N-Oxide) (9) clearly appears in spectra of urine collected the day after a fish diet intake, as well mannitol appears after chewing gum intake. At the same time many other events can cause a spike, as very intense physical activity (lactate), drug consumption (paracetamol-O-glucoronide), excessive alcohol intake (ethanol). Moreover, it is present in a BF subject a case of "anti-spike", due to the disappearance of citrate signals, probably associated to a kidney metal complexation. Nevertheless, a particular diet is one of the causes of some jumps, in particular for individual AW it is observed a jump for xanthosine (10) due to an excessive meat consumption during a period of several days. A great number of jumps and waves are really due to the modification of gut microflora. It is noted that the peaks involved in waves and jumps are principally due to metabolites as hippurate, *meta*-hydroxyphenylpropionate, formate and phenylacetylglycine that are usually products of the metabolism of intestinal bacteria (11-15) and, therefore, linkable to activity variations of gut microflora. This last consideration leads us to consider the individual metabolic phenotypes as a metagenomic entity strongly affected by host genotype, environmental factors and gut microbiome.

Bibliography

(1) Gavaghan C. L., Holmes E., Lenz E., Wilson I. D., Nicholson J. K.; *An NMR-based metabonomic approach to investigate the biochemical consequences of genetic strain differences; application to the C57BL10J and Alpk:ApfCDmouse* FEBS Lett. **2000** 484 169:74.

(2) Rezzi S., Ramadan Z., Fay L. B., Kochhar S.; *Nutritional metabonomics: applications and perspectives* J. Proteome Res. **2007** 6 513:25.

(3) Kussmann M., Raymond F., Affolter M. *OMICS-driven biomarker discovery in nutrition and health* J. Biotechnol. **2006** 124 758:87.

(4) Clayton T. A., Lindon J. C., Antti H., Charuel C., Hanton G., et al.; *Pharmaco-metabonomic phenotyping and personalised drug treatment* Nature **2006** 440 1535:42.

(5) Nicholson J. K., Wilson J. D.; *Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism* Nat. Rev. Drug Discovery **2003** 2 668:76.

(6) Nicholson J. K., Holmes E., Lindon J. C., Wilson I. D.; *The challenges of modeling mammalian biocomplexity* Nat. Biotechnol. **2004** 22 1268:74.

(7) Tiret L.; *Gene environment interaction: a central concept in multifactorial diseases* Proc. Nutr. Soc. **2002** 61 457:63.

(8) Assfalg M., Bertini I., Colangiuli D., Luchinat C., Schäfer H., et al.; *Evidence of different metabolic phenotypes in humans* Proc. Natl. Acad. Sci. U.S.A. **2008** 105 1420:4.

(9) Rabinowitz L., Berlin R., Yamauchi H.; *Plasma potassium and diurnal cyclic potassium excretion in the rat* Am. J. Physiol. **1987** 253 F1178:81.

(10) Young D. S., Epley J. A., Goldman P.; *Influence of chemically defined diet on the composition of serum and urine* Clin. Chem. **1971** 17 765:73.

(11) Phipps A. N., Stewart J., Wright B., Wilson I. D.; *Effect of diet on the urinary excretion of hippuric acid and other dietary-derived aromatics in rat. A complex interaction between diet, gut microflora and substrate specificity* Xenobiotica **1998** 28 527:37.

(12) Williams R. E., Eyton-Jones H. W., Farnworth M. J., Gallagher R., Provan W. M.; *Effect of intestinal microflora on the urinary metabolic profile of rats: a (1)H-muclear magnetic resonance spectroscopy study* **2002** 32 783:94.

(13) Goodwin B. L., Ruthven C. R., Sandler M.; *Gut flora and the origin of some urinary aromatic phenolic compounds* **1994** 47 2294:7.

(14) Nicholls A. W., Mortishire-Smith R. J., Nicholson J. K.; *NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats* Chem. Res. Toxicol. **2003** 16 1395:404.

(15) Samuel B. S., Gordon J. I.; *A humanized gnotobiotic mouse model of host-archeal bacterialmutualism* Proc. Natl. Acad. Sci. U.S.A. **2006** 103 10011:6.

6 METABOLOMICS STUDIES ON HUMAN PLASMA

Aim of the work

This epidemiological study on healthy subjects was born with the aim to better understand the complex biological mechanisms that occur in living organisms. In order to reach this goal it has been decided to analyze a great number of EDTA-plasma samples (not fewer than 1000), that are collected from healthy blood donors in the Hospital of Pistoia. Due to this high number of samples, the project is quite widespread and complex, and therefore some sub-projects are identified, with specific goals to reach:

1.Relationship study between altered blood parameters and metabolic fingerprints (Altered Blood Parameters) .

2.Relationship study between peculiar behaviours of sample donors and metabolic fingerprints (Peculiar Behaviours).

3.Relationship study between SNPs genetic polymorphisms and metabolic fingerprints (SNPs).

Altered Blood Parameters

A considerable amount of information is present in plasma NMR spectra. In particular, some molecules have peaks arising in plasma spectra, such as cholesterol, LDL, HDL, glucose and so on. These molecules are usually considered as “healthy state signals” for each individual, and for them it is given a range of confidence associated to the normal condition. However, they are not sufficient to completely define both healthy status and percentage of risk of cardiovascular pathologies onset. Therefore, current European guidelines for the prevention of coronary heart diseases recommend to base any intervention on the evaluation of total risk (1). In order to estimate such a risk, several predictive equations have been developed in the last decade (2-3). Possibly also metabolomics can be used as a method to define some of these factors of risk, so metabolomics spectra contain all the possible metabolic information that is present in blood. In order to check this hypothesis, a first step consists in the analysis of the relationship existing between the previous value determined by the standard method

of analysis and the same value predicted by the metabolomics statistical analysis. Practically, it is necessary to determine the existence of any kind of correlation between the first ones (that can be called true values for simplicity) and the second ones (that are predicted values), in order to verify the real capability of metabolomics to highlight standard blood values. The second step is the real definition of a metabolomics risk. Some other parameters, also associated with individual behaviour, can be considered to solve this goal.

Peculiar Behaviours

It is generally assessed that the healthy state of an organism is strictly related to the life style, besides genetic factors. Therefore, some of the so-called "proper" behaviours have been identified as well as "not-proper" ones. For instance, a balanced diet and a steady physical activity are usually considered as a healthy pattern for an individual. In the same way some habits are considered risky or harmful, such as smoking cigarettes, high alcohol intake and so on. Many of these "not-proper" behaviours are largely widespread in modern society, especially in the most developed countries. Thus their impact on the individual metabolism, and therefore, on the individual healthy state have been already studied, see for instance the studies on the increment of risk percentage of lung cancer in smokers. There has been an increasing interest in the application of the metabonomic approach also to these fields recently (4-6). Indeed there is obviously a great advantage in the application of the metabonomic approach to the exam of the "un-proper" life styles. It is clearly possible to supervise the contemporaneous alterations of various metabolic pathways. In this way it is not only given a complete metabolic definition of all specific metabolic alterations due, for instance, to smoke but it is also possible to investigate how individual metabotypes can react to this and be able to define a possible percentage of risks of the onset of correlated pathologies. Naturally this step is strictly related to the previous one (altered blood parameters); most of these "not-proper" behaviours cause a significant alteration of blood parameters before the pathology.

SNPs

The last goal is to discover a relationship between genome and metabolome. It is assessed how the genome is one of the principal causes of the healthy state of a subject. Moreover, some gene mutations can cause serious pathologies that are not compatible with life in many cases.

Nevertheless, some other mutations are not strictly correlated with pathologies. Some of them are called SNPs, or Single Nucleotide Polymorphism, and generally are responsible for different enzymes (isoforms) in different subjects. Basically a SNP is a single variation in DNA sequence generally due to the changing of one of the nucleotides. In details SNPs may be changed (substitution), removed (deletion) or added (insertion) to a polynucleotide sequence. When there is this kind of modification, two different alleles exist for a gene and almost all common SNPs have only two alleles. It is supposed that these single variations in the DNA can affect the way humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine (7). It is noted that the existence of SNPs is relatively common for genes coding for enzymes involved in glucidic (8) and lipidic (9-10) metabolism. Therefore, the polymorphisms that will be considered in these projects are associated to genes with an important metabolic role. In details 4 genes are selected: i) PPARγ2 (Peronisome Profiletor Activated Receptor γ2 ii) LIPC (hepatic LIPase gene) iii) FADS1 (Fatty Acid DeSaturase) iv) SREBP-1c (Sterol Regulatory Element Binding Protein 1c).

PPARγ2 is a gene involved in the metabolism of lipids and glucose. The most common polymorphism is Pro12Pro, while the other is Pro12Ala. In a recent study it is noted that Pro12Ala polymorphism is associated with a diminution of activity of the hepatic receptor for the uptake of glucose and a higher sensibility to insuline (11).

LIPC is involved in HDL metabolism. The selected polymorphisms for the gene are rs12593008, rs261342 and rs4775041. It seems that their expression is associated with a low level of HDL, especially in women (9).

FADS1 is related to the enzyme Delta-5-Desaturase, that plays a fundamental role in the fatty acids metabolism. Indeed it is the responsible for the desaturation of acilic

chains of fatty acids. Two different polymorphisms are selected: rs174548 and rs3834458. The first one seems to be associated with level variations of HDL, LDL, and free cholesterol in serum (10). The second one seems to induce the increasing of the selected substrates for the enzyme (12).

SREBP-1c is a gene that codifies for an important transcription factor of both lipidic and glucidic metabolism. The selected polymorphisms are rs2297508 and rs11868035. They are associated with type 2 diabetes (13) and LDL altered levels (8).

Partial Results and Perspectives

Preliminary results are obtained in a widespread number of blood samples, even if not definitive. More than 800 samples are collected and analyzed (809). All these samples are EDTA-plasma. At the same time some important clinical data are collected from the recruited donors. These important data are anonymized for ethical reasons. Thus for every sample (and therefore every donor), a few items of information are present such as gender, smoke and drug intake, blood pressure, but also glycemia, cholesterol and triglicerides. Moreover, also LDL and HDL values are present for a lower number of samples/subjects (about 200). Initially it is decided to start with a simple classification of samples based on some of these classical blood parameters such as cholesterol, glycemia, LDL, HDL and ratio cholesterol-HDL. The last one is obviously easily derivable from the other data and it is an important blood parameter; it explains the ratio between the so-called “worst-cholesterol” and the so-called “good-cholesterol”. Therefore, if this ratio is not high, the subject does not have theoretically any risk of onset of cardiovascular disease even if cholesterol level is higher than normal. In order to perform the first preliminary statistical analysis, it is decided to constitute two classes of samples/subjects based on each of these blood parameters. For instance, the two classes of samples for glycemia parameter are formed with the following criterium: they contain the 10 per cent of samples having respectively the low values and the high values of the parameter (tails). All classes are correctly divided, see **Table 5.1**. This is important for many reasons. First of all, it is checked the quality of metabolomic analysis of collected samples and, furthermore, it is demonstrated the metabolomic skill to clearly highlight specific alterations of important blood

parameters, even if they are not out of classical "range of normality": for instance, LDL normal values have to be comprised between 70 and 180 mg per 100 mL, but the chosen classes present respectively samples with value of LDL lower than 110 (and not 70) and higher than 140 (and not 180), and, therefore, subjects that are not defined at risk in a classical way are present in the analysis too and they are separated on the basis of their metabolomic profile.

		CPMG Spectra			Noesy Spectra		
		% Correct	% Wrong	Accuracy	% Correct	% Wrong	Accuracy
Glycemia	HIGH (> 105)*	94.78%	5.22%	90.18%	88.13%	11.87%	86.99%
	LOW (< 78)	85.97%	14.03%		88.32%	11.68%	
Cholesterol	HIGH (>255)	96.84%	3.16%	96.52%	94.64%	5.36%	95.95%
	LOW (< 160)	98.32%	1.68%		96.43%	3.57%	
LDL	HIGH (> 140)	92.51%	7.49%	90.72%	96.10%	3.90%	95.36%
	LOW (< 110)	89.24%	10.76%		95.74%	4.26%	
HDL	HIGH (> 70)	91.58%	8.42%	91.88%	92.15%	7.85%	92.50%
	LOW (< 39)	92.46%	7.54%		92.92%	7.08%	
Triglicerides	HIGH (> 164)	95.38%	4.62%	96.45%	94.96%	5.04%	96.42%
	LOW (< 51)	96.54%	3.46%		97.11%	2.89%	
Chol./HDL	HIGH (> 4.49)	92.35%	7.65%	93.78%	95.18%	4.82%	96.41%
	LOW (<3.51)	95.21%	4.79%		98.39%	1.61%	

* = each value is expressed in mg/100 mL with exception of Chol./HDL that is a ratio

Table 5.1 : SVM classification of samples based on PLS-CA scores

What happens for LDL, also happens for all the other considered parameters. Starting from these findings, a second level of analysis is to clearly identify the signals responsible for these correct classifications. Obviously, the principal candidates are the signals due to the same parameters which are selected as classifiers. Indeed it is normal to expect that the principal cause of separation is glucose signal in case of glycemia classes, and so on. Moreover, it is also possible for some other signals to be

quite differently represented in each of the two classes of spectra. This supposition is also suggested by the fact that different lipidic classes (as high LDL/low LDL such as high cholesterol/ low cholesterol) are correctly classified also taking into account CPMG spectra, that do not have a great contribution of lipid signals. This finding suggests that probably other signals and, therefore, other metabolites are involved in the classification of the two classes. Before trying to evaluate which these metabolites are, another interesting test can be carried out. The samples which have been considered for glycemia are re-classified again in the same classes as before taking off glucose signals. Basically, it has been tried to demonstrate that subjects with low value of glycemia are significantly metabolomic different from subjects with high value of glycemia, even if the glucose and other carbohydrates signals are cutting off.

		CPMG Spectra			Noesy Spectra		
		% Correct	% Wrong	Accuracy	% Correct	% Wrong	Accuracy
Glicemia (No Glucose)	HIGH (> 105)*	81.72%	18.28%	82.83%	81.80%	18.20%	81.96%
	LOW (< 78)	84.55%	15.45%		82.17%	17.83%	

*= mg/100 mL

Table 5.2 : Classification of samples basing on glycemia (glucose peaks cut off).

The quite high value of correct classification obtained, more than 80% (**Table 5.2**), definitively suggests that specific alterations of metabolites different from glucose are present between the two glycemia classes and, thus, the metabolism of a low glycemia subject is completely different from a high glycemia subject. The identification of these metabolites is the following step in order to highlight which metabolic pathways are altered by high value (or low value) of glycemia, rather than cholesterol, LDL and so on.

A third approach has been used to investigate the data, starting from previous results. Indeed, the finding of a strict relationship between the considered tails of each parameter (10% high and low for each one) and metabonomic profiles of the same samples leads to have the same relationship for each value of each parameter, even if not comprised in the previously taken tails. Practically it is tried to check if a linear regression between the value of parameters coming from classical blood analysis and

the value of the same parameters determined by using multivariate statistical analysis on metabolomic data (PLS-CA scores) exists. As it is possible to see in the **Table 5.3**, these relationships exist for all considered parameters (I. E., glycemia, cholesterol and triglicerides are not yet analyzed). The first conclusion that can be done is that metabolomics could be used as predictive tool of these blood values. Clearly this is an important finding, even if the used classical methods are probably more accurate and, above all, are cheaper and easily applicable.

	CPMG Spectra	Noesy Spectra
LDL	0,8374	0,8803
HDL	0,8297	0,8130
Chol./HDL	0,8577	0,8824

Table 5.3 : R^2 for correlation between predicted values (NMR metabolomics) and true values (classical blood analysis).

However, the existence of this relationship suggests the possibility to define various metabotypes basing on these parameters, but also on other ones, such as BMI, smoke intake and so on. These metabotypes could be related with the risk of onset of cardiovascular disease and, therefore, lead to the development of an innovative strategy to define the risk in subjects that have, or seem to have, all classical blood parameters inside the normal values of confidence. In order to do this it will be necessary: i) to check the "presence" of all these clinical parameters inside the NMR spectra, through the verification of a strict relationship between their real values and the predicted ones through NMR analysis, as it has been done for cholesterol, triglicerides, LDL, HDL and ratio cholesterol-HDL, ii)to consider also other parameters, such as smoke, alcohol and drugs intake and check their influence in the metabonomic profiles of the previously chosen classes, iii) to obtain one value, that can be considered as a score, that is sensitive of the metabolomic risk of cardiovascular disease and that is useful as evaluation parameter of the same risk, iv) highlight the

difference and the similarity between the “newborn” metabolomic risk and the classical risk, although actually it does not exist a clear and unique parameter of risk accepted by the international community (1), in order to understand difference, similarity, strength and weak points of the new approach, v) validate the results.

Bibliography

(1) Graham I. Atar D., Borch-Johnsen K., Boysen G., Burell G., et al.; *European guidelines on cardiovascular disease prevention in clinical practice: executive summary* Eur. Heart J. **2007** 28 2375:414.

(2) Conroy R. M., Pyorala K., Fitzgerald A. P., Sans S., Menotti A., et al.; *Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project* Eur. Heart. J. **2003** 24 987:1003.

(3) Ferrario M., Chiodini P., Chambless L. E., Cesana G., Vanuzzo D., et al.; *Prediction of coronary events in a low incidence population. Assessing accuracy of the CUORE cohort study prediction equation* Int. J. Epidemiol. **2005** 34 413:21.

(4) Vulimiri S. V., Misra M., Hamm J. T:, Mitchell M., Berger A.; *Effects of mainstream cigarette smoke on the global metabolome of human lung epithelial cells* Chem. Res. Toxicol. **2009** 22 492:503.

(5) Loulhelainen N., Myllernierni M., Rohman I., Kinnula V. L.; *Airway biomarkers of the oxidant burden in asthma and chronic obstructive polmonary disease: current and future perspectives* Int. J. Chron. Obstruct. Pulmon. Dis. **2008** 3 585:603.

(6) Mi P. E., Lee E., Jin T. H., Oh E., Lee J., et al.; *Inter- and intra-individual variations of urinary endogenous metabolites in healthy male college students using (1)H NMR spectroscopy* Clin. Chem. Lab. Med. **2009** 47 188:94.

(7) Mitani Y., Lezhava A., Sakurai A., Horikawa A., Nagakura M., et al.; *Rapid and cost-effective SNP detection method: application of SmartAmp2 to pharmacogenomics research* Pharmacogenomics **2009** 10 1187:97.

(8) Liu J. X., Liu J., Li P. Q., Xie X. D., Guo Q.; et al.; *Association of sterol regulatory element-binding protein-1c gene polymorphism with type 2 diabetes mellitus, insuline resistance and blood lipid levels in Chinese population* Diabetes Res. Clin. Pract. **2008** 82 42:7.

(9) Feitosa M. F., Myers R. H., Pankow J. S., Province M. A., Borecki I. B.; *LIPC variants in the promoter and intron 1 modify HDL-C levels in a sex-specific fashion* Atherosclerosis **2009** 204 271:7.

(10) Schaeffer L., Gohlke H., Muller M., Heid I. M., Palmer L. J., et al.; *Common genetic variants of the FADS1/FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids* Hum. Mol. Genet. **2006** 15 1745:56.

(11) Honka M. J., Vanttinen M., Iozzo P., Virtanen K. A., Lautamaki R., et al.; *The Pro12Ala polymorphism of the PPARgamma2 gene is associated with hepatic glucose uptake during hyperinsulinemia in subjects with type 2 diabetes mellitus* Metabolism **2009** 58 541:6.

(12) Martinelli N., Girelli D., Malerba G., Guarini P., Illig T., et al.; *FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease* Am. J. Clin. Nutr. **2008** 88 941:9.

(13) Dentin R., Girard J., Postic C.; *Carbohydrate responsive element binding protein (ChREBP) and sterol regulatory element binding protein-1c (SREBP-1c): two key regulators of glucose metabolism and lipid synthesis in liver* Biochimie **2005** 87 81:6.

7 CONCLUSIONS AND PERSPECTIVES

Besides the perspectives opened by each project, that are reported in proper section of each proper chapter, it is generally demonstrated the feasibility of metabolomics in the study and interpretation of various pathology. Therefore the metabolomics approach can be easily extended to some other pathologies. Indeed some other metabolomics projects are already started or are in a starting-phase. These new projects involve some widespread pathologies such as diabetes (type II), lung carcinoma, Obstructive and Chronic BroncoPathy (BPCO), liver cyrrhosis and carcinoma. The study of these new pathologies lead us to define better their characteristic, but also allow us to clearly understand what are the ideal targets and the limits of the metabolomic approach. Moreover, regarding the definition of metabotypes, it is necessary to univocally define what are the metabotypes and how exactly different metabotypes are related to different metabolic capacities of individuals.

Furthermore it is necessary to continuously improve the step of the standardization of the samples. In general the metabolome changes depending on the body fluid tested, on the method of analysis, and on sample collection and handling. This highlights the need for the optimisation of pre-analytical tools for metabolomic analysis of biological samples, both because this is a relatively new technique and because the analytical technique itself is still undergoing optimisation. In other words, the technique itself cannot be advanced without the standardisation of sample selection and preparation. This is particularly true and necessary when it is analyzed some peculiar fluids, such as exhaled breath condensate or saliva for instance. Nevertheless, much research work is still needed in order to define the best protocols for sample collection and preparation. Sample handling and storage may strongly affect metabolomic profiles due to differing stabilities of the various metabolites. Even in the presence of "stability-optimised" samples, a strong limiting factor in the practical evaluation of diseases using a metabolomic approach lies in the intrinsic variability of human metabolic samples. Any study aimed at the identification of relevant metabolites should be presented with reference to the normal or control population. In order to be able to identify relevant metabolic changes, identification of metabolic fingerprint is usually needed as opposed to changes in concentration of a single biomarker.

However, the variations in metabolic fingerprinting are usually extremely small and therefore multiple sampling is required to eliminate the background noise due to personal variability. Sample collection, handling and storage are all critical steps for the detectability of metabolites by NMR and need to be optimised separately for each biofluid and tissue extract. For example NMR spectra of urine are dominated by thousands of sharp lines from predominantly low molecular mass metabolites, and the detectability of each metabolite is limited only by its concentration and stability over time. On the contrary, blood plasma and serum contain both low molecular and high molecular mass components. Tendency toward aggregation of proteins and protein-small molecule interactions may cause disappearance of the signals of metabolites due to line broadening, even if the metabolites remain stable over time.

Moreover this is valid also for statistical and NMR methods. The development of new simple 1D sequences of acquisition, an example is constituted by the 1D diffusion filtered acquisition (a sort of dosy1d), can be very useful to single out some information presented in samples, in case of 1D diffusion it is possible to obtain information about only macromolecules presented in serum, and in general in a biofluid. Moreover the development of new statistical analysis could further increase the quantity of obtainable data.

8 MATERIALS AND METHODS

Sample Preparation

All samples are collected in hospital structures and, immediately, frozen to – 80°C in order to avoid metabolites degradation. On the day of the analysis samples are thawed at room temperature and shaken before us. The following steps are slightly different basing on type of biofluids. Six hundred and thirty microliters of urine are added to 70 microliters of a sodium phosphate buffer in deuterium oxide ($2H_2O$). The buffer is principally constitued by Na_2HPO_4 (0.2 M) and NaH_2PO_4 (0.2 M); pH is standardized at 7.0 to minimize variations in metabolite NMR chemical shifts. Moreover, the buffer also contains sodium trimethylsilyl [2,2,3,3-$2H_4$]propionate (TSP) (10mM) and sodium azide (NaN_3) (30mM), the first one is used to center the spectra at 0.00 ppm, the second one is a bacteriostatic. Samples are centrifuged at 14 000*g* for 5 minutes to remove any solid debris. Blood samples, both serum and plasma, are simply prepared adding 300 microliters of them to 300 microliters of another sodium phosphate buffer in 20% (v/v) of $2H_2O$. The buffer contains Na_2HPO_4 (70mM) and the pH is standardized at 7.4. Moreover, sodium trimethylsilyl [2,2,3,3-$2H_4$]propionate (TSP) (0.8% w/v) and sodium azide (NaN_3) (30mM) are present in the buffer. A total of 450 microliters of urine supernatant or serum are transferred into 4.25 mm outer-diameter NMR tubes.

NMR Experiments and Bucketing

One dimensional (1D) ^1H NMR spectrum has been acquired at 600 MHz spectrometer. A spectrum of each sample is acquired with water peak suppression using a standard noesypresat1D pulse sequence. This sequence has been chosen in order to optimize sample acquisition basing on some parameters such as sensitivity, reproducibility and robustness. In particular it is chosen a presaturation sequence to partially erase the water peak because it is less invasive than other techniques as water gate-based (peaks are much easier to integrate). Furthermore, the noesypresat1d sequence presents some advantages also with respect to standard zgpr and excitation sculpting (es). Indeed it is more sensitive than the sculpting for the absence of the shape pulses, and with respect to zgpr it is possible to experimentally have a better baseline.

Furthermore, serum samples are also acquired using a Carr-Purcell-Meiboom-Gill (CPMG) spin-echo sequence to suppress signals arising from high molecular weight molecules. This suppression is due to the capability of the mathematical CPMG filter to eliminate the signals with a short T2 such as macromolecules. Spectra are collected with 64 scans and 4 dummy scans. Thus the duration of each sequence is very short, about 8 minutes. Each 1D spectrum is segmented (in the range between 10.00 ppm and 0.02 ppm in to 0.02-ppm chemical shift buckets, and the corresponding spectral areas are integrated using AMIX software. Regions between 6.0 and 4.5 ppm, containing residual water and urea signals (in urine) are removed. The normalization is then carried out on the data prior to pattern recognition

Multivariate Statistical Analysis

Multivariate statistical analysis is a tool to examine relationships among a great number of statistical variables at the same time (multivariate data). Generally multivariate data are represented by a matrix. Each row of this matrix corresponds to an object, while each column is a peculiar and observable characteristic of the objects, for instance a bucket or a bin in metabolomics. Many methods can be applied in order to perform a multivariate statistical analysis. Substantially the aim of these methods is obtaining a clusterization without giving data information to the systems (unsupervised methods). Basically, they search for correlating variables that allow a division of the objects into two or more classes. The principal and most used of these methods is PCA (Principal Components Analysis), even if other methods as HCA (Hierarchical Cluster Analysis) and Kohonen maps (also called SOMs) are used. Additionally to the *X* matrix containing data, another property, called *y*, may be given for each subject. This property adds a piece of information about the "nature" of each sample, as the class. These methods are supervised and therefore on a priori knowledge about subjects. The most common methods used are PLS (Partial Least Square) regression (1) and their variants as O-PLS or K-PLS, but some other methods are also applied, as ANN (Artifical Neural Networks) and an informed version of Kohonen maps. Moreover, both supervised and un-supervised methods are

frequently applied in combination with methods as Canonical Analysis (CA) and Regularized Canonical Analysis (RCC). These methods are usually applied on the PCA/PLS scores to enlarge the separation between the groups. After the separation step, it is necessary to classify the samples in order to decide which class each of them belongs to.. The principal tools to be applied in classification are k-Nearest Neighbor (k-NN) method, Linear Discriminant Analysis (LDA) (2) and derived methods, Support Vector Machine (SVM) method (3) and Random Forest method (4).

The following step is the validation of the results. The key of the validation of the NMR classification on class identity relies on Test Set Validation (TSV) approach which requires that models do not have any knowledge of the existence of any test set data. Monte Carlo cross-validation (MC), Leave One Out (LOO) validation and k-Fold validation are the principal methods used in this step (5-6). MC methods randomly split the dataset into two different set of data: training set and validation set. The training set is used to determine the statistical model, while the validation set is used to assess predictive accuracy. In k-Fold validation, the original dataset is divided into K subsets. K-1 subsets are used as training sets and the remaining subsets as validation sets. Finally LOO works taking only one sample as validation set, whilst all other samples constitute the training set. All these methods are usually applied with 1000 iterations, randomly varying the composition of both training and validation set, in order to obtain robust results.

Finally, to assess which bucket is significantly different among various classes a one-way analysis of variance is used. Normality of the data distributions is assessed using the Jarque-Bera normality test (7). Statistical significance of the means over the two groups is assessed using ANOVA or the non-parametric analogue Kruskal-Wallis test by using the Bonferroni correction on a nominal value of 0.005.

Principal Component Analysis (PCA)

PCA is an orthogonal linear combination of the original variables (normally called buckets or bins in metabolomics) in order to highlight the maximum variance among themselves. The mathematical formula can be expressed in this way:

$$Y^T = W X^T$$

where *X* is the original matrix data and *W* is the matrix of the eigenvectors (see below) for linear combination.

PCA was invented in 1901 by Karl Pearson, implemented in 1933 by Hotelling. The principal goal of PCA is the reduction of data, whilst it is retaining as much as possible the original variation present in the data set. This reduction is easily achieved by taking *p* variables from the data set (in metabolomics *p* buckets) X1, X2,…, Xp and linearly combining them to produce uncorrelated principal components (PCs) PC1, PC2,…, PCp. Moreover, these combinations are due so that the maximum variation among samples is expressed in PC1, the second greatest amount of variation in PC2 and so on. Clearly if the linear combination works well, the complete variability in the data set has to be adequately expressed by means of a few PCs. Se vuoi dire pochi few, alcuni a few

Mathematically the analysis is performed on a data set of p variables ($X_1, X_2, \ldots, X_p$) for n individuals, as indicated below in **Table 6.1.**

Variable / Individual	X_1	X_2	…	X_p
1	X_{11}	X_{12}	…	X_{1p}
2	X_{21}	X_{22}	…	X_{2p}
.	.	.		.
.	.	.		.
.	.	.		.
n	X_{n1}	X_{n2}	…	X_{np}

Table 6.1 Data matrix for principal component analysis

It is necessary to calculate the corresponding covariance matrix:

$$\mathrm{Cov}(X_j, X_k) = \frac{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)}{(n-1)}$$

$$\text{where } \overline{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n}, \text{ and } j, k = 1, 2, \ldots, p.$$

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} \ldots & s_{1p} \\ s_{21} & s_{22} & s_{23} \ldots & s_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ s_{p1} & s_{p2} & s_{p3} \ldots & s_{pp} \end{bmatrix}$$

where S is the covariance matrix, s_{jk} is the covariance of variables X_j and X_k when $j \neq k$ and the diagonal elements s_{jj} represents the variance of variable X_j when $j = k$. original variation present in the data set. PCs and their associated eigenvalues are found in the sample covariance matrix through an iterative calculation process.

In details, the first principal component (PC1) is then a linear combination of the original variables $X_1, X_2, \ldots, X_p$,

$$PC_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \ldots + a_{1p}X_p = \sum_{j=1}^{p} a_{1j}X_j$$

that varies as much as possible for the individuals, subject to the condition that

$$a_{11}^2 + a_{12}^2 + a_{13}^2 + \ldots + a_{1p}^2 = 1$$

where $a_{11}, a_{12} \ldots, a_{1p}$ are coefficients assigned to the original p variables for PC1.

Therefore, the eigenvalue of PC1 is as large as possible given this constraint on the constant a_{1j}. The constraint must be imposed in order to avoid the increasing of the eigenvalue of PC1 by simply increasing one or more of the a_{1j} values.

Similarly, the second principal component,

$$PC_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \ldots + a_{2p}X_p,$$

is such that eigenvalues of PC2, are as large as possible subject to the constraint that:

$$a_{21}^2 + a_{22}^2 + a_{23}^2 + \ldots + a_{2p}^2 = 1$$

and also on the condition that PC2 is uncorrelated with PC1. The third principal component:

$$PC_3 = a_{31}X_1 + a_{32}X_2 + a_{33}X_3 + \ldots + a_{3p}X_p,$$

is such that the eigenvalue of PC3 is as large as possible subject to the constraint that:

$$a_{31}^2 + a_{32}^2 + a_{33}^2 + \ldots + a_{3p}^2 = 1,$$

and also on the condition that PC3, PC2 and PC1 are uncorrelated. All other principal components are obtained in a similar way. When the eigenvectors (or PCs) are obtained, the following step is to select the components that better describe the system. Generally in PCA the number of extracted components is equal to the number of analyzed variables, but in general the last few components do not account for much of the variance and can be ignored. It is not easy to identify the correct number of components that must be chosen to correctly describe a system. Someone suggests that the first 6 components explain 70% to 80% of the total variation. Probably the best method to assess the adequate number of PCs is the so-called eigenvalue-one criterion (Kaiser criterion). Basing on this criterion, only the PCs that have eigenvalues higher or equal to one are retained. Indeed if a PC has an eigenvalue lower than 1, this means that contains less information than one of the original variables and it is

discarded. Actually the value of 1 as cut-off is considered too selective and the correct cut-off value is experimentally determined in 0.7. However, the results of PCA are currently expressed in terms of scores (PCs) and loadings (eigenvalues).

Partial Least Square (PLS) regression and derived methods

The PLS is a common supervised method that was first introduced by Wold in 2001. It commonly relates to two different matrices: *X*, that usually contains spectral or chromatographic data, and *Y*, comprising quantitative characteristics of samples, as for instance class belongings. Substantially PLS finds the multidimensional direction in the *X* space that explains the maximum multidimensional variance direction in the *Y* space. PLS modelling is mathematically expressed by the following formulas:

$$X = TP^T + E$$

$$Y = TC^T + F$$

where *T* is a score, or component, matrix (as see above for PCA), *P* and *Q* are respectively the loading matrices of *X* and *Y* and the terms *E* and *F* are the errors. A recent development of PLS potentiality is obtained with OPLS. There is in OPLS the separation of *X* contribution in two parts: one is linearly related to *Y* matrix and the other one is orthogonal and, therefore, independent.

$$X = T_p P_p^T + T_o P_o^T + E$$

$$Y = T_p C_p^T + F$$

Bibliography

(1) De Jong S.; *SIMPLS: An alternative approach to partial least squares regression* Chemom. Intell. Lab. Syst. **1993** 18 251:63.

(2) Friedman J. H.; *Regularized Discriminant Analysis* J. Am. Stat. Ass. **1989** 84 165:75.

(3) Vapnik V. N.; *The nature of statistical learning theory* Springer-Verlag New York **1995**.

(4) Breiman R. F.; *Random Forest* Mach. Learn. **2001** 45 5:32.

(5) Picard R., Cook D.; *Cross-Validation of regression models* J. Am. Stat. Ass. **1984** 79 575:83.

(6) Efron B., Tibshirani R.; *Improvements on cross-validation: the .632 + Bootstrap method* J. Am. Stat. Ass. **1997** 79 575:83

(7) Jarque C. M., Bera A. K.; *Efficient tests for normality, homoscedasticity and serial independence of regression residual* Econ. Lett. **1980** 6 255:59.

9 PUBLICATIONS

List of publications

1. Bertini I., Calabrò A., De Carli V., Luchinat C., Nepi S., Porfirio B., Renzi D., Saccenti E., Tenori L., *The metabonomic signature of celiac disease* J. Proteome Res. **2009** 8 170:7.

2. Bernini P., Bertini I., Luchinat C., Nepi S., Saccenti E., Schäfer H., Schütz B., Spraul M., Tenori L., *Individual human phenotypes in metabolic space and time* J. Proteome Res. **2009** 8 4264:71.

3. Oakman C., Tenori L., Claudino W. M., Cappadona S., Nepi S., Battaglia A., Bernini P., Zafarana E., Saccenti E., Destefanis M., Fornier M., Morris P., Biganzoli L., Luchinat C., Bertini I., Di Leo A., *Identification of a serum-detectable metabolomic fingerprint potentially correlated with the presence of micrometastatic disease in early breast cancer patients at different risks of disease relapse by traditional prognostic factors* Breast Cancer Res. SUBMITTED.

Author's contribution to each work

Chapter 3, 4, 5 : Sample standardization and preparation, NMR spectra acquisition and standardization, NMR processing and assignment, Metabolites idenfication, Interpretation of metabolic data coming from statistical analysis, Theorical speculation.

Chapter 6 : Sample standardization and preparation, NMR spectra acquisition and standardization, NMR processing and assignment , Statistical analysis and development Metabolites idenfication, Interpretation of metabolic data coming from statistical analysis, Theorical speculation.