Alma Mater Studiorum
Università degli Studi di Bologna

Dottorato di Ricerca in

Sanità Pubblica e Medicina del Lavoro XX ciclo

Settore Scientifico Disciplinare: CHIM/01

MASS SPECTROMETRY-BASED PROTEIN PROFILING STRATEGIES FOR BIOMARKER DISCOVERY IN LIVER AND INFLAMMATORY BOWEL DISEASES

Tesi di Dottorato di:

PAOLO NANNI

Relatore:

Chiar.mo Prof. ALDO RODA

Coordinatore:

Chiar.mo Prof. MARIA PIA FANTINI

Esame finale anno 2008

CONTENTS

Abbreviations

1D/2D-PAGE	Mono/bidimensional electrophoresis
CD	Crohn's disease
CID	Collision induced dissociation
CIR	Cirrhosis
DIGE	Differential in gel electrophoresis
ESI	Electrospray ionization
HPLC	High-performance liquid chromatography
H	Healthy
HCC	Hepatocellular carcinoma
HV	Hepatitis virus
HBV	Hepatitis virus B
HCV	Hepatitis virus C
IBD	Inflammatory bowel disease
IC	Inflammatory controls
ICAT	Isotope coded affinity tags
IDA	Imminodiacetic acid
IECs	Intestinal epithelial cells
IMAC	Metal-ion affinity chromatography
iTRAQ	Isobaric tags for quantitative proteomica
LC	Liquid Chromatography
LDA	Linear discriminant analysis
LMW	Low molecular weight
m/z	Mass to charge ratio
MALDI	Matrix assisted laser desorption/ionization
MS	Mass spectrometry
MudPIT	Multidimensional protein identification technology
MWCO	Molecular weight cut-off
PCA	Principal component analysis
PIQS	Parent ion quantification scanning
Q	Quadrupole
RP	Reverse phase
SAX	Strong anion exchange

SELDI	Surface enhanced laser desorpion/ionization
TOF	Time-of-Flight
UC	Ulcerative colitis

Introduction

Proteomics studies represent an important tool for the characterization of the phenotype of a population of cells, a serum sample, a tissue sample and for the identification of specific biomarkers of a given disease. For this reason proteomics has emerged as one of the most important topics in all of bioscience and is a rapidly growing multidisciplinary field that combines separation techniques, mass spectrometry and bioinformatic approaches. Many factors are involved in this increasing interest, first of all the huge potential of clinical proteomic research for the discovery of diseases biomarkers, molecules whose concentrations in the biological fluids or tissues can help in the early diagnosis or therapy monitoring of pathologies, thanks to the possibility of the simultaneous analysis of all proteins expressed in cells, biological fluids or tissues in specific physiological conditions. Furthermore in the last years proteins became central in biological studies after the complete deciphering of human genoma. There was an important shift in the molecular biology dogma due the evidence of the non-direct correlation between genome sequence and protein function, well summarized in a recent editorial: “proteins, not genes, are the business end of biology” [1].

Protein analysis and identification is greatly improved in the last years but methodological problems still limited the identification of proteins and peptides in complex biological matrices. Several gel-free proteomics and labelling techniques have appeared as important alternative to conventional bidimensional gel electrophoresis and many technological advancements have recently occurred, most of them due to the introduction of new mass spectrometers, but to better characterize samples in term of protein expression and functional proteomics studies novel proteomics approaches still need to be developed.

In this thesis work the development of two mass spectrometry-based protein profiling strategies and their application to biomarker discovery in liver and inflammatory bowel diseases is described and discussed.

Chapter 1

1.1 The complexity of biological samples

Biological samples such as cells, saliva, urine and serum are very complex and contain a high number of proteins which concentration ranges of many order of magnitude. Furthermore the biological samples are dynamic because of the enzymatic activity, the post-translational modifications and protein-protein interactions.

An example of the complexity of the interactions among proteins and of the way they work together forming multi complexes to carry out specific functions is shown in figure 1.1. Figure 1.2 shows the distribution of serum proteins among highly-abundant proteins (HAP), moderately-abundant proteins (MAP) and low abundant proteins (LAP): few HAP represent the 96% of the total amount of serum proteins, while LAP, that are often the potential biomarkers of pathological states, are less than 1%.

For these reasons always more sophisticated analytical tools are needed for an in-depth analysis of biological samples proteome. In particular proteomics challenges for robust, automated, and sensitive high-throughput technologies able to resolve complex biological samples using the combination of different separation techniques.

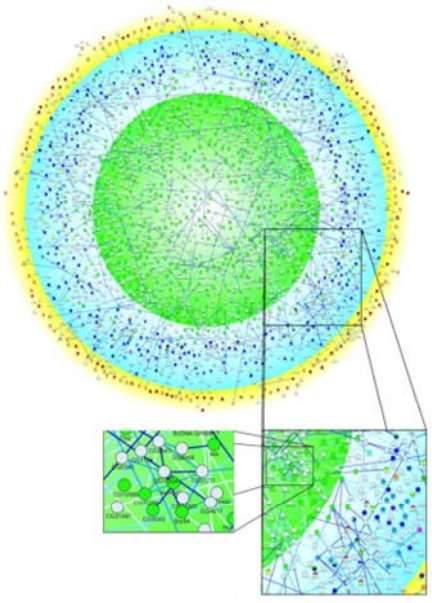

Figure 1.1: example of protein-protein interaction network. Image from Giot, Rothberg *et al*. [2]

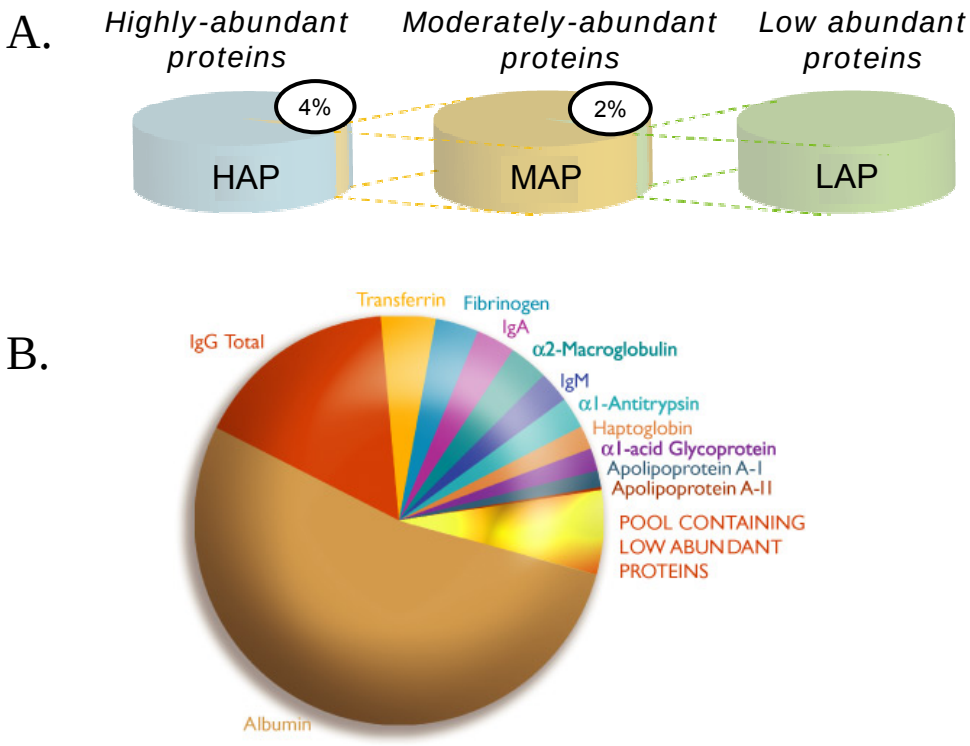

Figure 1.2: Part A. Highly-abundant (HAP), moderately-abundant (MAP) and low-abundant (LAP) proteins in human plasma. Relative abundance of HAP, MAP and LAP fractions is shown in percentage of the bulk mass of the proteins from each fraction to the bulk weight of the total plasma protein. Part B: list of HAP. Image from www.genway.com

1.2 Protein profiling strategies

The ability of protein biomarkers to give indications of physiological states or change make them an important diagnostic and predictive tool in many clinical settings.

Since disease processes involve very complex interactions of large numbers of proteins, there is a considerable interest in the technologies and data analysis techniques specially designed to handle this level of complexity, making it possible to study the entire complement of proteins, the “proteome”, of a blood or tissue sample. The simultaneous analysis of many proteins in a single sample may reveal patterns in their presence, abundance and modifications that result in a “protein signature” associated with the presence or absence of disease at a stage when it is otherwise undetectable. Such a signature, involving dozens or even hundreds of markers, is more likely to provide accurate predictive or diagnostic information than

a single marker. The same techniques may also be used to analyse a complex disease process and identify key molecules that could be targets for drug development.

Two main strategies are used for protein profiling studies, gel-based and gel-free proteomics, which differ on the use of mono-bidimensional gel electrophoresis as separation technique.

1.2.1 Gel-based protein profiling

In the gel-based approach the proteins are separated by bidimensional gel electrophoresis (2DE), an established technique since the late 1970s [3], according to their isoelectric point (pI) in the first dimension and to the molecular mass weight in the second dimension. After the separation the proteins maps obtained from the different samples can be visualized and compared using appropriate bioinformatic tools, and the differently expressed proteins can be excised, in-gel digested and identified by mass spectrometry.

Bidimensional electrophoresis is a technique widely used because of its potentiality to separate thousands of proteins, but is time consuming and needs expert and qualified personal to be performed. However, thanks to many improvements in the technique such as Differential In Gel Electrophoresis (DIGE), the feasibility, the reproducibility and the sensitivity of the experiments have been drastically increased and allowed the use of this approach for very in-depth studies of protein expression in different biological states.

1.2.2 Gel-free protein profiling

The gel-free protein profiling approaches are based on the study by mass spectrometry of the fingerprint (or protein pattern) of proteins characteristic of a specific sample or biological state.

A typical mass spectrometry protein profiling approach can be performed by enrichment of proteins according to their physical/chemical features followed by MALDI-TOF analysis. The comparison among spectra, one of the critical points of this approach, allows the evaluation of the proteins differentially expressed which can be identified by MALDI-TOF-TOF analysis or by isolation of the proteins of interest followed by tryptic digestion and MS analysis. The surface-enhanced laser desorption/ionization (SELDI) TOF MS technology, an extension of MALDI MS, can be also used for these purposes. This technique differs from MALDI-TOF MS

because of its sample target, which is derivatized with different planar chromatographic chemistries, the design of the analyser and the software tools used to interpret the data. On the sample target the proteins actively interact with the chromatographic array surface, and become sequestered according to their surface interaction potential as well as separated from salts and other sample contaminants by subsequent on-spot washing with appropriate buffer solutions.

Other gel-free protein profiling approaches are based on multidimensional separation of peptides from enzimatically digested proteins by bi/tri-dimensional liquid chromatography (LC) or by the combination of monodimensional gel and LC analysis, both of them followed by ESI-QTOF analysis. The experiment can be performed both in the MS/MS or MS operational mode. The former modality, the so-called “shot-gun” proteomics, was initiated with the introduction of the Multidimensional Protein Identification Technology (MudPIT) [4] and can yield to the identification of hundreds of proteins.

The latter modality allows the evaluation of the differentially expressed proteins among the proteins from different biological states and their identification by a following a MS/MS analysis. This comparative studies can be carried out using label-free approaches or isobaric-isotopic derivatizing strategies such as iTRAQ [5] or ICAT [6,7] that permit the simultaneous analysis of the different samples.

1.3 Mass spectrometry soft ionization techniques and analysers

Mass spectrometry (MS) is an analytical technique that measures the molecular weight of molecules based upon the motion of a charged particle in an electric or magnetic field.

The analytes are ionized in the gas phase and separated according to their mass/charge ratio (m/z). Due to its ability to acquire high-content of information, mass spectrometry has emerged among the proteomics techniques as the method of choice for analysing the study of the composition, regulation and function of protein complexes in biological systems [8-10].

Mass spectrometers are used either to measure simply the molecular mass of a polypeptide or to determine additional structural features including the amino acid sequence or the site of attachment and type of posttranslational modifications [11].
A mass spectrometer can be split into three main parts: the ionization source where the gas phase ions are produced from sample molecules, the mass analyser where the ion separation occurs and the mass detector where the signal is recorded.

1.3.1 Soft ionization techniques

For a long time mass spectrometry was restricted to small and thermostable compounds because of the lack of effective techniques to softly ionize and transfer the ionized molecules from the condensed phase into the gas phase without excessive fragmentation. Only in the late 1980s the situation evolved with the development of two techniques for the formation of molecular ions of intact biomolecules: electrospray ionization (ESI) [12] and matrix assisted laser desorption/ionization (MALDI) [13]. The proteins became accessible to mass-spectrometric analysis.

1.3.1.1 MALDI

MALDI is a pulsed ionization technique which utilises the energy from a laser to desorb and ionize the analyte molecules that are co-crystallised on a target with an excess of small organic molecules which absorbs light at the wavelength of the laser (matrix)[13].
The crystals are bombarded with a UV-laser with a typical wavelength of 337nm and the matrix adsorbs most of the energy transferring only a little amount of it to the analytes which sublimates into the gas phase and ionizes by protonation (positively charged ions) or by deprotonation (negatively charged ions).
Many different matrices have been adopted for bio-mass spectrometry, but the most important are α-cyano, 4-hydroxy cinnamic acid (CHCA) for peptides and sinapinic acid (SA) for proteins.
MALDI produces mostly single charged ions that are generated into high-vacuum ($5x10^{-6}$) and accelerated into the mass analyser, typically a Time-of-Flight (TOF).

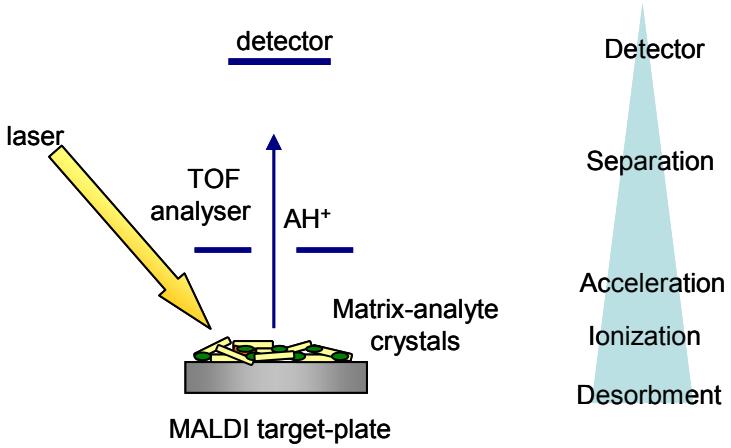

Figure 1.3: schematic of the MALDI process

1.3.1.2 ESI

Electrospray ionization is a technique that uses an electric field to yield a spray of fine droplets. A diluted solution of the sample is sprayed from a narrow capillary tube which carries a high potential (about 4–5 kV). If the needle carries a positive potential, the droplets will have an excess of positive charges, usually protons. Evaporation of the volatile solvent (i.e. H_2O, CH_3OH, or CH_3CN) results in increased Coulombic repulsion between the positive charges, which causes fragmentation of the droplet, generating smaller droplets. This process continues till nanometer sized droplets are produced.

The charges are statistically distributed over the analyte's potential charge sites, enabling the formation of multiply charged ions. Each multiply charged ion can be termed "charge state", and the distribution of charge states is characteristic of large macromolecules during ESI analysis [14].

This feature of ESI to produce predominantly multiply charged ions (opposed to singly charged ions produced by MALDI) yields a reduced mass for the analytes of interest. Figure 1.4 shows the comparison among MALDI and ESI spectra of Cytochrome C. From a MALDI spectra its possible to easily obtain information about the molecular weight of the proteins, while from ESI spectra a deconvolution need to be performed. Furthermore for complex biological samples a separation step prior to mass spectrometry analysis is necessary.

Since ESI ionizes the analytes out of a solution, it can be easily coupled to liquid chromatography and capillary electrophoresis, but for analysis of low abundant samples a low flow rate is needed (sub-micolitre). The use of a nanospray ionization

source with a flow rates in the order of nanolitres per minute increased drastically the sensitivity [15].

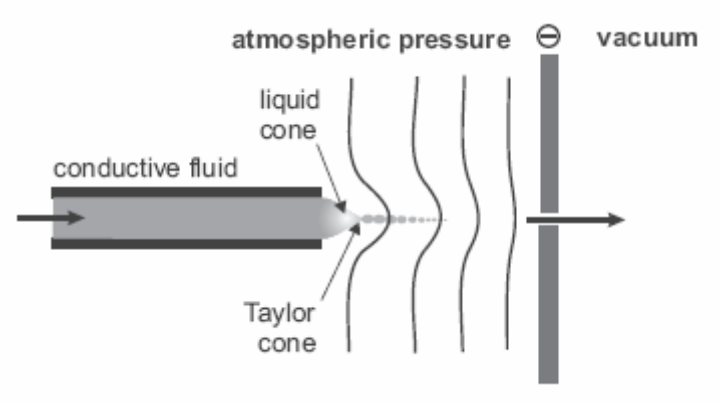

Figure 1.4: schematic of electrospray ionization.

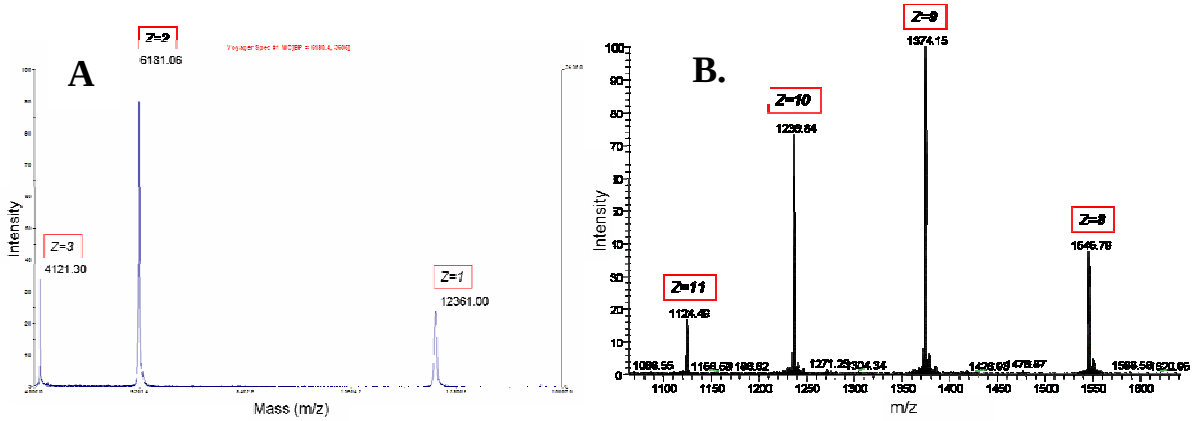

Figure 1.5: example of Cytochrome C spectrum after MALDI (a) and ESI (b) ionization. Differently charged ions are generated.

1.3.2 Mass analysers

The mass analyser is the region of the mass spectrometer where the ions are separated according to their mass to charge ratio (m/z), and can be considered as the central part of the technology. The main four different analysers that have been developed for proteomics applications are the Time-of-Flight (TOF), the ion-trap, the quadrupole and the Fourier Transform ion cyclotron (FT-MS). All of them differs for sensitivity, resolution and mass accuracy. In order to improve the analysis of complex proteins samples many different multistage combinations of these analysers have been developed, such as hybrid quadrupole - Time-of-Flight (QqTOF) and tandem Time-of-Flight (TOF-TOF) [16,17].

1.3.2.1 QUADRUPOLE

The quadrupole analyser consists of four parallel hyperbolic rods through which the gas phase ions have to achieve a stable trajectory [18]. A stable voltage and an oscillating voltage (radio frequency, rf) is applied to one pair of rods and an opposite polarity voltage and rf of different phase is applied to the opposite pair of rods. For every specific voltage the created electric field allows only to ions with a certain m/z value to cross the analyser and reach the detector. The quadrupole voltage can be set to allow a wide mass range to be observed or a single ions to be selected.

Few advantages can be observed in proteomics study if the analyser is constituted of a single quadrupole, but the coupling of three quadrupoles analysers in sequence or two quadrupoles followed by a TOF can yield to important structural information (see tandem and hybrid analysers section).

1.3.2.2 TIME-OF-FLIGHT (TOF)

TOF analyser separates the ions that have been produced in the ion source and accelerated by high voltage to a final velocity dependant on their mass. Since the accelerating voltage is constant, the mass measurement can be recorded by the Time-of-Flight of an ion into the TOF tube, that is proportional to the square root of its mass to charge ratio.

$$\text{Time-of-Flight} = k\sqrt{m / z}$$

The linear TOF analyser, the simplest one, can claim high sensitivity and a very broad mass range. However also if the resolution increases with longer flight tubes it's usually very poor and can hardly discriminate among isotopes. This kind of analyser is for this reason used only for high molecular weight proteins (more than 10kDa), while for peptides the reflectron TOF analyser is preferred.

A reflectron acts as an ion mirror reversing the trajectory of the ions back into the flight tube and for this reason increasing the length of the flight tube [19]. Moreover the reflectron focuses the ions that have a different kinetic energy due to their position in the source during the application of the accelerating voltage. The full width half maximum (FWHM) mass resolution obtained is usually more than 5000.

1.3.2.3 TANDEM AND HYBRID ANALYSERS

Different combinations of mass analysers have been developed. The most important for proteomic studies, schematized in figure 1.6, are the triple quadrupole (Q-q-Q), the quadrupole – ion trap (Q-Trap), the quadrupole – Time-of-Flight (Q-TOF) and the Time-of-Flight – Time-of-Flight (TOF-TOF) because they allow a wide range of research strategies with their capability in collecting useful information about protein identity and modifications.

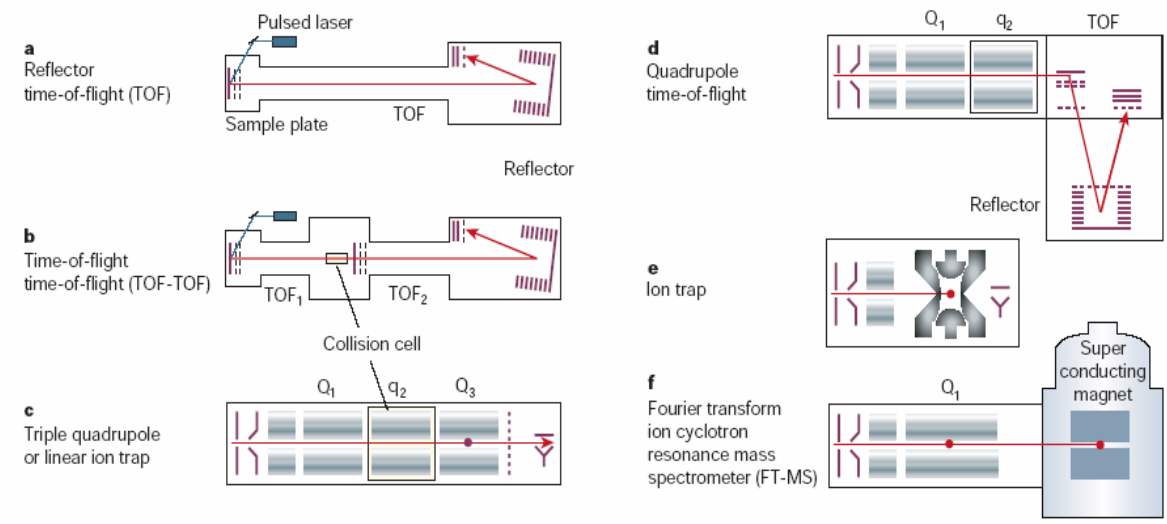

Figure 1.6: Different mass analysers configurations. Image obtained by Aebersold and Mann 2003 [17].

The introduction of these multiple analysers permits the use of two operation modes of a mass spectrometer: the MS mode where molecular weight are measured, and the MS/MS mode where the analyte of interest can be mass measured, selected and fragmented in the mass spectrometer by collision induced dissociation (CID) generating important structural information. MS/MS experiments using triple quadrupole (Q-q-Q), quadrupole – Time-of-Flight (Q-TOF) and Time-of-Flight – Time-of-Flight (TOF-TOF) analysers are widely performed in proteomics.

All of these analysers can be operated in four different ways to perform MS/MS experiments: product ion scanning, precursor ion scanning, neutral loss scanning and multiple ion monitoring (see figure 1.7).

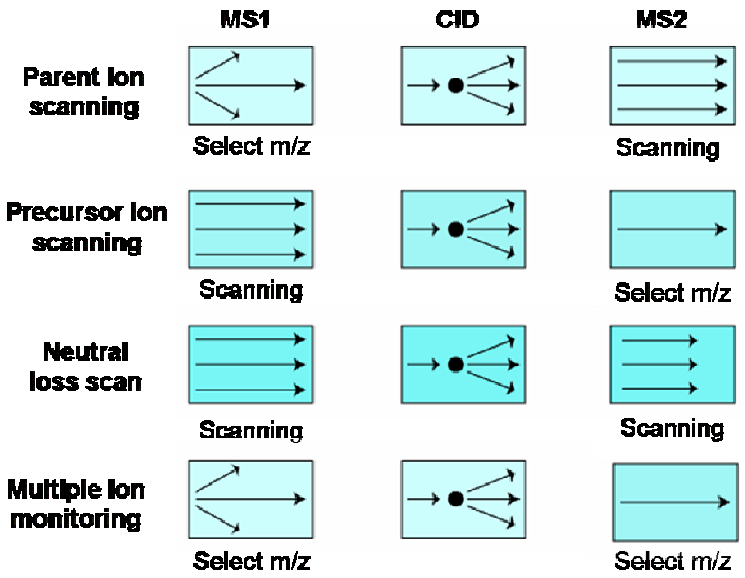

Figure 1.7: schematic representation of the four operation modes to perform MS/MS analysis

In the parent ion scanning the first analyser (MS1) selects a specific m/z (precursor ion) at a time and after CID in the collision cell the resulting fragments are analysed in the second analyser (MS2). This operation mode is particularly useful in proteomic for the identification of the aminoacidic sequence of specific peptides.

The precursor ion scanning operation mode scans the ions in the MS1 and after CID the MS2 selects only specific fragments. The signal is recorded only when the desired m/z fragment reaches the detector. This method is useful for the detection of peptides with functionalised groups such as phosphate esters or carbohydrate modification, since they produce specific ions after fragmentation.

In the neutral loss scanning method both the analysers are synchronised to allow only specific mass differences, a neutral fragment lost from the peptides in the collision cell, to pass through MS1 and MS2. The detection of peptides with specific functional groups can be performed, such as phosphorylation at serine residues via the loss of phosphoric acid.

The last operation mode, the multiple ion monitoring (MRM), is used for the detection of specific fragments in MS1 and of specific peptides in the MS2. This method is particularly useful for the detection of known analytes with known fragmentation patterns present into a complex mixture.

1.4 Mass spectrometry analysis of complex protein/peptides mixtures

Several strategies have been developed for mass-spectrometry based protein analysis that can all be summarized in three main steps: (I) isolation of protein from the biological sample, fractionation (optional), enzymatic digestion and further fractionation of the resulting peptides (very common); (II) qualitative and quantitative analysis of the peptide mixture by mass spectrometry; (III) bioinformatic analysis of the large data set generated by MS/MS analysis, aimed at the identification, and potentially quantification, of the proteins in the sample. The identification is performed by database searching according to specific guidelines [20,21].

Figure 1.8 shows the detailed workflow of the four main strategies used in MS-based proteomics. In part A the proteins are separated by 2D gel electrophoresis, digested, analysed by MS and the proteins identified by peptide mass fingerprint (PMF); part B shows the shotgun proteomics approach, where the protein identification and quantification are performed on randomly selected peptides which are subjected to product ion scanning. In part C a LC-MS analysis step followed by differential analysis of the peptide mixture is performed, with the aim to drive the choice of the peptides to identify and quantify in a second step of MS/MS analysis on the differently expressed ones. Finally, in part D of the figure the peptides chosen for the identification and quantification are selected according to hypothesis from previous experiments, and subjected to multiple ion monitoring (MRM) for a better precision.

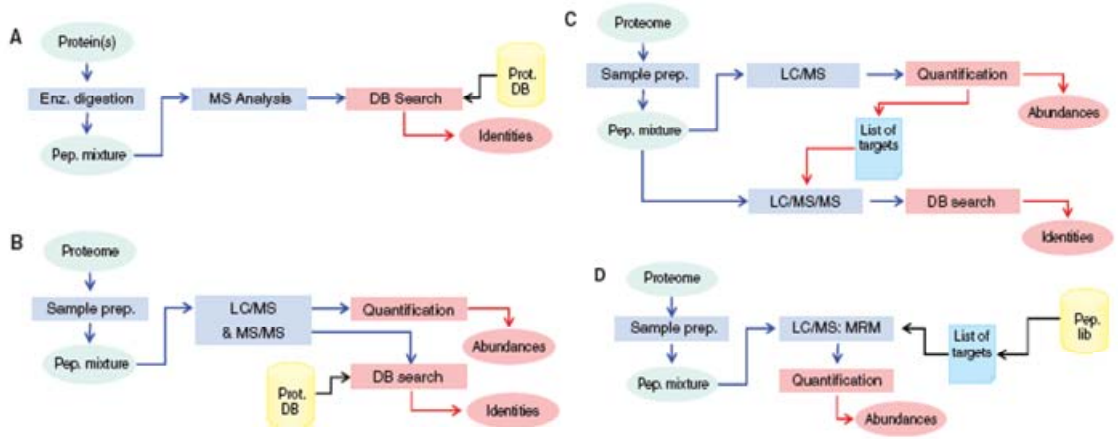

Figure 1.8: mass spectrometry based proteomics workflows. Image adapted from Domon B and Aebersold R [11]

1.5 Chemometric and bioinformatic tools for proteomic data analysis

The development of new analytical strategies, workflows and instruments allows to acquire for every experiment a huge amount of proteomic data that need new statistical and bioinformatic tools to be used and correlated properly. In particular in this thesis work the software PARVUS [22], for the multivariate analysis of MALDI mass spectra, msInspect [23] for the visualization of LC-MS data and the database server PROTEIOS [24] for the conversion of data and the differential analysis of LC-MS runs have been used.

1.5.1 Parvus

In protein profiling studies the comparison of data such as MALDI spectra, the classification of samples and the evaluation of the discriminating peptides/proteins need complex multivariate statistical techniques to be used. PARVUS is a package of programs for explorative data analysis, classification and regression analysis that, among the others, allows to perform the linear discriminant analysis (LDA), the principal components analysis (PCA) and the selection of variables with high disciminant power.

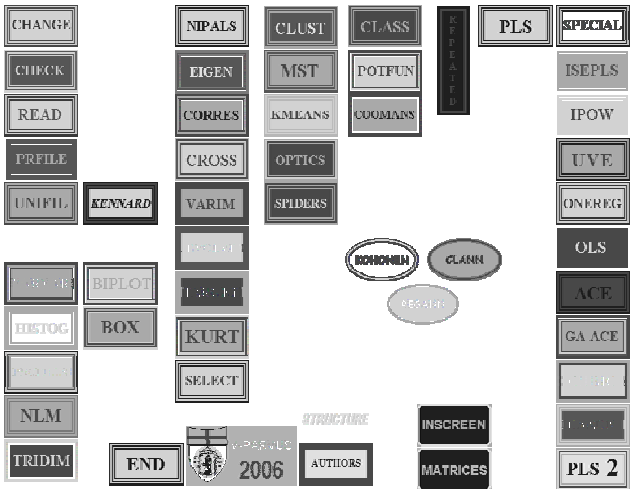

Figure 1.9: screenshot of the PARVUS software, that allows the use of many different multivariate analysis techniques.

1.5.2 MsInspect

Successful application of differential proteomics by liquid chromatography mass spectrometry (LC-MS) requires extraction of peptide features, estimation of peptide abundances, relative quantification, alignment e normalization across multiple related runs, and identification of features.

MsInspect is an open-source software that applies new algorithm for each step of this process, looking for peptide signatures rather than isolated peaks in LC-MS data. It can be used both by command line for batch processing of very large data set or by graphical interface (figure 1.10), thus allowing a better comprehension of the LC-MS data visualization [23].

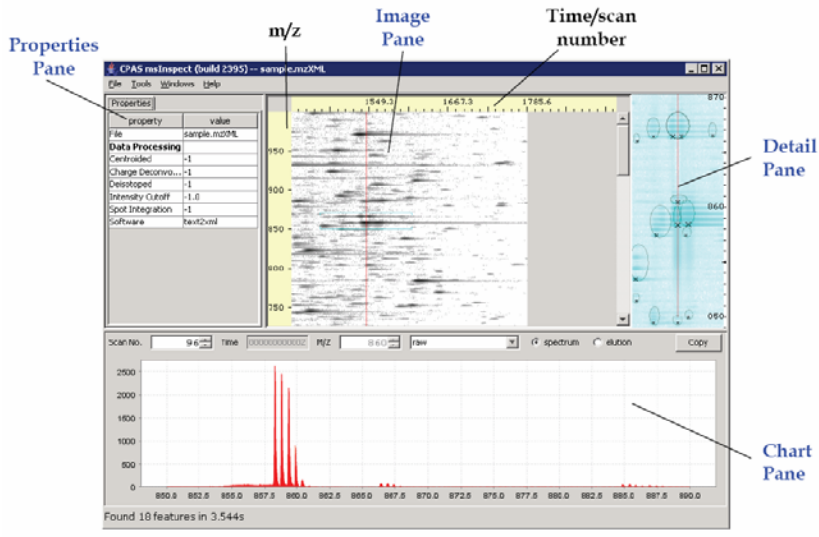

Figure 1.10: the msInspect window has four panes: the *Image Pane* displays an image of the data from MS1 run, the *Detail Pane* displays a zoomed in view of the area in the Image Pane, the *Chart Pane* displays m/z spectra and elution profiles corresponding to the scan and m/z value, the optional *Properties Pane* displays properties of the open mzXML or Feature Set file or the peptide feature selected in the Detail Pane.

1.5.3 Proteios

Proteios is a comprehensive database server to manage and track all the massive amounts of data generated by proteomics experiment and analysis [24]. In particular it can manage the whole workflow from images, raw data and database search results. Furthermore it can be used for data format conversion and statistical analysis on alignment files produces by msInspect software, in order to evaluate the differently expressed peptides in various LC-MS runs.

Figure 1.11: screenshots of PROTEIOS database server interface

1.6 Liver and inflammatory bowel diseases: ethiopathogenesis and existing biomarkers

Mass spectrometry-based protein profiling is a relative young field of analysis that observed a massive increase in the available technologies. Many efforts are made for the discovery of new biomarkers, but frequently the studies become only a proposal of new technology approaches, losing the biological meaning of what is found out. "Bio" must be put back into biomarkers [25].

For this reason in this thesis mass spectrometry based protein profiling strategies have been developed and applied to biomarker discovery in two important gastrointestinal diseases: liver and inflammatory bowel diseases.

1.6.1 Inflammatory bowel diseases (IBDs)

Ulcerative colitis (UC) and Crohn's disease (CD) are clinico-pathologic constructs that collectively are termed inflammatory bowel diseases (IBDs). Given the non-specific nature of the symptoms that herald disease onset (diarrhea, often bloody, abdominal pain, malaise, fever, weight loss), the diagnosis is usually made by endoscopic, histologic or radiographic testing.

Serum antibody testing represents another diagnostic avenue for distinguishing IBD from non-IBD patients. While individual tests may suffer from low sensitivity or specificity, the combination of multiple antibodies into a summary panel holds promise. IgA antibodies to the outer membrane porin C of E. coli, (anti-OmpC); Antibodies to the bacterial flagellin Cbir1(anti-Cbir1); and Perinuclear anti-neutrophilic cytoplasmic antibodies (pANCA) based on three measures:

autoantibody by ELISA; perinuclear pattern of ANCA by immunofluorescence; and DNAse sensitivity. The clinical relevance of such panel testing, however, is uncertain based on two important features.

The same lack of specificity seen in endoscopic, histologic, and radiographic testing impacts not just the distinction between IBD and non-IBD patients, but also distinguishing ulcerative colitis from Crohn's colitis. The term "indeterminate colitis" refers to the inability to distinguish ulcerative colitis from Crohn's disease even after colectomy [26]. The clinical importance in these patients lies in the consequences of medical, and more importantly, surgical therapy.

Several studies have recently been published using proteomic approaches for serum profiling in IBD. Meuwis *et al*. analysed serum samples from IBD patients and controls with SELDI-TOF mass spectrometry [27]. They found that a proteomic approach had greater sensitivity and specificity than testing for pANCA and ASCA. The markers they found, however, seemed to correlate best with activity of disease. Proteomic strategies have also been applied to isolated intestinal epithelial cells from IBD mucosa [28].

1.6.2 Liver diseases

Liver diseases consist in at least two classes: one in which the hepatocytes are damaged (viral, alcholic, autoimmune…) and one in which the bile flow is interrupted (colestatic diseases).

The developed work is focused on the hepatocellular damage and in particular in the viral infection of the hepatocytes. Viral infections of the liver are very common diseases and can be caused by different virus such as hepathytis C virus (HCV), B (HBV), A (HAV).

HAV is more common under 25 years age and is usually an acute infection without any consequences if treated correctly, whereas HCV and HBV are more common after the first quarter of life. These infections can be acute or chronic if persist for more than 6 months.

Different antiviral treatments exists for those infections but the main issue is that if the infection of the liver persists the cells will be totally damaged, the architecture of the liver will be completely altered and the tissue will be replaced with fibrosis. Therefore the infection can cause cirrhosis (CIR) of the liver and the cirrhosis itself is a way to the development of the hepatocellular carcinoma (HCC).

The HCV virus may cause a chronic infection in 60-70% of the people who get infected, while the HBV virus may cause a chronic infection in 1%. Thus, the development of cirrhosis and consequently the hepatocellular carcinoma is more frequent in HCV patients.

Diagnosis for the infection is supported by the detection of viral markers in the blood, while the ultra-sounds are the gold standard for the detection of cirrhosis and the development of the carcinoma. The damage of the liver is also evident in a blood routine analysis (coagulation, cholesterol, bilirubin, platelets….). The prediction of the switch from chronic infection to cirrhosis, and must of all the development of the hepatocellular carcinoma, is often a challenge.

The characteristic tumoral marker for HCC is α-fetoprotein (AFP), but its concentration is very variable and for this reason both sensitivity and specificity are very low. For this reason new markers have been proposed, such as des-gamma-carboxy prothrombin (DCP). The combination of these two markers increases a lot both specificity and sensitivity, but new biomarkers are still needed. Many protein profiling studies for biomarker discovery, well reviewed by Chignard N. and Beretta L. [29], have been performed both on serum and cellular samples, allowing the identification of different signalling pathways in liver carcinogenesis and providing source of novel molecular targets for new therapies [30].

Chapter 2

Serum protein profiling by solid-phase bulk extraction MALDI-TOF MS and chemometric data analysis

Serum is a complicated matrix containing more than 20000 proteins with a wide range of concentrations exceeding even 10 decades [31]. Furthermore the presence of high abundant proteins such as albumin, masks the low abundant proteins which usually are of diagnostic utility since they change their expression level and regulation in relation to a given disease.

Despite the powerful of the new soft impact mass spectrometry (MS) technologies such as matrix-assisted laser desorption/ionization – Time-of-Flight (MALDI-TOF/TOF) and electrospray ionization quadrupole Time-of-Flight (ESI-QTOF), the main problem in clinical proteomics is the difficulty to efficiently isolate the proteins of interest from the matrix to produce a "clean" sample suitable for MS analysis. Differences in the protein expression patterns between diseased and normal samples is generally carried out using two dimensional polyacrylamide gel electrophoresis (2D-PAGE). After gel staining the protein(s) of interest are removed, proteolytically or chemically digested and identified by MS analysis and proteomic databases search. The need for protein staining often limits the sensitivity of the overall approach. The development of two-dimensional nano-liquid chromatography (nano-LC) coupled with ESI-QTOF MS greatly improved the analysis of proteins at low concentration levels, but sample preparation and clean-up procedures are still necessary [32].

Methods for sample preparation [33], including size cut-off membrane-based methods like dialysis and ultracentrifugation, selective protein precipitation and solid phase extraction (SPE) have gained importance, but none of them yet fulfil criteria of applicability.

SPE is the most used approach for selective protein extraction both on-line with LC ESI-MS [34,35] and, in particular, off-line with MALDI MS [36,37]. Many different devices for the off-line coupling such as micropipet tips packed with SPE media [38], microtiter plates with integrated SPE [39] and modified MALDI target plates for the purification of the sample directly on-target like Integrated Selective Enrichment target (ISET) [40] are now available.

Recently other approaches based on technologies such as SELDI (surface-enhanced laser desorption/ionization) or MassPREP PROtarget (Waters, Milford, MA) have been developed [41-43].

As alternative approach we developed a simple and rapid top-down analytical proteomic method, based on the selective solid phase extractions of serum proteins in bulk using differently derivatized beads prior to the analysis by MALDI-TOF MS, followed by chemometric analysis for the recognition of a pattern of proteins more relevant for the discrimination between diseased/non-diseased samples and the classification of samples.

The developed method was applied to the study of MALDI-TOF MS serum protein expression profiles of two important types of diseases: liver and inflammatory bowel diseases (IBDs, Crohn's disease and ulcerative colitis). Because of the not specific nature of the symptoms, the differential diagnosis in IBDs cannot be made with the common serological tests and requires endoscopic, histological or radiological examination; also if recently new tests have been introduced none of them alone can be of diagnostic utility [44,45]. Liver diseases are extremely common and there is still a strong demand for specific markers of switch from chronic infection to cirrhosis and hepatocellular carcinoma.

Linear discriminant analysis (LDA), a multivariate classification technique, was here used as classification method. LDA is a probabilistic parametric classification technique based on the estimates of the probability distributions which allows the classification of an object in the class with the higher assignment probability. Since an excessive number of variables prevents the use of LDA, the initial large amount of variables (m/z values), due to the high resolution of TOF analyser, was reduced by features selection. The selection of a pattern of 10-20 relevant m/z values not only allows the application of LDA, but also the direct correlation between the selected m/z and proteins involved in the discrimination between healthy subjects and among the IBD patients, opening the possibility of their identification as biomarkers.

2.1 Introduction

The search for disease protein biomarkers by MALDI MS needs the development of serum sample preparation techniques and the use of statistical analysis methods for their identification and validation.

A simple methodological approach able to evaluate the differences in MALDI-TOF MS serum protein expression profiles among liver and inflammatory bowel diseases was developed.

The developed procedure uses the interaction of proteins with different functionalized beads (SAX, C18, IDA-Cu(II)) added directly to the serum samples. Selective solid-phase bulk extraction, purification and concentration of proteins were simultaneously performed and the suspension of beads-bound proteins was directly analysed by MALDI-TOF MS. Feasibility studies for the application of functionalized carrier materials for direct MALDI-TOF MS serum protein profiling were already performed by other research groups [46-48], but to our knowledge none of them applied it for biomarkers discovery. In this thesis the first application in clinical proteomics of this methodological approach combined to chemometric analysis of the obtained mass spectra is reported.

The use of beads in bulk procedure instead of derivatized surfaces (such as SELDI chips) for the solid phase extraction increases the interaction area between analytes and functionalized groups allowing to extract more proteins with a faster kinetic due to polydispersion of the microsized beads reducing diffusion problems when the solid phase is a surface with a consequent steric hindrance problem.

The elution step was performed by the application of the matrix on the suspension of beads-bound proteins previously spotted directly on the MALDI-target plate, thus preventing the potential loss of proteins/peptides typical of the extraction during the elution step. The whole workflow is shown in figure 2.1.

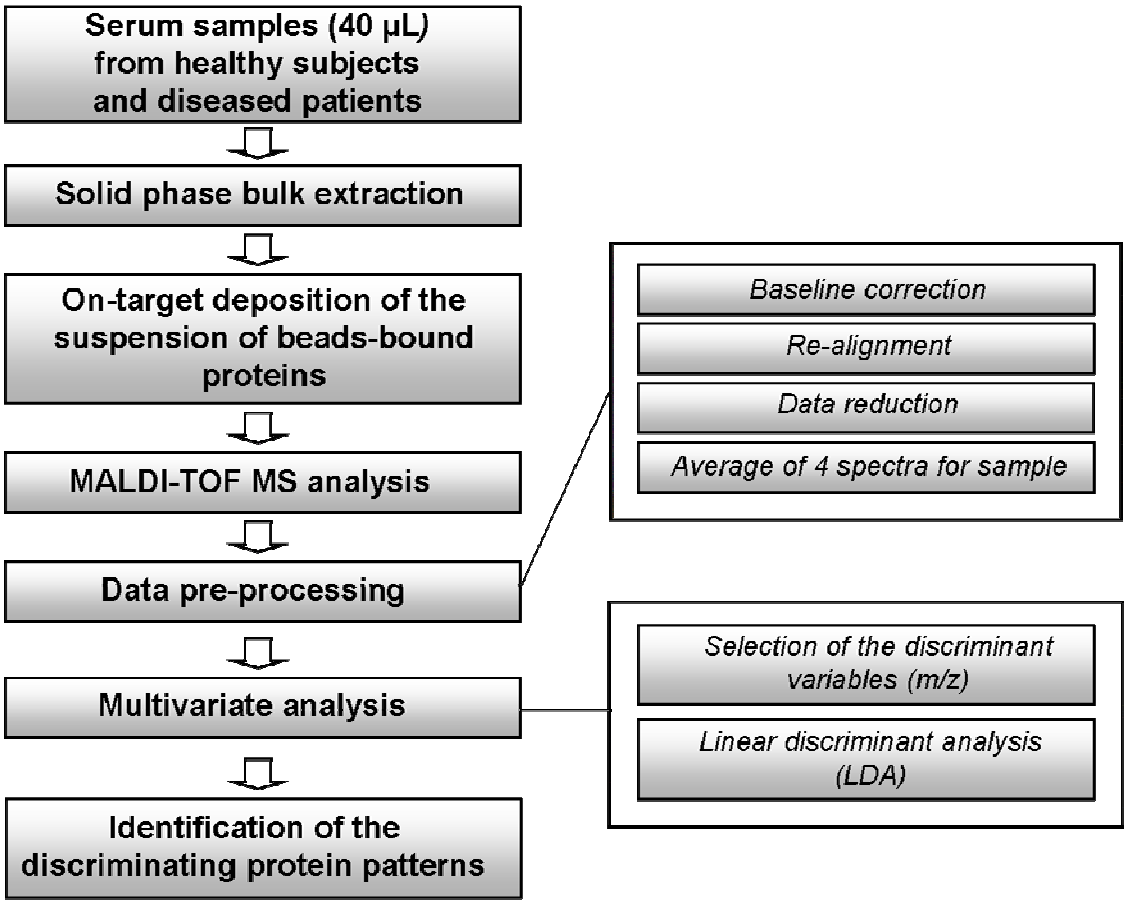

Figure 2.1: workflow of the developed MALDI-TOF MS protein profiling approach.

2.2 Experimental

2.2.1 Materials

Silica C18/Corasil beads (Bondapak® 37-50µm) were purchased from Waters and strong anion-exchange (SAX, 40-50µm) beads were from Varian (Palo Alto, CA). All the chemicals and solvent were purchased from Sigma Aldrich (St. Louis, MO, USA) and used without further purification.

2.2.2 Patients

Blood samples were provided by the Gastroenterology Unit at S.Orsola University Hospital of Bologna (Italy) after obtaining informed consent from the patient.

A total of 63 subjects, including 22 healthy donors volunteers (H; 9 M and 13 F), 15 Crohn's disease (CD; 9 M and 6 F), and 26 ulcerative colitis patients (UC; 13 M and 13 F), were admitted to the preliminary step of the study of the inflammatory bowel diseases while a total 129 subjects were included in the follow-up study (50 H; 20 M

and 30 F; 30 CD; 21 M and 9 F, 40 UC; 20 M and 20 F, 9 non-specific inflammatory controls (IC) such as irritable bowel, celiac disease, arthritis, liver disease; 5 M and 4 F).

A total of 61 subjects, including 17 healthy donors (H; 9 M and 8 F), 19 hepatitis C (HCV; 6 M and 4 F) and hepatitis B (HBV; 6 M and 3 F), 13 cirrhosis of the liver without malignant or suspicious nodules (CIR; 8 M and 5 F) and 12 hepatocellular carcinoma diagnosed during chronic hepatopathy (HCC; 8 M and 4 F) patients were admitted to the study of liver diseases.

All the subjects were ranging in age from 20-65 years. The diagnosis of the IBD was achieved by clinical symptoms, clinical chemistry data and conventional endoscope procedures and patients with either active and inactive disease were included. The clinical diagnosis of the disease was assessed according to the CDAI (Crohn's disease activity index) and to the Truelove & Witt classification.

The diagnosis of liver diseases was achieved by clinical symptoms, blood analysis and ultrasound examination.

From each subject five millilitres of blood sample were collected and centrifuged at 4000 rpm for 10 min. The obtained serum samples were distributed into aliquots and stored frozen in plastic vials at -80°C until use.

2.2.3 Sample preparation

All the serum samples were subjected to solid phase extractions. In the study of IBD three different SPE were used: reverse phase (silica C18), anion exchange (SAX) and immobilized metal ion affinity-Cu(II) (IMAC-Cu(II)). In the liver diseases study only IMAC-Cu(II) extraction was carried out. The SPE was performed directly in 500 µL assay tube using 5 mg of derivatized beads. Each sample was analysed in duplicate.

Reverse phase extraction

Silica C18 beads were washed two times with 150 µL acetonitrile (ACN), mixed for 2 minutes and centrifuged at 13.400 rpm for 1 min. The surnatant was removed and the residue was equilibrated twice using 150 µL 0.1% trifluoroacetic acid (TFA), then 40 µL of serum sample was added.

After 10 min of room temperature incubation two washing steps with 150 µL 0.1%TFA for the removal of unbound proteins were performed and a 10 µL suspension of beads-bound proteins residue in the washing solution was obtained.

Anionic exchange extraction

Strong anion exchange beads were washed two times with 150 µL methanol (MeOH), mixed for 2 minutes and centrifuged at 13.400 rpm for 1 min. The surnatant was removed and the residue was equilibrated twice using 150 µL 10 mM ammonium bicarbonate buffer (pH 7.7) then 40 µL of serum sample was added. After 10 min of room temperature incubation two washing steps with 150 µL 10 mM ammonium bicarbonate buffer (pH 7.7) for the removal of unbound proteins were performed and a 10 µL suspension of beads-bound proteins residue in the washing solution was obtained.

Metal ion affinity - Cu(II) extraction

For immobilized metal ion affinity chromatography (IMAC), Cu(II) was loaded on derivatized cellulose particles. Both the procedure for derivatization of cellulose with iminodiacetic acid (IDA) and the saturation with Cu(II) ions were previously described by Feuerstein *et al* [49].

The beads were washed two times with 150 µL 50 mM sodium acetate buffer pH 4.0, mixed for 2 minutes and centrifuged at 13.400 rpm for 1 min. The surnatant was removed and the residue was equilibrated twice using 150 µl phosphate buffered saline buffer (PBS) pH 7.4.

At the same time 40 µl of serum sample was treated with 30 µl 8 M urea containing 1% 3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate (CHAPS) in PBS by mixing for 2 min. Afterwards 100 µl 1 M urea containing 0.125% CHAPS was added, the mixture was diluted to 850 µl with PBS and vortexed at 4 °C for 10 min. Four hundred microlitres of the diluted serum sample were added after the equilibration step and the whole suspension mixture was incubated on a platform shaker at 1500 rpm for 2 h at 30 °C. To remove unbound proteins two washing steps with 150 µl equilibration buffer, followed by a quick washing step with 200 µl deionized water were performed. A 10 µL suspension of beads-bound proteins residue in deionized water was obtained.

2.2.5 Mass spectrometry

One microlitre of each suspension of beads-bound proteins was directly applied onto a standard 100-spot stainless steel MALDI-target plate, mixed with 1 µL of sinapinic acid matrix (30 mg/mL, 50% ACN/0.1% TFA) and analysed. It is important to add the matrix solution before the sample is air-dried.

Proteins bound to the derivatized beads were analysed by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) using an Applied Biosystems Voyager-DE PRO mass spectrometer equipped with a 337-nm nitrogen laser. Analysis were performed in linear positive-ion mode, using delayed extraction. The acceleration voltage was 25 kV, guide wire was 0.15 of the accelerating voltage, grid voltage was 93%, and the delay time was 200 ns. Four spectra for each sample were collected manually in the m/z range 2000-10000 averaging 120 laser shots.

External mass calibration was performed using the calibration mixture 2 of the Sequazyme peptide mass standard kit (Applied Biosystems, Darmstadt, Germany), containing bovine insuline (3 pmol/μL), ACTH (1.5 pmol/μL) and angiotensin I (2 pmol/μL).

Intra-assay repeatability and inter-assay reproducibility studies were performed using healthy donor serum samples. For the repeatability study, solid phase extractions were performed in duplicate and four MALDI spectra were acquired for each replicate. The reproducibility of the procedure was assessed by analysing the same samples in different days.

The raw spectra were processed with the Voyager Data Explorer software (version 4.0.0.0, Applied Biosystems).

2.2.5 Data pre-processing

Before statistical analysis data pre-processing steps were needed. Beside the external calibration a re-alignment process was done. A healthy donor serum sample, considered as quality control, was internally calibrated by deposition on MALDI-target plate of 1 μL of the suspension of beads-bound proteins with 1 μL of calibration mixture diluted 1:6 in matrix solution for serum samples extracted by SAX and 1:2 for serum samples extracted by IDA-Cu(II) and C18. Four peaks present in all the spectra obtained from the same SPE were considered as references peaks and used for a re-alignment process (m/z = 4213.15, 5341.26, 7772.62, 9296.50 for SAX; m/z = 3955.76, 4644.01, 6631.03, 8929.64 for C18; m/z = 4283.2, 6432.21, 7764. 67, 8932.41 for IDA-Cu(II)).

No peak detection step was performed on the mass spectra. Data were exported from Voyager Data Explorer software as ASCII text files and m/z values were reduced from more than 16000 to 1600 data points by average every 5 m/z values.

For each sample 8 spectra were collected (4 spectra for sample, each sample in duplicate), 4 were selected according to signal/noise ratio and averaged in order to obtain a single representative spectra for each subject.

2.2.6 Statistical analysis

After pre-processing of MALDI spectra a data matrix for each extraction procedure was obtained. The three final data matrices contain as many rows as samples and as many columns as m/z values obtained after data pre-processing (1600).

The three data sets were imported in the chemometrical software package V-PARVUS [22] and the multivariate analysis was performed.

Normalization of the data by row profile and column centering was applied to correct the drift sensitivity [50].

Linear Discriminant Analysis (LDA) [51] was used as classification method after applying SELECT [52], a features selection technique for the reduction of the original number of variable that permits to retain only the useful variables (m/z).

2.3 Results and discussion

The main goals of profiling studies are the evaluation of protein patterns that distinguish between classes such as diseased/non-diseased, the classification of samples and the identification of the protein biomarkers involved in these patterns [53]. For this comparative analysis of protein profiles spectra reproducible MALDI protein profiling strategies are pivotal, but has been recently demonstrated that also the data pre-processing step and the statistical analysis are critical points that must be improved [54].

In figure 2.2 it's possible to notice the differences among the typical MALDI MS protein profiles obtained by the three serum proteins extraction principles.

In order to evaluate the variability of the MALDI-TOF MS spectra obtained an intra-assay repeatability and inter-assay reproducibility study were performed. Figure 2.3 shows how among a given extraction principle the protein profiles showed a stable pattern of peaks with similar intensity ratio between them, thus allowing the use of this analytical method for MALDI MS protein profiling.

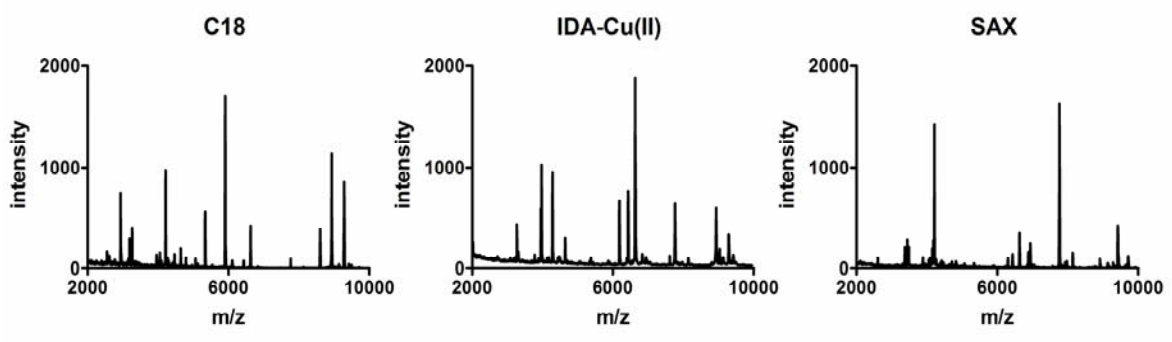

Figure 2.2: example of MALDI-TOF mass spectra obtained after sample preparation using C18, IDA-Cu(II) and SAX solid-phase bulk extractions.

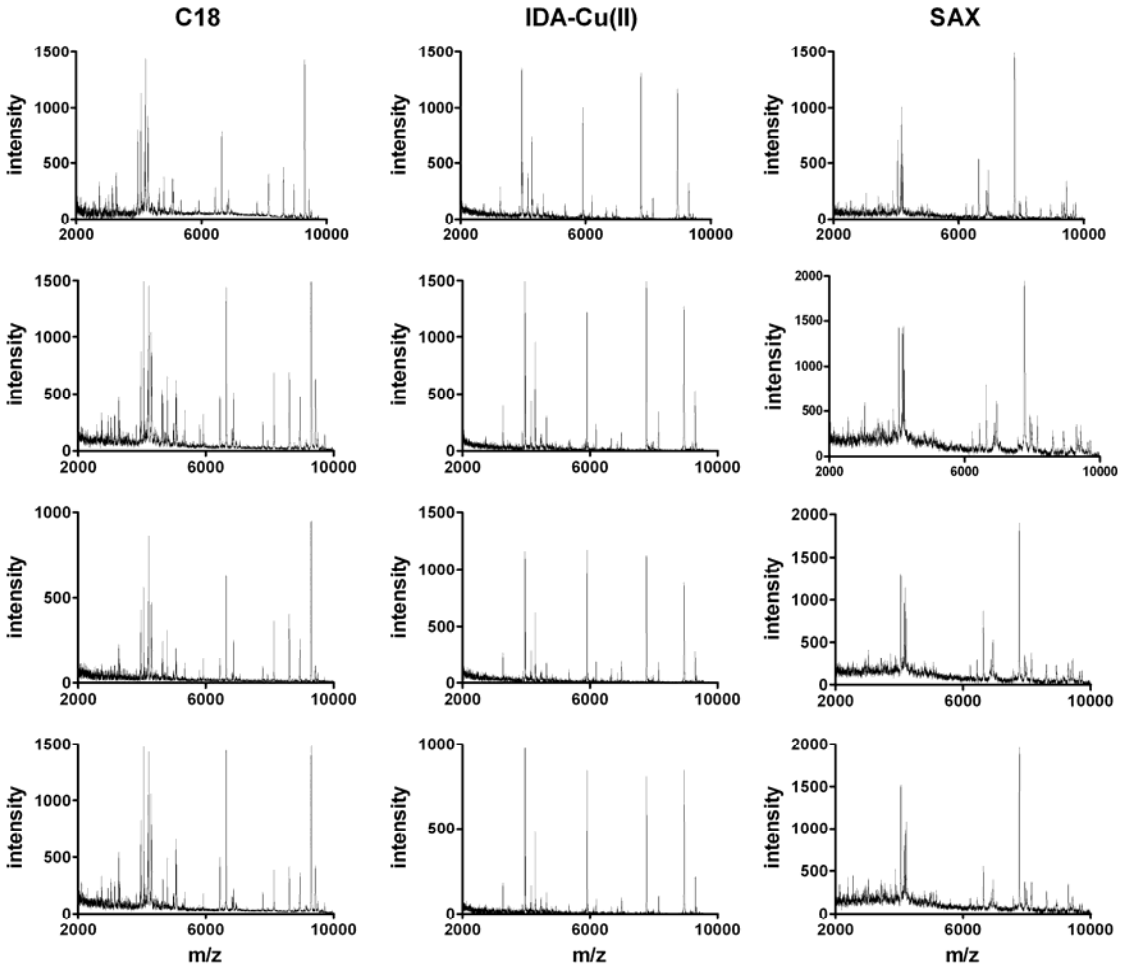

Figure 2.3: reproducibility of serum protein profiles obtained by C18, IDA-Cu(II) and SAX solid-phase bulk extraction combined with MALDI-TOF MS analysis. As described in the Results and Discussion section, each serum sample was extracted in duplicate with every bulk-SPE, spotted twice on the MALDI-target plate and two spectra were acquired for each spot (eight spectra from four spots). The figure shows for each serum sample the four spectra with the higher signal/noise ratio.

A well designed data pre-processing and multivariate analysis step must be performed for the recognition of the proteins evaluated as significant in the classification, otherwise from the same data set different results can be obtained. For example in many recent comparative studies concerning the same tumor/non-tumor protein patterns [55-61] a prediction ability >90% was always obtained, but there was no agreement about the proteins involved in the discrimination.

After baseline subtraction the re-alignment of the peaks was done, since during the acquisition of the data a mass error up to 200 ppm was sometimes obtained despite the external calibration step. A pattern of peptides was evaluated as stable between the spectra obtained from the same extraction principle and for this reason used for the re-alignment. Usually a number of about 16000 m/z values (variables) in the selected m/z range (2000-10000) was obtained for each raw mass spectrum. To each m/z value is associated one intensity value.

In our approach we reduced the variables number to 1600 by averaging every 5 m/z values, since the TOF analyser resolution doesn't allow the discrimination between peaks closer than this value. Using this procedure the data dimension was extremely reduced.

With the aim to decrease the variability due to sample preparation steps and the crystallization of the protein-bound beads with the matrix every sample was analysed in duplicate. Each sample was spotted twice on the MALDI-target plate and two spectra were acquired. A total of eight spectra were obtained. The spectra with the higher signal/noise ratio for each spot were taken into account and the average (arithmetic mean of the intensities for each m/z value) among the obtained four spectra was calculated to reduce the number of data without loss of information. In this way a representative spectrum for each serum sample was obtained facilitating the subsequent multivariate analysis.

LDA was applied to the obtained data set. LDA is a classification method that searches for directions (discriminant functions) with maximum separation among categories. The first discriminant function is the direction of maximum ratio between inter-class and intra-class variances.

In these works classification and prediction abilities were computed for every class (H, CD, UC and IC for the IBDs study and H, HCC, HCV, CIR for the study of liver diseases). When the prediction ability is almost equal to the classification ability validates the classification rule built on the training set. To evaluate the prediction rate the cross-validation procedure was performed with five cancellation groups

(CV): the classification rule was computed five times, each time with the objects of four CV. The classification ability was calculated as the percentage of training set objects correctly classified while the prediction ability was measured on the objects of the left-out group. The number of CV, five, is considered to provide a more realistic, not too optimistic, evaluation of the prediction rate, as usually happens using the leave-one-out validation procedure [62].

Since the high number of variables in the MALDI spectra (more than 16000 m/z values) prevents the use of LDA a variables reduction is needed.

This can be done applying the Principal Component Analysis (PCA) and using the scores as input for LDA [63]. In these studies the variables selection is proposed as an alternative to the use of the significant principal components because it presents some advantages such as the easy identification of a pattern of m/z values with high discriminant power.

Features selection is an important aim of chemometrics: its goal is to discard variables without discriminant information and to maintain only those variables which really improve classification ability. SELECT is a feature selection technique based on the stepwise decorrelation of the variables and it is implemented in the V-PARVUS; it generates a set of decorrelated variables ordered according to their Fisher weights. SELECT searches, at each step, for the variable with the largest classification weight. This variable is selected and decorrelated from the other variables; then the algorithm is repeated until a fixed number of variables is selected. In these works different protein patterns with different numbers of retained variables (10, 15, 20, 25, 30) corresponding to m/z values were used as input for the following LDA.

2.3.1 Inflammatory bowel diseases

The study of inflammatory bowel diseases was performed in two steps. The first step, a preliminary study where only 63 subjects were included, was necessary for the evaluation of the extraction procedure able to obtain the serum protein profiles with the best classification ability. In the second step the work was carried out with the previously selected extraction procedure by analysing serum samples from 129 subjects.

2.3.1.1 PRELIMINARY STUDY

In table 2.1 are reported the variables selected for the three data matrices while table 2.2 shows the LDA results for each data set. It's possible to observe that LDA results improve by using the pattern of 20 variables. In general all the extraction methods gave similar results, but overall we preferred C18 principle both for the prediction ability achieved and for the easy and fast extraction procedure. In particular using this procedure the 100% of samples were correctly classified and the 96.9% of average prediction ability with 5 CV was achieved. The prediction ability was 95.5% for healthy subjects, 100% for Crohn's disease patients and 96.3% for ulcerative colitis patients. A graphical display of C18, SAX and IDA-Cu(II) results, where the objects are projected on the first two canonical variables of LDA, is here reported (figure 2.4). Results show a perfect separation among the three categories (H, CD and UC), thus proving the capability of this methodological approach for the identification of discriminating protein profiles in the detection of IBDs.

SPE	Number of selected variables	Selected variables
C18	10	2555, 3505, 3525, 3745, 3815, 3885, 4635, 4650, 5755, 7610
	15	2555, 2755, 2975, 3505, 3525, 3745, 3815, 3885, 4635, 4650, 4735, 5755, 7610, 8605, 9155
	20	2555, 2615, 2755, 2975, 3505, 3525, 3600, 3745, 3815, 3885, 4635, 4650, 4735, 4740, 5755, 6610, 6850, 7610, 8605, 9155
SAX	10	2275, 3030, 3775, 3995, 4305, 5165, 5270, 5770, 5815, 6830
	15	2275, 3030, 3125, 3775, 3950, 3995, 4305, 4740, 5165, 5270, 5770, 5815, 6830, 8210, 8760
	20	2275, 3030, 3125, 3225, 3775, 3950, 3995, 4000, 4305, 4740, 5165, 5270, 5470, 5475, 5770, 5815, 6830, 8210, 8760, 8935
IDA-Cu(II)	10	2420, 2795, 2935, 4155, 4180, 4215, 4785, 5925, 8960, 9305
	15	2340, 2420, 2440, 2795, 2935, 3940, 4155, 4180, 4190, 4215, 4785, 5925, 6685, 8960, 9305
	20	2205, 2340, 2420, 2440, 2795, 2935, 3260, 3940, 4155, 4180, 4190, 4215, 4785, 4795, 5580, 5925, 6075, 6685, 8960, 9305

Table 2.1: variables (m/z values) selected by SELECT features selection technique.

SPE	Number of selected variables	LDA results							
		Classification ability (%)				Prediction ability (%)			
		H	CD	UC	Average	H	CD	UC	Average
C18	10	94.5	88.8	88.1	**90.5**	90.9	81.5	82.5	**84.6**
	15	98.2	88.8	97.0	**95.4**	90.9	81.5	92.6	**89.2**
	20	100.0	100.0	100.0	**100.0**	95.5	100.0	96.3	**96.9**
SAX	10	95.5	76.0	86.7	**87.2**	95.5	73.3	87.5	**86.9**
	15	97.3	91.0	100.0	**96.7**	95.5	86.7	83.3	**88.6**
	20	98.2	93.3	100.0	**97.7**	86.4	93.3	95.8	**91.8**
IDA-Cu(II)	10	92.7	96.0	99.1	**96.0**	90.9	86.7	82.6	**86.7**
	15	95.5	98.7	100.0	**98.0**	90.9	80.0	95.7	**90.0**
	20	100.0	100.0	100.0	**100.0**	86.4	93.3	95.7	**91.7**

Table 2.2: classification and prediction ability results of LDA.

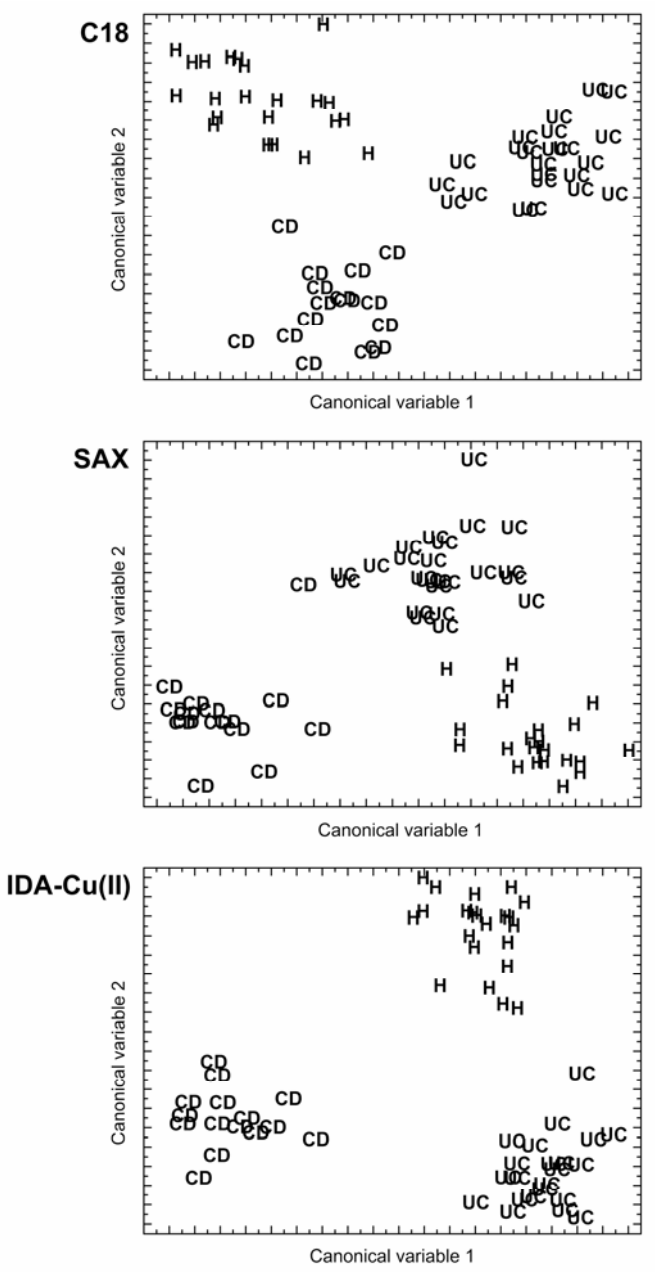

Figure 2.4: graphical display of LDA results obtained from MALDI-TOF MS serum protein profiles of healthy donors and IBD patients. Objects are projected on the first two canonical variables.

2.3.1.2 C18 SOLID-PHASE BULK EXTRACTION

After the selection of C18 solid-phase bulk extraction the follow-up to the study was performed including more serum samples for every category (n° 50 healthy subjects (H), n° 30 Crohn's disease (CD) and n° 40 ulcerative colitis (UC) and an optimum classification and prediction ability was achieved (97.0% and 87.5% respectively, data shown in table 2.3; figure 2.5). For this more representative set of data 30 variables were selected.

Number of selected variables	LDA results							
	Classification ability (%)				Prediction ability (%)			
	H	CD	UC	Average	H	CD	UC	Average
30	97.6	96.7	98.5	**97.0**	86.0	86.7	90.0	**87.5**

Table 2.3: classification and prediction ability values obtained by the analysis of 120 serum samples from H donors, CD and UC patients

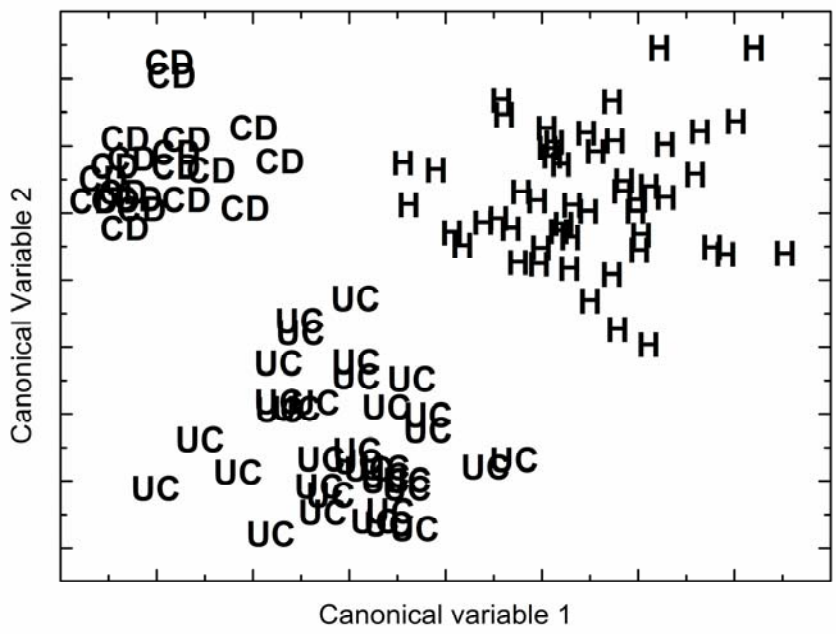

Figure 2.5: graphical display on the first two canonical variables of the results of LDA of 120 serum spectra from healthy subjects and patients affected from IBD (Crohn's disease and ulcerative Colitis, both in the active and inactive phases) treated with C18 solid phase.

The patients included in the study were both in active and inactive disease phases, hence the obtained results are particularly promising for further studies. Indeed performing LDA including only healthy donors and CD and UC patients in the active disease phases (data reported in table 2.4 and figure 2.5) results even better were obtained. This can be explained by the higher number of differences among the protein expression profiles in the active phase.

Number of selected variables	LDA results							
	Classification ability (%)				Prediction ability (%)			
	H	CDact	UCact	Average	H	CDact	UCact	Average
25	100.0	100.0	100.0	**100.0**	92.0	94.7	77.7	**89.0**

Table 2.4: classification and prediction ability values obtained by the analysis of 91 serum samples from H donors and CD and UC patients in the active phases of the disease.

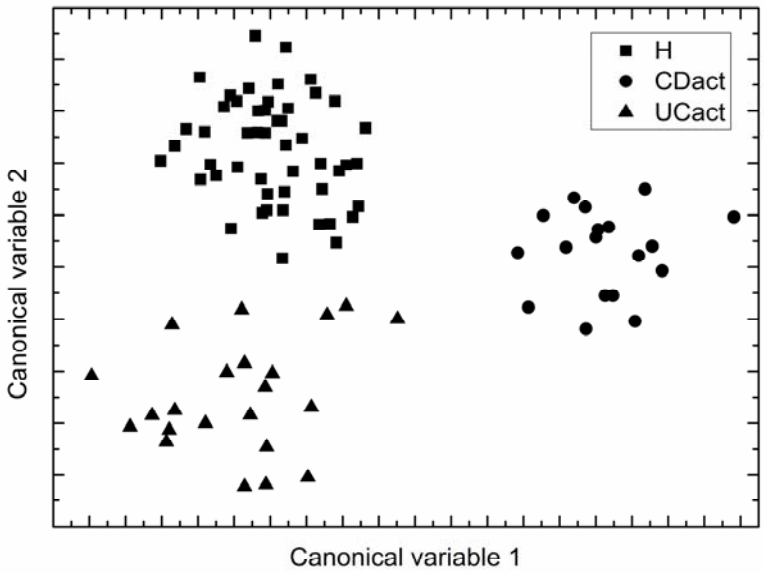

Figure 2.6: graphical display of LDA results obtained from MALDI-TOF MS serum protein profiles of healthy donors (H) and IBD patients in the active phase of disease (Crohn's disease active (CDact) and ulcerative colitis active (UCact)). The objects are projected on the first and second canonical variables.

Afterwards a little number of non-specific inflammatory controls (IC: irritable bowel, celiac disease, arthritis, liver diseases…), analysed in the same lack of time of H, CD and UC samples, has been introduced in the classification model and the classification and prediction ability were calculated (table 2.5).

Figures 2.7A and B show the LDA results on different canonical variable planes and a perfect discrimination of the inflammatory controls from the other categories can be observed. However the inclusion of the inflammatory controls decreases the prediction ability of all the other groups. This can be due to the little number of samples in this category compared to the number of samples in the other ones: one

more category is added but it's not strongly represented. For this reason more inflammatory control serum samples need to be analysed in order to better validate the classification model.

Number of selected variables	LDA results									
	Classification ability (%)					Prediction ability (%)				
	H	CD	UC	IC	Average	HV	CIR	HCC	H	Average
30	94.0	92.7	83	100	**90.7**	76.0	66.7	67.5	100	**72.9**

Table 2.5: classification and prediction ability values obtained by the analysis of 129 samples from H, CD, UC and IC serum samples

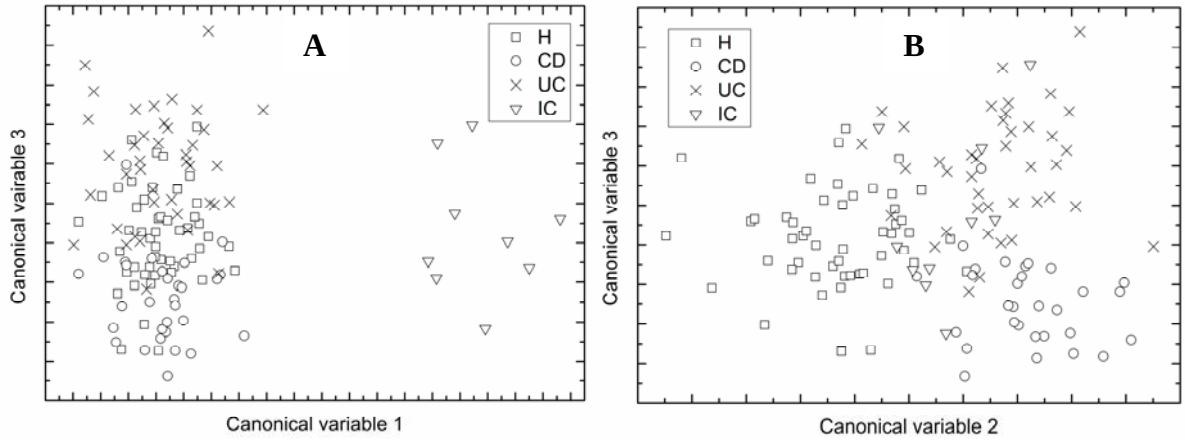

Figure 2.7: graphical display of LDA results obtained from 129 MALDI-TOF MS serum protein profiles of healthy donors (H), IBD patients (Crohn's disease (CD) and ulcerative colitis (UC)) and inflammatory controls (IC). In figure A the objects are projected on the first and third canonical variables, while in figure B on the second and third canonical variables.

An in-depth analysis of the final results was carried out to calculate the sensitivity (true positive), the specificity (true negative) and the accuracy of the method (table 2.6). The comparison of the classification results of IBD samples and controls (H and IC) allowed to obtain a sensitivity and a specificity of 87.1% and 76.0% respectively, thus meaning that the protein expression profiles in IBD and controls samples are very different. Otherwise a lower sensitivity and specificity have been obtained comparing CD and UC samples (both of them about 67%) that can be explained by the higher similarity among the respective protein patterns. Finally the classification results of both active and inactive samples from CD and UC samples

were compared among them: a high accuracy was obtained for CD samples while only a 60% accuracy was calculated for UC. The method seems to be more sensitive and specific for Crohn's disease than for ulcerative colitis.

C18	Sensitivity (true pos)		Specificity (true neg)		Accuracy	
IBD vs H	61 / 70	**87.1%**	38 / 50	**76.0%**	99 / 120	**82.5%**
IBD vs all control	61 / 70	**87.1%**	47 / 59	**79.7%**	108 / 129	**83.7%**
CD vs UC	20 / 30	**66.7%**	27 / 40	**67.5%**	47 / 70	**67.1%**
Active CD vs inactive CD	17 / 19	**89.5%**	8 / 11	**72.7%**	25 / 30	**83.3%**
Active UC vs inactive UC	12 / 22	**54.6%**	12 / 18	**66.7%**	24 / 40	**60.0%**

Table 2.6: sensitivity (true positive), specificity (true negative) and accuracy results.

2.3.2 Liver diseases

The developed method has been applied to the analysis of serum samples from healthy donors and patients affected from different liver diseases. The serum proteins were extracted only by IDA-Cu(II) bulk-solid phase extraction due its capability to extract phosphorylated proteins, and the MALDI spectra acquired. Figure 2.8 shows a MALDI spectrum from one sample of each group analysed.

Protein phosphorylation is a dynamic and reversible event essential to the proper functioning of physiological processes, including cell proliferation and programmed cell death. Since protein phosphorylation is a regulatory event, it follows that the protein kinases that catalyze phosphorylation, should themselves be subject to regulation. The improper regulation of protein kinases has been implicated in many human pathologies, including cancer.

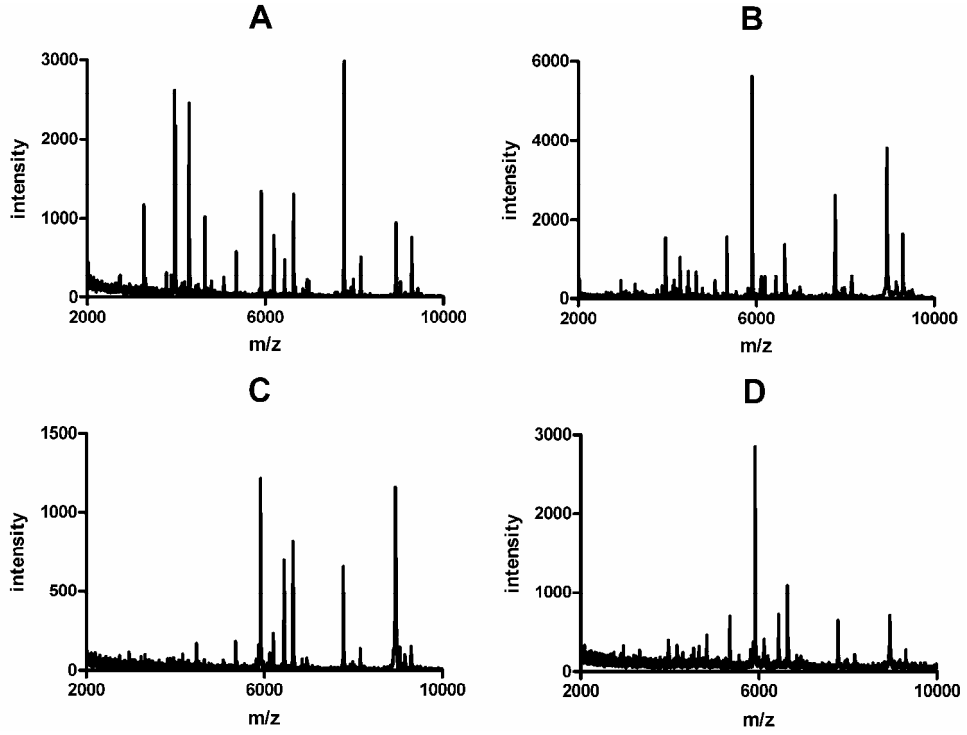

Figure 2.8: example of MALDI-TOF mass spectra obtained after sample preparation using IDA-Cu(II) solid-phase bulk extractions on serum from healthy donors (A) and patients affected by hepatitis C (B), cirrhosis (C) and hepatocellular carcinoma.

Two classification models have been calculated for the discrimination among hepatitis, cirrhosis and hepatocellular carcinoma patients (A), and for the discrimination among healthy donors, hepatitis, cirrhosis and hepatocellular carcinoma patients (B).

In table 2.7 and 2.8 the LDA results of the two classification models are reported. Figure 2.9A shows the perfect classification among the different diseases, with a classification and prediction ability of 97.3% and 88.6% respectively. This model has been calculated after selection of 10 relevant variables.

Number of selected variables	LDA results							
	Classification ability (%)				Prediction ability (%)			
	HV	CIR	HCC	Average	HV	CIR	HCC	Average
10	93.7	100	100	**97.3**	78.9	92.3	100	**88.6**

Table 2.7: classification and prediction ability results of LDA for the classification model A: hepatitis (HV), cirrhosis (CIR) and hepatocellular carcinoma (HCC) patients

In the classification model B (15 variables selected) the healthy donors category has been included, thus decreasing both the classification and prediction ability to 95.1% and 80.6%. This can be explained by the low number of serum samples analysed and to the greater differences among healthy and diseased serum protein profiles than the one existing among the different diseases (data shown in figure 2.9B).

Number of selected variables	LDA results									
	Classification ability (%)					Prediction ability (%)				
	HV	CIR	HCC	H	Average	HV	CIR	HCC	H	Average
15	91.6	96.9	91.7	100	**95.1**	78.9	69.2	75.0	94.4	**80.6**

Table 2.8: classification and prediction ability results of LDA for the classification model B: healthy donors (H) and hepatitis (HV), cirrhosis (CIR) and hepatocellular carcinoma (HCC) patients.

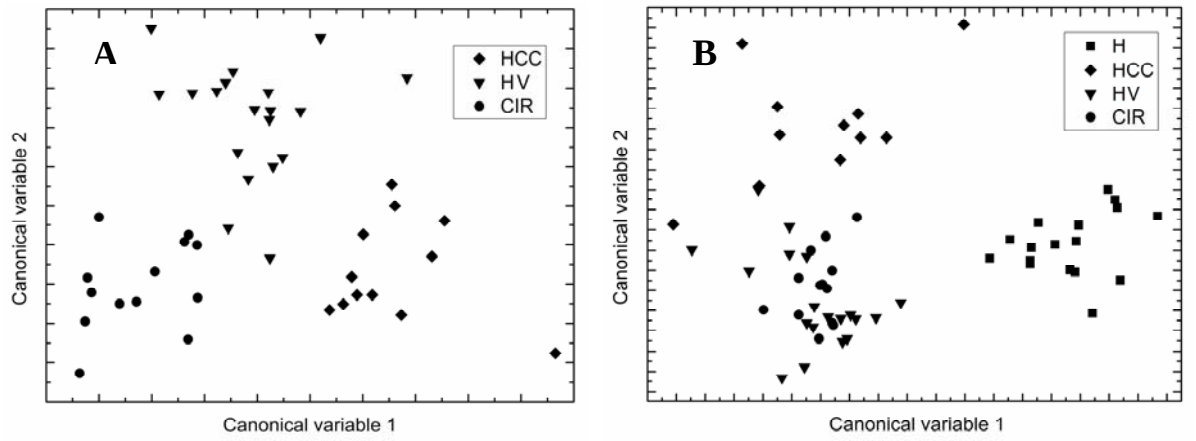

Figure 2.9: graphical display of LDA results obtained from MALDI-TOF MS serum protein profiles of hepatitis, cirrhosis and hepatocellular carcinoma patients (A), and healthy donors, hepatitis, cirrhosis and hepatocellular carcinoma patients (B). Objects are projected on the first two canonical variables.

2.3.3 Protein profiles specificity

A very common problem in protein profiling studies is the evaluation of the specificity of the obtained protein profiles, indeed the studies are often performed without the inclusion of sample to be used as a control (physiological state or other type of diseases). For this reason some of the data from both the studies have been

combined in a single dataset and analysed by LDA. In particular, since the two works have been performed in different lack of time, the MALDI-TOF MS spectra of hepatitis, cirrhosis and hepatocellular carcinoma were combined with a new set of MALDI-TOF MS spectra of ulcerative colitis serum samples analysed in the same period. The results (figure 2.10) demonstrate a good discrimination among ulcerative colitis and the different liver diseases, with an average classification and prediction ability of 92.5% and 75.8% respectively.

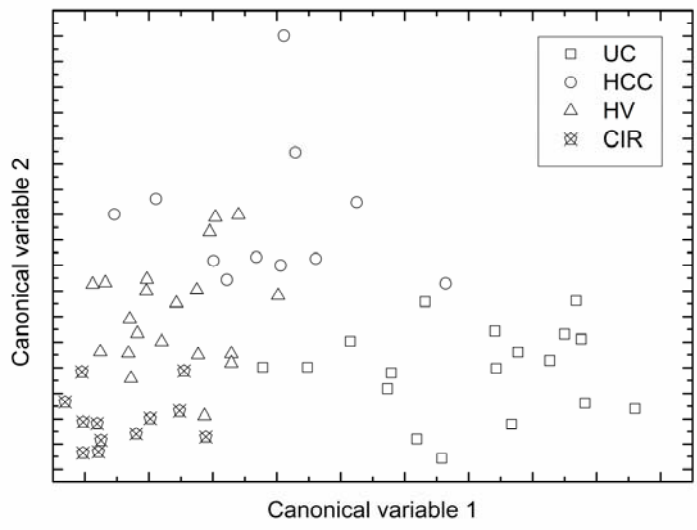

Figure 2.10: graphical display of LDA results obtained from MALDI-TOF MS serum protein profiles of hepatitis (HV), cirrhosis (CIR), hepatocellular carcinoma patients (HCC) and ulcerative colitis (UC). Objects are projected on the first two canonical variables.

2.4 Conclusions

The bulk solid-phase extraction was proved to be an useful methodological approach for the sample preparation and clean-up prior to MALDI-TOF MS analysis. This approach allows to obtain an enrichment of serum protein/peptides due to the high interaction surface between analytes and functionalized groups and a high recovery due to the elution step performed directly on the MALDI-target plate.

The application of LDA to the analysis of MALDI-TOF MS serum proteins profiles from IBD and liver diseases patients allows the classification of serum samples. In particular among healthy subjects and Crohn's disease and ulcerative colitis patients the serum protein profiles obtained using the reverse phase extraction gave the best

prediction ability results (96.9%) and for this reason this technique has been the first choice for the prosecution of the IBDs serum protein profiling studies. After the analysis of 129 samples from H, CD, UC and inflammatory controls using C18 a 90.7% of classification ability and a 72.9% prediction ability were obtained. Furthermore the use of SELECT features selection technique for the selection of the variables with high discriminant power permitted the identification of a pattern of 25-30 m/z values involved in the differentiation and classification of serum samples. The recognition of this m/z values may give useful information in order to identify protein biomarkers involved in the disease.

The application of the method to serum biomarker discovery of viral hepatitis, cirrhosis and hepatocellular carcinoma using IDA-Cu(II) as extraction procedure allowed to identify a pattern of 15 discriminating peptides an to classify the samples with a 80.6% of correct prediction ability.

Finally, a cross-study was performed for the evaluation of the obtained profiles and the MALDI spectra from serum of healthy donors, ulcerative colitis, hepatocellular carcinoma, viral hepatitis and cirrhosis patients were combined in a single data set, obtaining a 75.8% of prediction ability.

The obtained results are very promising for the use of this analytical method as a simple tool for diagnostic and biomarker discovery in liver and inflammatory bowel diseases.

Chapter 3

Label-free liquid chromatography electrospray ionization-quadrupole/Time-of-Flight mass spectrometry differential analysis of protein profiles

3.1 Introduction

The development of new mass spectrometry protein profiling strategies is a fundamental step for the discovery, identification, quantification and validation of disease related alterations of protein abundance in clinical samples. For this reason recently several comparative proteomics approaches have been developed, most of them based on the use of isotopes or mass tag labelling of the peptides in the samples to be compared [5, 6, 64, 65]. Despite some disadvantages such as the high cost of isotopic labelling and the impossibility of retrospective comparisons among samples due to their pairwise analysis, a high accuracy can be obtained.

Quantification methods based on a label-free strategy have been proposed [66-68] as promising alternatives. It's here described a label-free liquid chromatography electrospray ionization / Quadrupole - Time-of-Flight (LC ESI/QTOF) differential mass spectrometry (MS) approach combined with targeted MS/MS analysis of only identified differences for the study of inflammatory bowel diseases, and in particular Crohn's disease.

Two parallel studies in serum samples and epithelial cells (IECs) isolated from fresh biopsies or surgical specimen of Crohn's disease patients (CD) and healthy controls (H) (endoscopical screening for colorectal cancer, diverticulitis cancer) were carried out with different aims.

In serum samples from H an CD the study was performed for biomarker discovery on the low molecular weight (LMW) proteins, enriched by ultrafiltration using molecular filter devices with a 10 kDa cut-off,

The purpose of the work on IECs was the study of the protein involved in the mechanism of inflammation on the isolated subcellular fractions, while the evaluation of new potential biomarkers for CD, which could be initially produced at the site of disease, was only a sub-aim of this second work because of the invasive

procedure necessary for sample collection. Monodimensional electrophoresis was performed on the compartmental proteins, and afterwards the slices were cut in three parts and digested with trypsin.

After sample preparation reverse phase (RP) - LC ESI/QTOF MS analysis was performed for both works, and using appropriate bioinformatic tools the LC-MS runs were aligned. The differentially expressed peptides in the pooled samples were statistically evaluated and identified by RP-LC ESI/QTOF MS/MS analysis followed by database search.

3.2 Experimental

3.2.1 Materials and reagents

All the reagents, the Lowry and Bradford protein assay kits and the solvents for high-performance liquid chromatography were purchased from Sigma Aldrich (St. Louis, MO, USA).

3.2.2 Serum

3.2.2.1 SERUM SAMPLES

Blood samples were provided by the Gastroenterology Unit at S.Orsola University Hospital of Bologna (Italy) after obtaining informed consent from the patient.

A total of 65 subjects, including 50 healthy donors volunteers (H; 22 M and 28 F) and 15 Crohn's disease patients (CD; 9 M and 6 F), were admitted to this study. All the subjects were ranging in age from 20-65 years. The diagnosis of the IBD was achieved by clinical symptoms, clinical chemistry data and conventional endoscope procedures and only patients with active disease were included.

From each subject five millilitres of blood sample were collected and centrifuged at 4000 rpm for 10 min. The obtained serum samples were distributed into aliquots and stored frozen in plastic vials at -80°C until use.

Six pools of serum samples from healthy donors (250 µL from 8 different serum samples each) and five pools of serum samples from Crohn's disease patients (250 µL from 3 different serum sample each) were created and aliquoted.

3.2.2.2 LOW MOLECULAR WEIGHT SERUM PROTEINS ENRICHMENT

Protein quantification of each pooled serum sample was performed in triplicate by Lowry protein assay using bovine serum albumin (BSA) as a standard according to the manufacturers' instructions (Sigma-Aldrich).

For each sample 25 mg of proteins amount was diluted to 2 mL of 25 mM NH_4HCO_3 pH 8.2 containing 10% (v/v) acetonitrile (ACN), 10 pmol of Alcohol deydrogenase (ADH) tryptic digest were added as internal standard and incubated at room temperature for 30 minutes to disrupt the protein-protein interactions [69].

The diluted serum sample was transferred to a Centricon centrifugal filters with a 10 kDa molecular weight cut-off (MWCO)(Millipore Corporation, Bedford, MA, U.S.A.) to deplete the high molecular weight proteins [70]. The sample was centrifuged at 3000 g until 90% of the diluted serum had passed through the membrane. At the end of the process approximately 1400 µL of LMW protein enriched serum sample was collected and distributed into aliquots. Two hundreds microlitres of the pooled serum collected filtrate was lyophilized to dryness and resuspended in 30 µL of deionized H_2O containing 0.1% formic acid (FA) for the following LC-MS analysis.

3.2.3 Cells

3.2.3.1 CELL LINE

HT29 colorectal adenocarcinoma cell line (American Type Culture Collection) was used. HT29Cl.16E is a goblet cell line and was grown in Dulbecco's Modified Eagle Medium (DMEM) with 10% Fetal Calf Serum (FCS) supplemented with 50 units of penicillin/ml and 50 µg of streptomycin/ml (Sigma). For stimulation experiments, 1% FCS was used. HT29 were grown in Falcon flasks (75-cm2) and the medium was changed every 2 or 3 days.

HT29 cell line was cultured in the absence of interferon γ (IFNγ, Sigma) or treated for 5 hours with a concentration of IFNγ 75 ng/mL.

3.2.3.2 ISOLATION OF HUMAN INTESTINAL EPITHELIAL CELLS

Intestinal epithelial cells (IECs) were isolated from surgical specimens of patients undergoing operative procedures for cancer and diverticulitis (healthy control) and of patients affected by Crohn's disease obtained directly from the operating room at the Gastroenterology Unit at S.Orsola University Hospital of Bologna (Italy) after informed consent from the patients. IECs were isolated as described previously [71]. Briefly, specimens were washed extensively with PBS. The mucosa was stripped

from the submucosa, cut into small pieces and placed in 1 mM dithiothreitol, (DTT, Sigma) for 15 minutes at room temperature to remove mucous. The tissue was then washed in PBS and incubated twice in Dispase II 3 mg/ml (Roche Diagnostics, Mannheim, Germany) in RPMI1640 (Sigma) for 30 minutes in a 37°C incubator, vortexing every 5 minutes. The supernatant (released IECs) was collected and washed in medium (RPMI). The viability of isolated IECs was >95%. Only IECs derived from colon were analysed.

3.2.3.3 PROTEIN EXTRACTION FROM SUBCELLULAR FRACTIONS

Proteins from cytosolic, membrane, nuclear and cytoskeletal fractions were extracted from $5x10^6$ of stimulated and non-stimulated HT29 cells and healthy and affected by Crohn's disease IECs using QProteome cell compartment kit (Qiagen, Milano, Italy) according to manufacturer instructions and purified from contaminants by acetone precipitation. Four volumes of ice-cold acetone were added to every sample and proteins were allowed to precipitate for 1 hour at -20°C. After centrifuging the protein pellets were air dried and resuspended with ammonium bicarbonate 100 mM. The quantification was performed by BCA method (Cyanagen, Bologna, Italy) and every sample diluted to a final concentration of 2 µg/uL.

3.2.3.4 MONODIMENSIONAL GEL ELECTROPHORESIS

Proteins from all the isolated compartments were separated on 9% sodium-dodecyl sulphate (SDS)-PAGE monodimensional gels. Fifteen micrograms of proteins for each compartment were loaded for treated/non-treated HT29 cells while five micrograms of proteins for H and CD IECs isolated compartments. Each sample was diluted with Lemmli buffer and loaded on the gel in a way to easily compare same compartments from both H and CD samples. The gel staining was performed using Comassie Blu (Bio-rad, Hercules CA), while the destaining was done by washing with a mixture 50% methanol, 30% water and 20% glacial acetic acid. Each lane was cut into 2 slices (same area for every couple of samples to be compared) before reduction with 10 mM DTT (Sigma) in 100 mM ammonium bicarbonate for 45 minutes at 56°C. Afterwards alkylation was performed by addition of 55 mM iodoacetamide in 100 mM ammonium bicarbonate for 30 minutes at room temperature in the dark. The gel slices were washed with 100 mM ammonium bicarbonate and acetonitrile and incubated overnight in a 50:1 (w/w) ratio with 12 ng/µL sequencing-grade-modified trypsin (Promega, Falkenberg, Sweden). The day

after the digested peptides were extracted from the slices by following addictions of ammonium bicarbonate 25 mM, acetonitrile and 5% formic acid. Every sample were lyophilized to dryness and resuspended with 16 µL of 0.1% formic acid.
The reproducibility of the whole digestion process has been tested and the coefficient of variation calculated by using bovine serum albumine (BSA) as a standard.

3.2.3.5 WESTERN BLOT ANALYSIS OF IECs AND HT29 CELL LINES

To evaluate the subcellular fractionation of each protein sample one more monodimensional gel separation was performed as described above. Membrane protein Gp180, nuclear protein SOX9 and cytoskeletal protein β-actin were used as controls of the subcellular fractionation.
Equal amounts of proteins from each subcellular fraction (20 µg) were subjected to sodium dodecyl sulfate (SDS)-9% polyacrylamide gel electrophoresis. The proteins were transferred onto a nitrocellulose membrane (Hybond-C extra, Amersham), that was blocked with PBS/Tween 0.1% + 5% milk for 2 hours and incubated overnight at 4°C with the primary antibodies anti-Gp180 (kindly provided by the immunobiology center, Mount Sinai Hospital, NY, 10 µg/mL in in PBS/Tween 0.1% + 5% milk), the rabbit-polyclonal antibody anti-SOX9 (Santa-Cruz, CA, USA, diluted 1:100 in PBS/Tween 0.1% + 5% milk) and the monoclonal antibody anti-β-actin produced in mouse (Sigma, diluted 1:2000 in PBS/Tween 0.1% + 5% BSA). Afterwards the incubation with the appropriate horseradish peroxidase-conjugated secondary antibodies (DAKO, Glostrup, Denmark, diluted 1:200 in PBS/Tween 0.1% + 5% milk) was performed. The presence of bands was revealed with the enhanced chemiluminescence detection kit Westar Nova (Cyanagen).

3.3.3 Liquid chromatography and mass spectrometry

3.3.3.1 LMW SERUM PROTEINS

LC-MS analysis was performed on a CapLC with flow splitting from 5 µL/min to 250 nL/min, connected with a nano electrospray interface to a QTOF Ultima (Waters, Manchester, U.K.) using MassLynx v4.0 software as operating software. An Atlantis dC18 NanoEase column (150 × 0.75 mm, 3 µm) (Waters) with a C18 Intersil precolumn (0.3 × 5 mm, 3 µm particle size) (LC-Packings, Skandinaviska Genetec AB) was used.

The mobile phase had a constant concentration of 0.1% formic acid, with an acetonitrile gradient after 10 min from 5 to 60% over 125 minutes, followed by 10 minutes column cleaning at 80% acetonitrile, and 15 minutes equilibration. Two blank injections and a 60 minute gradient plus column washing were run between samples to minimize sample carry over. For quantitative experiments, the QTOF was set to scan in profile mode m/z 400-1800 with 1.9 seconds per scan and 0.1 seconds of scan delay.

Each sample was analysed in triplicate, each with a 3 µl injection. For targeted MS/MS, 6 µl sample was injected, the same LC gradient was run and the survey scan used 1 second scan time and a peak limit of 20 count to switch to MS/MS mode. For inclusion lists the time tolerance was set to 180 s.

3.3.3.2 CELL PROTEINS

LC-MS analysis was performed on a CapLC with flow splitting from 4.8 µL/min to 260 nL/min, connected with a nano electrospray interface to a QTOF micro (Waters, Manchester, U.K.) using MassLynx v4.1 software as operating software. An Atlantis dC18 NanoEase column (150 × 0.75 mm, 3 µm) (Waters) with an Atlantis dC18 NanoEase precolumn (0.3 × 5 mm, 5 µm particle size) (Waters) was used.

The mobile phase A for the nanoLC separation was H_2O/acetonitrile (95:5) 0.1% FA while the mobile phase B was acetonitrile/H_2O (95:5) 0.1% FA. The chromatograpic gradient was set up to give a linear increase after 3 min from 2% B to 10% B in 7 min, from 10% B to 60% B in 15 min and from 60% B to 80% B in 10 min. After 8 min at 80% B the column is conditioned again at 5% B for 20 min. The time of a single run was 100 min.

A 30 min gradient blank injection was run between sample to minimize carry over. For quantitative experiments, the QTOF was set to scan in profile mode m/z 400-1800 with 1.9 seconds per scan and 0.1 seconds of scan delay.

Each sample was analysed in triplicate, each with a 3 µl injection. For targeted MS/MS, 4 µl sample was injected, the same LC gradient was run and the survey scan used 1 second scan time and a peak limit of 15 count to switch to MS/MS mode. For inclusion lists the time tolerance was set to 120 s.

3.3.3.3 MS DATA ANALYSIS AND INCLUDE LIST GENERATION

Both LC-MS MassLynx v4.0 and v4.1 raw data files were converted to mzXML using the appropriate massWolf version [72]. Peptide feature finding was performed using

msInspect version 1.01 [23]. The resulting feature lists were aligned using msInspect, (setting a mass window of 0.2 m/z and a time window of 75 scans) and the detailed results files were used for further processing. For the alignments of LMW serum protein analysis the triplicates of each pool of serum sample from healthy donors were analysed with the triplicates of each pool of serum sample from Crohn's disease patients, while for the alignments of cell subfraction proteins the triplicates of each control gel slice were analysed with the triplicates of the corresponding sample gel slice. The normalization was performed using the msInspect algorithm during the alignment step.

The aligned peak lists were analysed using newly developed plugins for the PROTEIOS 2 analysis platform [24]. First the MsInspect Details Analyser (version 0.91) plugin was used to find features which were significantly upregulated in the treated samples and to automatically produce include lists for MS/MS identification. For a feature to be identified as significantly up regulated it had to be present in at least two sample replicates and have a p-value of less than 0.05 in a homoscedastic student's t-test. For the t-test the total intensities, which represents the peak volume, were used. For features where peaks could not be found in the control samples, a value of 50 ion counts was used, which was an estimate for the detection level in the present setup. Features which were upregulated at least 1.5 times in the treated sample compared to the control and with a significant pValue were sorted according to intensity and put into include lists with a maximum of 150 peaks per include list. The retention time of the lastly acquired sample was used for the include list.

3.3.3.4 FEATURE PEPTIDE IDENTIFICATION

To generate peak lists for peptide identification, ProteinLynx Global Server 2.2 (Waters) was used. The xml format peak lists were converted to mzData using PROTEIOS. Mascot version 2.1.02 (www.matrixscience.com) was used for peptide identification.

For LMW serum proteins the Sprot human database, version 53.0 was used, 191913 sequences in total. The search settings were 0.1 Da precursor and fragment tolerances, no fixed modification carbamidomethylation of cysteine and none enzyme. For cell subfraction proteins the Sprot human database, version 54.7.0 was used (333445 sequences in total) with 0.2 Da precursor and fragment tolerances, carbamidomethylation of cysteine as fixed modification and trypsin as enzyme.

The search results were exported as XML and analysed using the "ProteinLynx Global Server to mzdata converter" PROTEIOS plugin. Using the plugin, peptide search results were matched with the MS features, with a retention time tolerance of 100 s and a mass tolerance of 0.12 Da.

3.3 Results and discussion

The goal of these works was the development of a methodological approach for the analysis of the protein expression profiles in serum samples and intestinal epithelial cells from healthy donors (H) and Crohn's disease patients (CD), based on a label-free LC ESI/QTOF differential MS approach combined with targeted MS/MS analysis of only identified differences.
Different sample preparation approaches have been developed for the two studies, but similar mass spectrometry and data analysis strategies have been used. The general workflow is shown in figure 3.1.

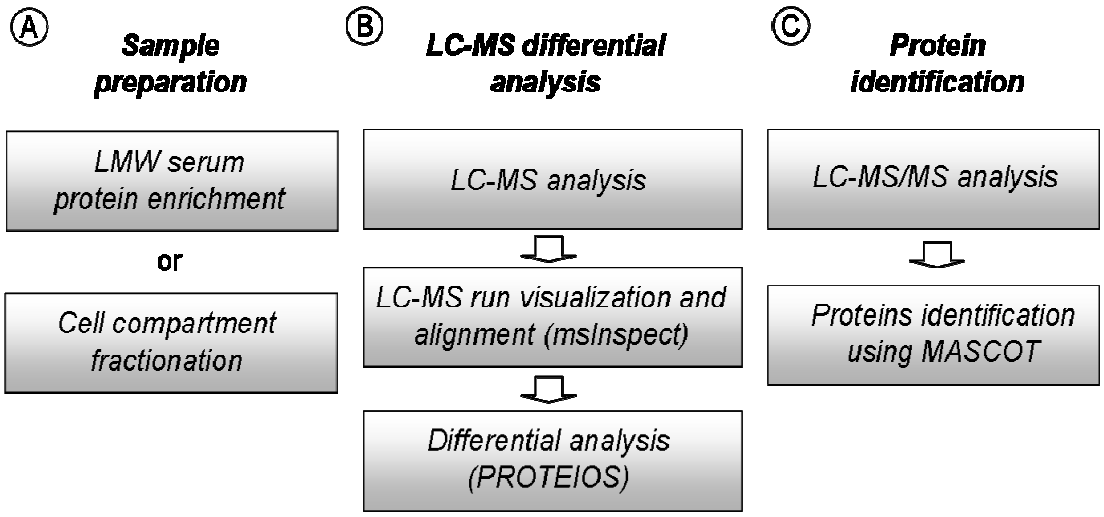

Figure 3.1: Outline of the sample preparation procedure (low molecular weight serum proteins enrichment or cell compartment fractionation) (A), label-free LC ESI/QTOF MS differential analysis (B) and protein identification (C).

Protein samples were analysed in triplicate by LC-MS using scan mode in the 400-1800 m/z and the data from the samples to be compared have been converted to mzXML format in order to be imported into the opensource software msInspect. Bidimensional maps of the LC-MS runs have been created for each sample and a

specific algorithm for feature finding was used with the aim to evaluate and filtrate the protein/peptides signals. Figure 3.2B shows a bidimensional image obtained from the LC-MS run (3.2A) analysis of a serum sample. In the box the zoom of an area is proposed as an example, in order to highlight the feature finding step: the red points represent the m/z values which satisfy the parameter described in the material and method section (minimum number of scan, intensity, charge, peaks…).

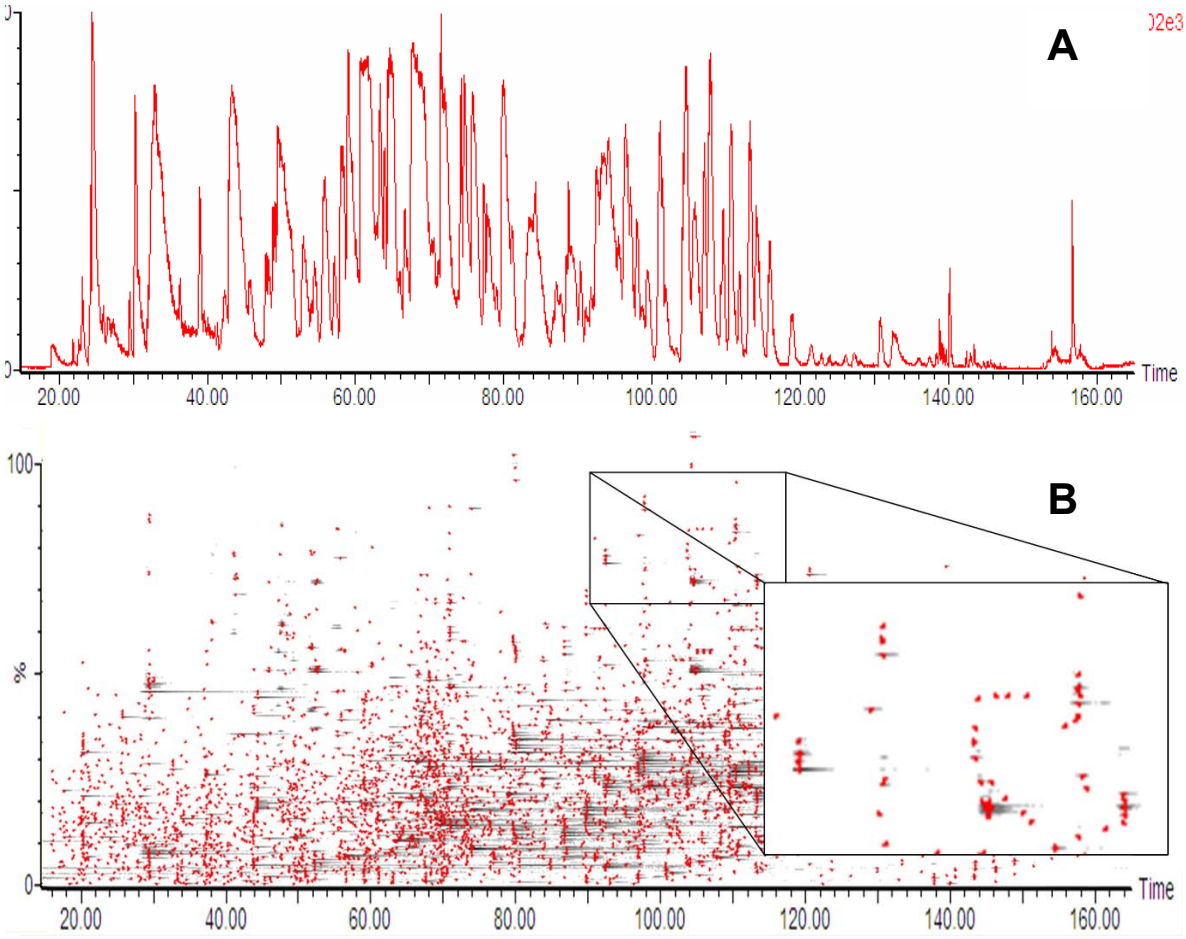

Figure 3.2: example of LC ESI/QTOF MS run of a serum protein sample. Part A: BPI visualization of a chromatogram; part B: msInspect bidimensional data visualization. The spot dimension is the ion intensity. The extracted features (peaks) are highlighted in red.

Afterwards the features from each sample were normalized and aligned by an algorithm implemented in the software and the resulting peptide arrays were used for further analysis by PROTEIOS with the aim to evaluate the differences in protein expression (at least 1.5-fold change) and the creation of a list of peptides to include in the following MS/MS analysis. At the time this approach allowed the comparison

of only two categories of samples (healthy donors and Crohn's diseases for both the study performed).
Two alignment were done for every set of data to be compared: one to the last acquired healthy donor sample and one to last acquired Crohn's disease sample. This permitted to create for each category an include list containing the overexpressed peptides, reducing the normal problems of retention time shift due to the lag of time between MS and MS/MS analysis.
The details concerning sample preparation and the results of the works are discussed in the following sections.

3.3.1 Serum protein profiling in IBD by LC ESI-QTOF MS differential analysis

The Low Molecular Weight (LMW) serum proteome was investigated. Since its high variability, due to different factors such as sample collection, clotting time, storage and presence of high abundant proteins [73-77], some researchers consider it as "biological trash" or noise [78,79]. Despite this, the LMW proteome contains a huge amount of disease-related information which only need to be understood [80,81]. New approaches have been recently proposed for its comprehension, and in particular Villanueva *et al* [82] suggested and demonstrated that some biomarkers are not directly expressed by the diseased tissues, but are ex-vivo products from exoprotease activity in the coagulation and complement activation pathways. In his work he focused in particular on different types of cancers (breast, ladder and prostate) because they involve the transformation of altered cells types that produce specific proteins, but the same theory can be potentially applied to every type of disease, such as IBDs.
There are many factors related to serum proteins analysis that must be controlled, first of all the serum sampling procedure. For this reason a standard protocol concerning clotting time, storing and aliquoting has been used. The consequence of a non-well designed sample collection procedure can be the serum transformation that brings to high level of degradation peptides presents in the sample caused by the ex-vivo enzymatic activity. Furthermore it must be considered that the presence of high abundance proteins (HAP) such as albumine or immunoglobuline hides the low molecular weight proteome which need to be enriched. Different approaches have been developed for their depletion (most of them based on immunoaffinity [83-

86], but a lack of reproducibility and specificity [87-92], and the removal of LMW proteins bound to albumine [93,94] have been reported [95]. Moreover the high cost of the columns prevents their use for large-scale studies.

In this work the LMW serum samples proteins enrichment was performed by ultrafiltration using a molecular filter devices with a 10 kDa cut-off, after testing the reproducibility of the whole process using Universal Proteomics Standard (Sigma) digest as a standard.

The Lowry assay from Sigma for the protein quantification of the filtrates was tested, but since this assay is not suitable for the quantification of peptides (such as Bradford or Nanodrop), it was tested to evaluate the order of magnitude of concentration of the peptides in the filtrate. The obtained results allowed to estimate a peptide concentration around 150 ng/µL, in agreement with the results presented in other paper where a similar approach is described [96], but due to the extremely low precision and accuracy in the quantification of the peptides in the filtrate, the normalization of the protein amount was performed in the pooled serum samples before the ultrafiltration step.

Ten picomoles of Alcohol dehydrogenase (ADH) tryptic digest were added as internal standard to 25 mg proteins of each pooled serum sample to evaluate the sample-to-sample variations in the whole process. Afterwards samples were diluted to 2 mL with 25 mM ammonium bicarbonate containing 10% (v/v) acetonitrile and incubated 30 min at room temperature in order to disrupt the protein-protein interactions [96,97].

After the LC-MS analysis the msInspect software was used for the feature finding in all the LC-MS run, and afterwards all the peaks were filtered, aligned and normalized as described in the Material and Methods section. Due to the length of the LC analysis (165 min) and to the flow that is passively split in the CapLC (from 5 µL/min to 250 nL/min) the retention time varied of almost one minutes among the LC runs, and for this reason the alignment was performed using a quite wide retention time windows (75 scans), with an associated m/z window of only 0.1 Da.

The differently expressed LMW proteins in CD and H serum samples were evaluated by statistical analysis, performed using PROTEIOS. One include peak list for the overexpressed peptides in CD samples and one for the overexpressed peptides in H samples were created.

The LC-MS/MS analysis of the peaks inside the include lists followed by MASCOT database search allowed the identification of many peptides resulting from some proteins commonly involved in the inflammatory processes (table 3.1).

Protein	Entry name	MW (Da)	N° of AA	Overexpression
Apolipoprotein A-IV	APOA4	45399	396	H
Apolipoprotein E	APO E	36154	317	H
Alpha-2-HS-glycoprotein (Fetuin A)	FETUA	39325	367	H
Fibrinogen alpha chain	FIBA	94973	866	CD
Fibrinopeptide A	FPA	1537	16	CD
Complement C3	C3	187148	1663	CD
Complement C3f fragment	C3f	2021	17	CD
Complement C4-A	C4-A	192771	1744	CD

Table 3.1: overexpressed proteins in CD and H samples identified by LC-MS/MS analysis and MASCOT database search

At first glance these results didn't seem to be interesting, but after an in-depth analysis of the MASCOT results we discovered that most of the identified differently expressed peptides in CD and H subjects fall into tight clusters (table 3.2). These clusters seem to be generated by exopeptidase activities allowing the discrimination between CD and H patients based on a fingerprint of peptides resulting from specific proteolytic and complement degradation pathways as described by Villanueva *et al* [82]. Moreover most of the differently expressed proteins identified are exactly the same he discovered as involved in breast, ladder and prostate cancer (table 3.3).
Fibrinopeptide A, a peptide released from the amino end of fibrinogen by the action of thrombin to form fibrin during clotting of the blood, is the most overexpressed peptide in Crohn's disease: for its identification a cluster of 8 peptides was identified with an impressive medium variation of almost 700-fold between CD and H. The identification of proteins such as fibrinogen-α and complement 3 that have a molecular weight much higher than 10 kDa (the MWCO applied to the serum) can be explained only by the activity of serum protease and the consequent production of their degradation peptides. The peptides from all these proteins had a 3-fold increase in CD samples. Among the peptides from complement 3 protein, that plays a central role in the activation of the complement system, three peptides from its fragment C3f have been identified with a 3 fold-change.

	m/z		H	CD
FPA	655.27	DSGEGDFLAEGGGV		
	733.33	DSGEGDFLAEGGGVR		
	675.79	SGEGDFLAEGGGVR		
	632.29	GEGDFLAEGGGVR		
	603.79	EGDFLAEGGGVR		
	539.28	GDFLAEGGGVR		
	510.72	DFLAEGGGVR		
	453.25	FLAEGGGVR		
FIBA	574.25	MADEAGSEADHEGTHSTKRGHA		
	624.92	KMADEAGSEADHEGTHST		
	679.28	YKMADEAGSEADHEGTHST		
	708.3	SYKMADEAGSEADHEGTHST		
	695.82	NRGDSTFESKSY		
	560.74	GDSTFESKSY		
	532.23	DSTFESKSY		
	725.82	STSYNRGDSTFES		
C3f	549.78	HWESASLLR		
	471.74	HWESASLL		
	568.8	THRIHWESA		
C3	641.31	ENEGFTVTAEGK		
C4-A	542.64	GFKSHALQLNNRQI		
	647.36	SHALQLNNRQI		
APOA4	886.44	SLAELGGHLDQQVEEF		
	786.39	AELGGHLDQQVEEF		
	750.85	ELGGHLDQQVEEF		
	686.34	LGGHLDQQVEEF		
	629.79	GGHLDQQVEEF		
APO E	803.74	AATVGSLAGQPLQERAQAWGERL		
	756.38	TVGSLAGQPLQERAQAWGERL		
FETUA	694.34	HTFMGVVSLGSPSGEVSHPR		

Table 3.2: Serum peptide signatures for Crohn's disease (CD) and healthy donors (H). The peptides are listed in clusters of overlapped sequences and the differences in their expression in CD and H serum samples is reported (arrows).

Apoliprotein A-IV and E have been found less expressed in diseased patients with a difference of about 2.5-fold. This is perfectly in agreement with recent studies that showed how apoliprotein A-IV inhibits experimental colitis in vivo [98] and can be considered as an independent predictor of disease activity in patients with inflammatory bowel diseases [99].

Apolipoprotein E (apoE) is a polymorphic multifunctional protein with three common isoforms in humans (E2, E3, and E4): ApoE3 is the wild-type and most common isoform, but the identification of the isoform involved in this study was not possible. The role of ApoE as "inflammatory imbalance" between pro- and anti-inflammatory mediators has been reported [100-101].

Table 3.3 lists and compares the clusters of peptides found both in this and in Villanueva work: the overlapping among peptides involved in different diseases

underlines the importance of the abundance of the proteins in the sample as substrate for protease activity. Indeed by looking at the differences in the peptides patterns it's possible to observe how high abundance proteins such as fibrinogen-α are only the target of the exoproteases and are not themselves really overexpressed. This means that specific exoproteases can be involved in Crohn's disease and the peptide signature they produce can be potentially used as biomarker.

Further studies need to be performed for the validation of the obtained results, and for this aim the sequence-specific exopeptidase activity test (SSEAT) [102] could be used. Briefly, it consists in the addiction to a serum sample of a standard of cluster precursor peptide (Fibrinopeptide A, C3f…), followed by a short incubation and MALDI-TOF MS analysis at different times. In this way the standard peptide is degraded and the degradation product signature acquired. By using in a proper combination more cluster precursor peptides it seems possible to create a fingerprint of peptides able to discriminate between healthy donors and Crohn's disease patients.

			Cancer			IBD	
			Prostate	Bladder	Breast	Crohn's disease	Healthy
FPA	**1536.68**	ADSGEGDFLAEGGVR	+	+	+		
	1465.65	DSGEGDFLAEGGVR	+	+		+	
	1350.64	SGEGDFLAEGGVR	+	+	+	+	
	1263.6	GEGDFLAEGGVR	+	+	+	+	
	1206.57	EGDFLAEGGVR	+	+		+	
	1077.53	GDFLAEGGVR	+	+		+	
	1020.47	DFLAEGGVR	+	+		+	
	905.5	FLAEGGVR		+		+	
	758.45	LAEGGVR	+	+			
Fibrinogen α	3261.43	(K) SSSYSKQFTSSTSYNRGDSTFESKSYKMA	+	+			
	3190.36	(K) SSSYSKQFTSSTSYNRGDSTFESKSYKM	+	+			
	2931.2	(K) SSSYSKQFTSSTSYNRGDSTFESKSY		+			
	2768.26	(K) SSSYSKQFTSSTSYNRGDSTFESKS		+			
	2553.01	(K) SSSYSKQFTSSTSYNRGDSTFES	=	=	=		
	2379.03	.SSYSKQFTSSTSYNRGDSTFE		+			
	1390.62	NRGDSTFESKSY				+	
	1120.47	GDSTFESKSY				+	
	1063.45	DSTFESKSY				+	
	1450.53	STSYNRGDSTFES				+	
	2816.25	(R) GSESGIPTNTKESSSHHPGIAEFPSRG (K)					
	2293.97	MADEAGSEADHEGTHSTKRGHA				+	
	1872.74	KMADEAGSEADHEGTHST				+	
	2035.33	YKMADEAGSEADHEGTHST				+	
	2022.87	SYKMADEAGSEADHEGTHST				+	
	3239.22	SYKMADEAGSEADHEGTHSTKRGHAKSRPV (R)	/	/	/		
	2659.03	DEAGSEADHEGTHSTKRGHAKSRPV (R)					
C3f	1098.55	HWESALLR				+	
	2021.06	SSKITHRIHWESASLLR	=	=	=		
	1864.95	SSKITHRIHWESASLL.			+		
	1777.93	SKITHRIHWESASLL.					
	1690.9	KITHRIHWESASLL.					
	1562.84	ITHRIHWESASLL.			+		
	1449.76	THRIHWESASLL.					
	1348.7	HRIHWESASLL.	/	/	/	+	
	1211.7	RIHWESASLL.					
	1055.6	IHWESASLL.					
	942.43	HWESASLL.			+		
	1851.88	SSKITHRIHWESASL..					
	1136.55	THRIHWESA.				+	
C4a	1895.99	RNGFKSHALQLNNRQI (R)					
	1739.93	NGFKSHALQLNNRQI (R)					
	1626.85	NGFKSHALQLNNRQI.					
	1498.91	NGFKSHALQLNNR..					
	1625.85	GFKSHALQLNNRQI.				+	
	1293.68	SHALQLNNRQI.				+	
	3200.52	(R) GLEEELQFSLGSKINUKVGGNSKGTLKVLR	/	/	/		
	2704.13	(R) GLEEELQFSLGSKINUKVGGNSKGTL					
	2305.2	(R) GLEEELQFSLGSKINUKVGGNS					
	1762.87	(R) GLEEELQFSLGSKINUKV		+			
ApoA-IV	2508.16	ISESEEELRQRLAPLAEDVRGNL (K)					
	2755.2	(K) GNTEGLQKSLAELGGHLDQQVEEFR					
	1927.94	SLAELGGHLDQQVEEFR					
	1771.81	SLAELGGHLDQQVEEF.					+
	1571.75	AELGGHLDQQVEEF.					+
	1500.71	ELGGHLDQQVEEF.					+
	1371.66	LGGHLDQQVEEF.					+
	1258.57	GGHLDQQVEEF.					+
ApoE	2565.45	(R) AATVGSLAGQPLQERAQAWGERLR					+
	2409.13	(R) AATVGSLAGQPLQERAQAWGERL.					
	2267.12	TVGSLAGQPLQERAQAWGERL.					+

Table 3.3: serum peptides patterns in Crohn's disease and healthy controls, compared to peptides signature for breast, prostate and bladder evaluated in other research work by Villanueva *et al*. [82]. Coloured box corresponds to peptide presence, "+" to overexpression, "=" and "/" respectively to equally and randomly observed.

3.3.2 Protein profiling in HT29 cell lines and intestinal epithelial cells

The mucosal immune system uses a large number of mechanisms to protect the host against an aggressive immune response to luminal constituents.

These include a physical barrier, the luminal enzymes that alters the nature of the antigen itself, the presence of specific regulatory T cells in both the organized and disorganized lymphoid tissue of the gut, and the production of an antibody, the secretory immunoglobulin A (IgA). All these factors work in concert.

There is a balance between aggressive factors and endogenous protective and adaptive mechanisms in the intestinal mucosa. Mucosal inflammation is a process that represents a physiological response to microbes, foreign antigens but, if exaggerated, can represent the presence of a disease. In inflammatory bowel diseases (IBDs) the inflammation is very intense and there is a loss of barrier function.

The intestinal epithelium provides a barrier against both endogenous commensal microorganisms of the gut flora and enteropathogenic bacteria and viruses. In response to injury, intestinal epithelial cells increase their barrier activity by upregulating the production of anti-microbial factors such as α and β-defensins, and release chemokines that recruit other effector cells representing a connection between innate and adaptive immunity. Intestinal epithelial cells express various receptors for microorganisms (PRR), such as Toll like receptors (not in the normal state). Its activation (NF_kB pathway) provokes secretion of cytokines, antimicrobial defensins and tissue remodeling enzymes.

The mucosal barrier is important because it represents a system that alters antigen exposure with the help of enzymes and the emulsifying effect of bile salts, produce mucin glycoproteins that line the surface of epithelium and represent a protection against mechanical damage induced by the passage of fecal material or solid food.

Intestinal epithelial cells (IECs) function as a first line of defense, signaling the presence of noxious stimuli or pathogens to other effector cells, and separate the antigens in the lumen from the lymphoid tissue in the lamina propria.

Lamina propria lymphocytes (LPLs) are below the basement membrane and they are in contact with the projections of IECs that represent a way to interact with immune cells. That these cells play a role in regulating inflammation is seen in patients with IBD.

For example in ulcerative colitis there is an increase in IEC proliferation and an alteration in electrolyte secretion; in Crohn's disease T cells are more activated and as a consequence there is an increase in IEC proliferation. The inflammation process by itself alters the permeability of intestinal epithelial cells and as a consequence alters the function of other cells.

All of these processes are linked together and they trigger a cascade of alterations. Thus, the proliferation of IECs is the result of inflammation but is also involved in interactions with other cells of the innate immune system such as neutrophils.

The epithelial cells are joined together by tight junctions apically and basolaterally. Only ions can pass between the epithelium. In contrast in an inflamed state cells, proteins, luminal contents can pass because the tight junctions are less tight.

Lastly IECs are able to function as antigen-presenting cells for both CD4+ T (IBD) cells and CD8+ T cells (normal state).

Several studies have demonstrated that normal IECs selectively stimulate CD8+ T cells that are suppressive in function [103]. In contrast IECs from patients with IBD stimulate CD4+ T cells and these cells proliferate and secrete interferon-γ.

Due to the importance of IECs in IBD, the study of variations in their protein expression profiles in normal or diseased states is crucial for understanding new mechanisms insight the pathology, with the ultimate goal to develop clinically relevant biomarkers of the specific disease status in individual patients and to identify potential targets for therapeutic intervention. For this reason the developed MS-based protein profiling method has been applied to this clinical problem.

Since the work is based on the comparison among different samples, the preliminary phase of the work consisted in the evaluation of the yield of the tryptic digestion and peptide extraction from the gel and the calculation of the reproducibility. Six samples of 10 µg of Bovine serum albumine (BSA) were loaded on a 9% SDS-PAGE monodimensional gel, digested by trypsin and the resulting peptides analysed in triplicate by LC-MS. The intensity of five peaks for every spectrum was considered and their coefficients of variation (CV%) were calculated. A medium CV% of 34.9 was obtained (data not shown), thus allowing to fix to 1.5 the fold-change threshold for the peptides.

Another parameter evaluated in the preliminary phase of the study was the efficiency of the subfractionation protocol. Three proteins from different compartments were used for western-blot analysis: β-actin is a predominantly cytoskeletal protein, but findable in every compartment, while gp180 (membrane

protein) and SOX9 (nuclear protein) were chosen because both of them are involved in Crohn's disease. In particular gp180 seems to be overexpressed in membranes from healthy donors and SOX9 in nuclei from Crohn's disease patients. Figure 3.3 shows the compartmental distribution of the proteins obtained by the analysis of compartments from treated and normal HT29 cells. As expected gp180 was detected only in membranes from normal cells while SOX9 was more concentrated in nuclei from treated samples. β-actin is present in every compartment with the same concentration ratio among normal/treated samples.

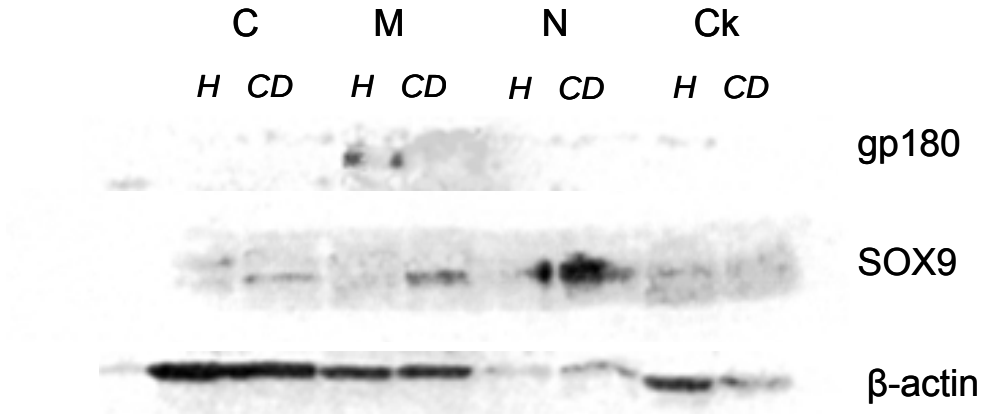

Figure 3.3: subcellular fractions from HT29 cells analysed by western blot. Gp180 is a membrane protein and SOX9 is a nuclear protein, while β-actin is present in every compartment (C: cytosol; M: membrane, N: nuclei; Ck: cytoskeleton)

Once the method was developed, before analysing IECs from intestinal biopsies, it was applied to a cellular inflammation model of intestinal HT29 cells treated with cytokines at the concentrations of IFNγ 75 ng/mL to simulate the chronic inflammation (concentration evaluated in previous studies conducted in our laboratories). The subcellular fractionation was performed on five million of cells both treated and normal and the cytosolic, nuclear, membrane and cytoskeletal compartments were obtained. Afterwards the proteins were concentrated by acetone precipitation and for each fraction 15 µg of proteins were loaded on a monodimensional gel (figure 3.4A). The different protein patterns in the different compartments confirm the subcellular fractionations, while the similarity among the corresponding compartments of treated and normal cells was expected.
Cytosol and nuclei lanes were divided into three slices in order to perform a further fractionation of the sample prior to tryptic digestion and LC-MS analysis.

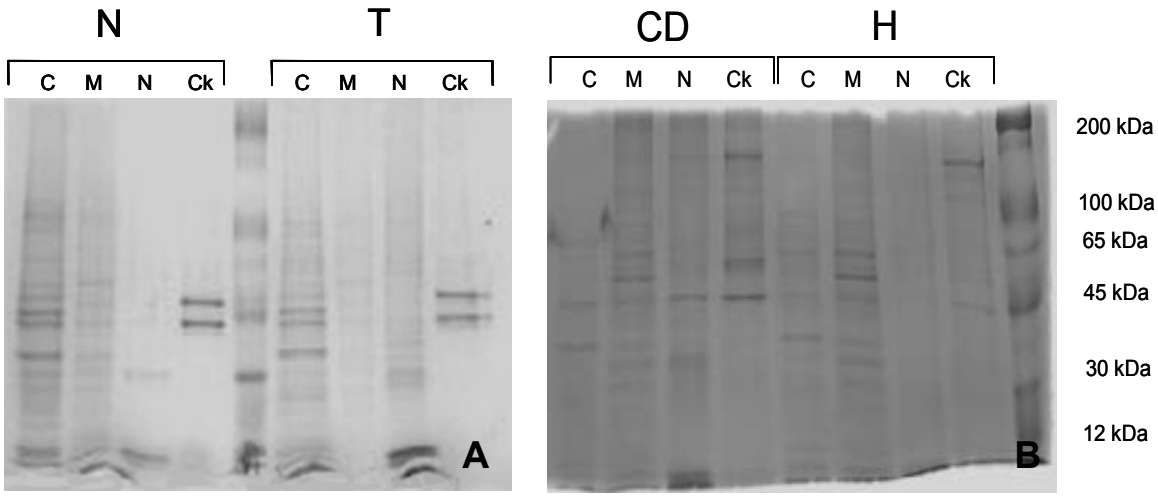

Figure 3.4A: monodimensional gels of cytosolic (C), membrane (M), nuclear (N) and cytoskeletal (Ck) cells subfractions. Part A.: cell compartments of intestinal HT29 cells normal (N) and treated with cytokines(T). The middle lane is the marker. Part B.: cells compartments of IECS from healthy donors (H) and Crohn's disease patients (CD). The right lane is the marker.

The triplicate analysis of every digested sample, followed by the comparison of the LC-MS runs of corresponding slices for the two categories of sample and by the identification of the selected proteins by MS/MS permitted to identify proteins involved in the inflammation both in the cytosolic and in the nuclear fractions of treated HT29 (table 3.4).

Most of the differences were found in treated cells as expected, because of the higher number of proteins produced compared to the ones produced during the basal activity. The most interesting upregulated protein is adenosylhomocysteinase (SAHH), a protein that may play a key role in the control of methylations via regulation of the intracellular concentration of adenosylhomocysteine, a competitive inhibitor of S-adenosyl-L-methionine-dependent methyl transferase reactions. The implication of SAHH in inflammation processes in epithelial cells was already reported in previous works from other groups, that brought to its choice as a specific target for design of immunosuppressive and anti-inflammatory agents [104]. It has also been found to be involved in the aldosterone-induced activity of epithelial Na^+ channels [105].

The nuclear histone H2A type 1-B and the heterogeneous nuclear ribonucleoprotein C-like 1 were also found overexpressed.

HT29

	Protein	Entry name	Mass (Da)	Prot. score	Peptide sequences	m/z	Fold change
T	Histone H2A type 1-B	H2A1B_HUMAN	14127	112	R.AGLQFPVGR.V	472.62	7.98
					R.VTIAQGGVLPNIQAVLLPK.K	965.71	15.45
	Heterogeneous nuclear ribonucleoprotein C-like 1	HNRCL_HUMAN	32180	68	R.VFIGNLNTLVVK.K	658.73	20.25
	Actin, cytoplasmic 1	ACTB_HUMAN	42052	61	K.IWHHTFYNELR.V	505.91	6.76
	Adenosyl-homocysteinase	SAHH_HUMAN	48255	58	K.WLNENAVEK.V	551.77	2.61
					K.VPAINVNDSVTK.S	628.84	1.86
N	Heterogeneous nuclear ribonucleoproteins A2/B1	ROA2_HUMAN	37464	39	R.GGGGNFGPGPGSNFR.G	689.11	2.33

Table 3.4: list of the upregulated proteins and the corresponding peptides in treated (T) and normal (N) HT29 cell lines.

The same protocol was applied to isolated IECs both from healthy donors and Crohn's disease surgical specimens, subjected to subcellular fractionation as described above. The proteins were concentrated by acetone precipitation, but due to the lower number of cells compared to HT29 ($4x10^6$ instead of $5x10^6$) and to the higher number of dead cells that avoided to obtain an optimum yield of compartment subfractionation, a lower amount of proteins was extracted for every fraction. For this reason only 5 µg of proteins for each compartment were loaded on the gel for the separation (figure 3.4B), and every lane was divided only into two slices.

The results obtained by LC-MS and the following LC-MS/MS analysis allowed the identification of many proteins (table 3.5) involved in the inflammatory processes. In the nuclear compartment Annexin A1 (ANXA1), a protein also known as Lipocortin 1 which regulates phospolipase A2 activity, was found upregulated in healthy donors (13.3 fold-change). The involvement of ANXA1 in Crohn's disease is known since many years, indeed its presence or absence can be used for the discrimination among two clinical forms of IBDs (perforating and non-perforating respectively) [106]. Furthermore lipocortin 1 is a mediator of the anti-inflammatory actions of glucocorticoids, but the mechanism of its anti-inflammatory effects is still not clear [107,108]. It has been proved to regulate the MAPK pathway too [109] by inhibition of the interleukine-6 (IL-6), a mediator of inflammation.

Intestinal epithelial cells

	Protein	Entry name	Mass (Da)	Prot. score	Peptide sequences	m/z	fold change
CD	Albumine	ALBU_HUMAN	71317	52	K.CCTESLVNR.R	569.76	4.3
					K.VPQVSTPTLVEVSR.N	756.44	2.2

	DEP domain-containing protein 1B	DEP1B_HUMAN	62472	35	K.FIIHNVYSVSK.Q	653.37	5.2
	Cytochrome P450 2J2	CP2J2_HUMAN	57859	35	K.FTFRPPNNEK.L	625.32	15.2
					K.WPWQVSLR.V	536.37	8.3
					K.YHLGAYTGDDVR.I	683.89	5.9
	Tryptase alpha-1 precursor	TRYA1_HUMAN	31209	247	R.VTYYLDWIHHYVPK.K	917.63	23.8
					R.EQHLYYQDQLLPVSR.I	630.41	125.0
					R.EQHLYYQDQLLPVSR.I	945.13	53.0
					K.ASGPPVSELITK.A	599.90	3.8
	Histone H1.2	H12_HUMAN	21352	98	R.KASGPPVSELITK.A	663.95	2.7
					K.AAKPKVVKPK.K	532.83	12.7
					R.DAVTYTEHAK.R	567.83	2.7
	Histone H4	H4_HUMAN	11360	89	R.DNIQGITKPAIR.R	663.44	2.7
					K.RISGLIYEETR.G	668.93	2.0
					R.LLLPGELAK.H	477.37	3.3
	Histone H2B type 1-C/E/F/G/I	H2B1C_HUMAN	13811	86	K.ESYSVYVYK.V	569.33	1.8
					K.QVHPDTGISSK.A	584.85	1.9
					R.KESYSVYVYK.V	633.39	5.2
					R.LLLPGELAK.H	477.37	3.3
	Histone H2B type 1-B	H2B1B_HUMAN	13942	53	K.QVHPDTGISSK.A	584.85	1.9
					R.KESYSIYVYK.V	640.39	23.5
	Histone H3-like	H3L_HUMAN	15318	38	R.YRPGTVALR.E	516.84	2.2
	Histone H2AV	H2AV_HUMAN	13501	37	R.GDEELDSLIK.A	559.88	1.6
	Ubiquitin	UBIQ_HUMAN	8560	36	K.ESTLHLVLR.L	534.37	2.1
					K.TITLEVEPSDTIENVK.A	894.61	20.5
	Tetratricopeptide repeat protein 24	TTC24_HUMAN	64041	36	R.GLELLLR.A	407.28	1.7
	Lamin-A/C	LMNA_HUMAN	74380	35	R.LADALQELR.A	514.78	7.6
	ATP synthase subunit beta, mitochondrial precursor	ATPB_HUMAN	56525	127	R.FTQAGSEVSALLGR.I	718.33	82.3
	Protein disulfide-isomerase A6 precursor	PDIA6_HUMAN	48490	56	R.TGEAIVDAALSALR.Q	693.85	1.8
					K.AGFAGDDAPR.A	488.72	3.1
	Actin, cytoplasmic 1	ACTB_HUMAN	42052	37	R.AVFPSIVGRPR.H	599.85	2.6
					K.DSYVGDEAQSKR.G	677.81	12.4
					R.VAPEEHPVLLTEAPLNPK.A	652.00	8.2
	78 kDa glucose-regulated protein precursor	GRP78_HUMAN	72402	37	R.ITPSYVAFTPEGER.L	783.83	37.0
					R.IINEPTAAAIAYGLDKR.E	605.97	11.5
	Hemoglobin subunit alpha	HBA_HUMAN	15305	49	K.FLASVSTVLTSK.Y	626.86	9.4
					K.VGAHAGEYGAEALER.M	765.35	3.9
H	Malate dehydrogenase, mitochondrial precursor	MDHM_HUMAN	35969	58	K.GCDVVVIPAGVPR.K	669.92	15.7
					K.TIIPLISQCTPK.V	685.96	6.1
	Annexin A1	ANXA1_HUMAN	38918	82	K.GLGTDEDTLIEILASR.T	851.99	13.3
	Dermcidin precursor	DCD_HUMAN	11277	27	K.GAVHDVKDVLDSVL.	733.94	3.1
	Histone H3-like	H3L_HUMAN	15318	28	R.YRPGTVALR.E	516.87	4.0

Table 3.5: list of upregulated proteins identified in healthy (H) and Crohn's disease (CD) intestinal epithelial cells (IECs). For each peptide the sequence, m/z value and fold-change calculated is reported.

Malate dehydrogenase (MDHM), an enzyme in the citric acid cycle that catalyzes the conversion of malate into oxaloacetate, has been found overexpressed in healthy donors. This data is in agreement with other previously published results [110].

Among the more interesting upregulated proteins in Crohn's disease nuclear fraction, different histones (H1.2, H4, H2B type 1-C/E/F/G/I, H2B type 1-B, H3-like, H2AV), ubiquitin (UBIQ), tryptase alpha-1 precursor (TRYA1), ATP synthase subunit beta (ATPB) and Heat shock 70 kDa protein 5 (HSP70k5) were found. The overexpression of histones can be due to the activation of the transcription of the p21.3 tandem region of chromosome 6, while the upregulation of TRYA1 (the major neutral protease present in mast cell whose secretion is coupled to the activation-degranulation response of this cell type) can be associated with the increase of the paracellular permeability of the intestine [111]. Since mast cells are present in intestinal mucosa, they can be co-extracted during the enzymatic digestion step of the tissue for isolation of IECs.

Ubiquitin is a protein present in all the cells and tissues that marks damaged or obsolete proteins for destruction by proteasome, and its almost 10 fold-change overexpression is probably not specific for the discrimination among H and CD. However the upregulation of UBIQ can be explained by the increased activity of IECs, also confirmed by the overexpression of ATP synthase subunit beta.

The heat shock proteins (HSP) are a group of proteins whose expression increases after exposure of the cells to stress. HSP70k5 in particular has been demonstrated to be involved in the protection of intestinal epithelial [112], hence its upregulation in epithelial cells was expected.

3.4 Conclusions

The developed label-free LC ESI-QTOF MS comparative analysis approaches were proved to be useful for the study of protein expression profiles and the identification of the differently expressed proteins in serum samples and intestinal epithelial cells from Crohn's disease patients.

The combining of the LC-MS strategy with the low molecular weight serum proteins enrichment step allowed to evaluate a pattern of peptides derived from exoprotease activity in the coagulation and complement activation pathways. Among the obtained

results, particularly interesting was the discovery of eight peptides from fibrinopeptide A (FPA) that were found even 700-fold upregulated, thus meaning the overexpression of both FPA and of a specific exoprotease that degraded it. Since other research group found the same cluster of peptides involved also in breast, ladder and prostate cancers but with a different pattern than the one found in Crohn's disease, further study need to be performed to evaluate the specificity of the cluster and validate the results, in order to develop a fast serum diagnostic test.

The study of subcellular fractions of intestinal epithelial cells from Crohn's disease patients and healthy donors permitted to find many proteins involved in the inflammation process, such as heat shock protein 70 and tryptase alpha-1 precursor, and proteins whose upregulation can be explained by the increased activity of IECs in Crohn's disease. Follow-up studies based on immunochemical strategies will be performed to validate the obtained results and to further investigate the inflammation pathways involved in the disease.

General conclusions

The study of protein expression profiles in serum and in population of cells needs the continuous improvement and combination of proteins/peptides separation techniques, mass spectrometry, statistical and bioinformatic approaches. In this thesis work two different mass spectrometry-based protein profiling strategies have been developed and applied to liver and inflammatory bowel diseases for new biomarker discovery.

The bulk solid-phase extraction combined with MALDI-TOF MS and chemometric analysis allowed to enrich serum protein/peptides due the high interaction surface between analytes and functionalized groups and the high recovery due to the elution step performed directly on the MALDI-target plate. The use of chemometric algorithm for the selection of the variables with higher discriminant power permitted to evaluate patterns of 20-30 proteins involved in the differentiation and classification of serum samples from healthy donors and diseased patients. These proteins profiles permit to discriminate among the pathologies with an optimum classification and prediction abilities. In particular in the study of inflammatory bowel diseases, after the analysis using C18 of 129 serum samples from healthy donors and Crohn's disease, ulcerative colitis and inflammatory controls patients, a 90.7% of classification ability and a 72.9% prediction ability were obtained. In the study of liver diseases (hepatocellular carcinoma, viral hepatitis and cirrhosis) a 80.6% of prediction ability was achieved using IDA-Cu(II) as extraction procedure.

The identification of the selected proteins by MALDI-TOF/TOF MS analysis or by their selective enrichment followed by enzymatic digestion and MS/MS analysis may give useful information in order to identify new biomarkers involved in the diseases.

The search for biomarkers of inflammatory bowel diseases, and in particular of Crohn's disease, was also performed by using the developed label-free LC ESI-QTOF MS differential analysis strategy combined with targeted MS/MS analysis of only identified differences. The enriched serum peptidome and the subcellular fractions of intestinal epithelial cells (IECs) from healthy donors and Crohn's disease patients were analysed.

The combining of the low molecular weight serum proteins enrichment step and the LC-MS approach allowed to evaluate a pattern of peptides derived from specific exoprotease activity in the coagulation and complement activation pathways. Among

these peptides, particularly interesting was the discovery of clusters of peptides from fibrinopeptide A, Apolipoprotein E and A4, and complement C3 and C4. Further studies need to be performed to evaluate the specificity of these clusters and validate the results, in order to develop a rapid serum diagnostic test.

The analysis by label-free LC ESI-QTOF MS differential analysis of the subcellular fractions of IECs from Crohn's disease patients and healthy donors permitted to find many proteins that could be involved in the inflammation process. Among them heat shock protein 70, tryptase alpha-1 precursor and proteins whose upregulation can be explained by the increased activity of IECs in Crohn's disease were identified. Follow-up studies for the validation of the results and the in-depth investigation of the inflammation pathways involved in the disease will be performed.

Both the developed mass spectrometry-based protein profiling strategies have been proved to be useful tools for the discovery of disease biomarkers that need to be validated in further studies.

REFERENCES

1. **Proteomics' New Order,** *Nature* 2005; **417**: 169-170.
2. Giot L, Rothberg JM *et al*., **A protein interaction map of Drosophila melanogaster**, *Science* 2003; **302**: 1727-1736.
3. O'Farrel PH. **High resolution two dimensional electrophoresis for proteins**, *J. Biol. Chem.* 1975; **250**: 4007-4021.
4. Wolters DA, Washburn MP, Yates III, JR. **An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics**, *Anal. Chem.* 2001; **73**: 5683-5690.
5. Ross PL, Huang YL, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. **Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents**, *Mol. Cell. Proteomics* 2004; **3**:1154–1169.
6. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**, *Nat. Biotechnol.* 1999; **17**: 994–999.
7. Han DK, Eng J, Zhou H, Aebersold R. **Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry**, *Nat. Biotechnol*. 2001; **19**: 946–951.
8. Aebersold R, Mann M. **Mass spectrometry-based proteomics**, *Nature* 2003; **422**: 198–207.
9. Yates JR. **Mass spectral analysis in proteomics**, *Annu. Rev. Biophys. Biomol. Struct.* 2004; **33**: 297–316.
10. Cravatt BF, Simon GM, Yates III JR. **The biological impact of mass-spectrometry-based proteomics**, *Nature* 2007; **450**: 991-1000.
11. Domon B, Aebersold R. **Mass spectrometry and protein analysis**, *Science* 2006; **312**: 212–217.
12. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. **Electrospray ionization for mass spectrometry of large biomolecules**, *Science* 1989; **246**: 64-71.
13. Karas M, Hillenkamp F. **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**, *Anal. Chem.* 1988; **60:** 2299-2301.

14. Westermeier R, Naven T. **Proteomics in practice: a laboratory manual of proteome analysis**, Ed. Wiley, 2004: 115.
15. Wilm M, Mann M. **Analytical properties of the nanoelectrospray ion source**, *Anal. Chem.* 1996; **68**: 1-8.
16. Aebersold R, Goodlet DR**. Mass spectrometry in proteomics**, *Chem. Rev.* 2001; **101**: 269-295.
17. Aebersold R, Mann RM. **Mass spectrometry-based proteomics**, *Nature* 2003; **422**: 198-207.
18. Von Haller PD, Donohoe S, Goodlett DR, Aebersold R, Watts JD. **Mass spectrometric characterization of proteins extracted from Jurkat T cell detergent-resistant membrane domains**, *Proteomics* 2001; **1**: 1010-1021.
19. Mamyrin BA. **Time-of-flight mass spectrometry (concepts, achievements, and prospects)**, *Int. J. Mass Spectrom*. 2001; **206**: 251-266.
20. Baldwin MA. **Protein Identification by Mass Spectrometry: Issues to be Considered.** *Mol. Cell. Proteomics* 2004; **3**: 1-9.
21. Carr S, Aebersold R, Baldwin M, Burlingame A, Clauser K, Nesvizhskii A. **The Need for Guidelines in Publication of Peptide and Protein Identification Data: Working Group On Publication Guidelines For Peptide And Protein Identification Data**, *Mol. Cell. Proteomics* 2004; **3**: 531-533.
22. Forina M, Lanteri S, Armanino C, Oliveros MCC, Casolino C. ***V-PARVUS*** 2004. Freely available at http://www.parvus.unige.it
23. Bellew M, Coram M, Fitzgibbon M, Igra M, Randolph T, Wang P, May D, Eng J, Fang R, Lin CW, Chen J, Goodlett D, Whiteaker J, Paulovich A, McIntosh M. **A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS**, *Bioinformatics* 2006; **22**: 1902–1909. Free available at http://proteomics.fhcrc.org
24. Gärdén P, Alm R, Häkkinen J. **PROTEIOS: an open source proteomics initiative**, *Bioinformatics* 2005; **21**: 2085-2087.
25. Liotta LA, Petricoin EF. **Putting the "Bio" back into Biomarkers: Orienting Proteomic Discovery toward Biology and away from the Measurement Platform,** *Clinical Chemistry* 2008; **54**: 3-5.
26. Price A. **Overlap in the spectrum of non-specific inflammatory bowel disease–'colitis indeterminate**, *J. Clin. Pathol.* 1978; **31**: 567–577.
27. Meuwis MA, Fillet M, Geurts P, de Seny D, Lutteri L, Chapelle JP, Bours V, Wehenkel L, Belaiche J, Malaise M, Louis E, Merville MP. **Biomarker**

discovery for inflammatory bowel disease, using proteomic serum profiling, *Biochem. Pharmacol.* 2007; **73**:1422-1433.

28. Shkoda A, Werner T, Daniel H, Gunckel M, Rogler G, Haller D. **Differential protein expression profile in the intestinal epithelium from patients with inflammatory bowel disease**, *J. Proteome Res*. 2007; **6**: 1114-1125.
29. Chignard N, Beretta L. **Proteomics for hepatocellular carcinoma marker discovery**. *Gastroenterology* 2004; **127**: S120-125.
30. Villanueva A, Newell P, Chiang DY, Friedman SL, Llovet JM. **Genomics and signaling pathways in hepatocellular carcinoma**, *Semin. Liver Dis.* 2007; **27**: 55-76.
31. Steel LF, Trotter MG, Nakajima PB, Mattu TS, Gonye G, Block T. **Efficient and specific removal of albumin from human serum samples**, *Mol. Cell. Proteomics.* 2003; **2**: 262-270.
32. Luque-Garcia JL, Neubert TA. **Sample preparation for serum/plasma profiling and biomarker identification by mass spectrometry**, *J. Chromatogr. A* 2007; **1153**: 259-276.
33. Gilar M, Bouvier ES, Compton BJ. **Advances in sample preparation in electromigration, chromatographic and mass spectrometric separation methods**, *J. Chromatogr. A* 2001; **909**: 111-135.
34. Pyrzynska K, Pobozy E. **On-Line Coupling of Solid Phase Extraction Sample Processing with High-Performance Liquid Chromatography**, *Crit. Rev. Anal. Chem.* 2002; **32**: 227-243.
35. Naxing Xu R, Fan L, Rieser MJ, El-Shourbagy TA. **Recent advances in high-throughput quantitative bioanalysis by LC–MS/MS**, *J. Pharm. Biomed. Anal.* 2007; **44:** 342-355.
36. Zhang S, Van Pelt CK, Henion JD. **Automated chip-based nanoelectrospray-mass spectrometry for rapid identification of proteins separated by two-dimensional gel electrophoresis**, *Electrophoresis* 2003; **24**: 3620-3632.
37. Callesen AK, Mohammed S, Bunkenborg J, Kruse TA, Cold S, Mogensen O, Christensen RP, Vach W, Jørgensen PE, Jensen ON. **Serum protein profiling by miniaturized solid-phase extraction and matrix-assisted laser desorption/ionization mass spectrometry**, *Rapid Commun. Mass Spectrom*. 2005; **19**: 1578-1586.

38. Miyazaki S, Morisato K, Ishizuka N, Minakuchi H, Shintani Y, Furuno M, Nakinishi K. **Development of a monolithic silica extraction tip for the analysis of proteins**, *J Chromatog.* 2004; **1043**: 19-25.
39. Nissum M, Schneider U, Kuhfuss S, Obermaier C, Wildgruber R, Posch A, Eckerskorn, C. **In-gel digestion of proteins using a solid-phase extraction microplate**, *Anal. Chem*. 2004; **76**: 2040-2045.
40. Ekstrom S, Wallman L, Hok D, Marko-Varga G, Laurell T. **Miniaturized solid-phase extraction and sample preparation for MALDI MS using a microfabricated integrated selective enrichment target**, *J. Proteome Res.* 2006; **5**: 1071-1081.
41. Tang N, Tornatore P, Weinberger SR. **Current developments in SELDI affinity technology**, *Mass Spectrom. Rev.* 2004, **23**: 34-44.
42. Xu Y, Bruening ML, Watson JT. **Non-specific, on-probe cleanup methods for MALDI-MS samples**, *Mass Spectrom. Rev.* 2003; **22**: 429-420.
43. Brockman AH, Shah NN, Orlando R. **Optimization of a hydrophobic solid-phase extraction interface for matrix-assisted laser desorption/ionization,** *J. Mass Spectrom.* 1998; **33**: 1141-1147.
44. Beaven SW, Abreu MT. **Biomarkers in inflammatory bowel disease**, *Curr. Opin. Gastroen.* 2004; **20**: 318-327.
45. Abreu MT. **Serologies in Crohn's disease: can we change the gray zone to black and white?** *Gastroenterology* 2006; **131**: 664-667.
46. Gevaert K, Demol H, Puype M, Broekart D, De Boeck S, Houthaeve T, Vandekerckhove J. **Peptides adsorbed on reverse-phase chromatographic beads as targets for femtomole sequencing by post-source decay matrix assisted laser desorption ionization-reflectron time of flight mass spectrometry (MALDI-RETOF-MS)**, *Electrophoresis* 1997; **18:** 2950-2960.
47. Gevaert K, Eggermont L, Demol H, Vandekerckhove J. **A fast and convenient MALDI-MS based proteomic approach: identification of components scaffolded by the actin cytoskeleton of activated human thrombocytes**, *J. Biotechnol.* 2000; **78**: 259-269.
48. Rainer M, Muhammad NU, Huck CW, Feuerstein I, Bakry R, Huber LA, Gjerde DT, Zou X, Qian H, Du X, Wei-Gang F, Ke Y, Bonn GK. **Ultra-fast mass fingerprinting by high-affinity capture of peptides and proteins on derivatized poly(glycidyl methacrylate/divinylbenzene) for the analysis of**

serum and cell lysates, *Rapid Commun. Mass Spectrom.* 2006; **20**: 2954-2960.

49. Feuerstein I, Rainer M, Bernardo K, Stecher G, Huck CV, Kofler K, Pelzer A, Horninger W, Klocker H, Bartsch G, Bonn GK. **Derivatized cellulose combined with MALDI-TOF MS: a new tool for serum protein profiling**, *J. Proteome Res.* 2005; **4**: 2320-2326.
50. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PL, Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics: Part A*. Elsevier: Amsterdam, 1998; 168.
51. Massart DL, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PL, Smeyers-Verbeke J. *Handbook of Chemometrics and Qualimetrics: Part B*. Elsevier: Amsterdam, 1998; 213.
52. Forina M, Lanteri S, Casale M, Oliveros MCC. **Stepwise orthogonalization of predictors in classification and regression techniques: an "old" technique revisited**, *Chemometr. Intell. Lab.* 2007; **87**: 252-261.
53. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z,, Wright jGL. **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men**, *Cancer Res.* 2002; **62**: 3609-3014.
54. Norris JL, Cornett DS, Mobley JA, Andersson M, Seeley EH, Chauran P, Caprioli RM. **Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis**, *Int. J. Mass Spectrom*. 2007; **260**: 212-221.
55. Tatay JW, Feng X, Sobczak N, Jiang H, Chen CF, Kirova R, Struble C, Wang NJ, Tonellato PJ. **Multiple approaches to data-mining of proteomic data based on statistical and pattern classification methods**, *Proteomics* 2003; **3**:1704-1709.
56. Wagner M, Naik D, Pothen A. **Protocols for disease classification from mass spectrometry data**, *Proteomics* 2003; **3**: 1692-1698.
57. Slotta DJ, Heath LS, Ramakrishnan N, Helm R, Potts M. **Clustering mass spectrometry data using order statistics**, *Proteomics* 2003; **3**: 1687-1691.
58. Lee KR, Lin X, Park DC, Eslava S. **Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method**, *Proteomics* 2003; **3**: 1680-1686.

59. Markey MK, Tourassi GD, Floyd CE. **Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer**, *Proteomics* 2003; **3**: 1678-1679.
60. Zhu H, Yu CY, Zhang HP. **Tree-based disease classification using protein data**, *Proteomics* 2003; **3**: 1673-1677.
61. Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, Coombes KR. **A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples**, *Proteomics* 2003; **3**: 1667-1672.
62. Forina M, Lanteri S, Oliveros MCC, Millan CP. **Selection of useful predictors in multivariate calibration**, *Anal. Bioanal. Chem*. 2004; **380**: 397-418.
63. Ragazzi E, Pucciarelli S, Deraglia R, Molin L, Agostini M, Lise M, Traldi P, Nitti D. **Multivariate analysis approach to the plasma protein profile of patients with advanced colorectal cancer**, *J. Mass Spectrom* 2006; **41**: 1546-1553.
64. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol. Cell. Proteomics* 2002; **1**: 376-386.
65. Wåhlander Å, Arrigoni G, Wårell K, Levander F, Palmgren R, Maloisel JL, Busson P, James P. **Development of reagents for differential protein quantitation by subtractive parent (precursor) in scanning**, *Proteome Res.* 2007*;* **6:** 1101-1113.
66. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. **Comparison of Label-free Methods for Quantifying Human Proteins by Shotgun Proteomics**, *Mol. Cell. Proteomics* 2005; **4**: 1487–1502.
67. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. **Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics,** *J. Proteome Res.* 2006; **5**: 2909- 2918.
68. Wiener MC, Sachs JR, Deyanova EG, Yates NA. **Differential Mass Spectrometry: A Label-Free LC-MS Method for Finding Significant Differences in Complex Peptide and Protein Mixtures**, *Anal. Chem*. 2004; **76**: 6085-6096.

69. Zheng X, Baker H, Hancock WS. **Analysis of the low molecular weight serum peptidome using ultrafiltration and a hybrid ion trap-Fourier transform mass spectrometer**, *J. Chromatogr. A* 2006, **1120**: 173–184.
70. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. **Characterization of the Low Molecular Weight Human Serum Proteome**, *Mol. Cell. Proteomics* 2003; **2**: 1096–1103.
71. Mayer L, Shlien R. **Evidence for function of Ia molecules on gut epithelial cells in man**, *J. Exp. Med*. 1987, **166**:1471-1483.
72. http://sashimi.sourceforge.net
73. Coombes KR, Morris JS, Hu J, Edmonson SR, Baggerly KA. **Serum proteomics profiling-a young technology begins to mature**, *Nat. Biotechnol.* 2005, **23**: 291–292.
74. Diamandis EP. **Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations**, *Mol. Cell. Proteomics.* 2004; **3**: 367–378.
75. Check E. **Proteomics and cancer: running before we can walk?** *Nature* 2004; **429**: 496–497.
76. Ransohoff DF. **Opinion: bias as a threat to the validity of cancer molecular-marker research**, *Nat. Rev. Cancer.* 2005; **5**:142–149.
77. Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P. **Correcting common errors in identifying cancer-specific serum peptide signatures**, *J. Proteome Res*. 2005; **4**:1060–1072.
78. Diamandis EP. **Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics?**, *Clin. Chem.* 2003; **49**:1272–1275.
79. Sorace JM, Zhan M. **A data review and re-assessment of ovarian cancer serum proteomic profiling**, *BMC Bioinformatics* 2003; **4**: 24-36.
80. Liotta LA, Petricoin EF. **Serum peptidome for cancer detection: spinning biologic trash into diagnostic gold**; *J. Clin. Invest*. 2006; **116**: 26-30.
81. Liotta LA, Ferrari M, Petricoin E. **Clinical proteomics: written in blood**; *Nature* 2003; **425**: 905.
82. Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, Fleisher M, Lilja H, Brogi E, Boyd J, Sanchez-Carbayo M, Holland EC, Cordon-Cardo C, Scher HI, Tempst P. **Differential exoprotease activities**

confer tumor-specific serum peptidome patterns, *J. Clin. Invest.* 2006; 116:271–284.

83. Fu Q, Bovenkamp DE, Van Eyk JE. **A rapid, economical, and reproducible method for human serum delipidation and albumin and IgG removal for proteomic analysis**; *Methods Mol. Biol.* 2007; **357**: 365-371.
84. Dardé VM, Barderas MG, Vivanco F. **Depletion of high-abundance proteins in plasma by immunoaffinity subtraction for two-dimensional difference gel electrophoresis analysis**, *Methods Mol. Biol.* 2007; **357**: 351-364.
85. Dekker LJ, Bosman J, Burgers PC, van Rijswijk A, Freije R, Luider T, Bischoff R. **Depletion of high-abundance proteins from serum by immunoaffinity chromatography: a MALDI-FT-MS study**, *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 2007; **847**: 65-69.
86. Gao M, Zhang J, Deng C, Yang P, Zhang X. **Novel strategy of high-abundance protein depletion using multidimensional liquid chromatography**, *J. Proteome Res.* 2006; **5**: 2853-2860.
87. Zolotarjova N, Martosella J, Nicol G, Bailey J, Boyes BE, Barrett WC. **Differences among techniques for high-abundant protein depletion**, *Proteomics* 2005; **5**: 3304–3313.
88. Huang L, Harvie G, Feitelson JS, Gramatikoff K, Herold DA, Allen DL, Amunngama R, Hagler RA, Pisano MR, Zhang WW, Fang X. **Immunoaffinity separation of plasma proteins by IgY microbeads: meeting the needs of proteomic sample preparation and analysis**, *Proteomics* 2005; **5**: 3314–3328.
89. Chromy BA, Gonzales AD, Perkins J, Choi MW, Corzett MH, Chang BC, Corzett CH, McCutchen-Maloney SL. **Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins**, *J. Proteome Res.* 2004; **3**: 1120–1127.
90. Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW. **Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma**, *Proteomics* 2005; **5**: 3292–3303.
91. Yocum AK, Yu K, Oe T, Blair IA. **Effect of immunoaffinity depletion of human serum during proteomic investigations**, *J. Proteome Res.* 2005; **4**:1722–1731.

92. Brand J, Haslberger T, Zolg W, Pestlin G, Palme S. **Depletion efficiency and recovery of trace markers from a multiparameter immunodepletion column**, *Proteomics* 2006; **6**: 3236–3242.
93. Zhou M, Lucas DA, Chan KC, Issaq HJ, Petricoin EF 3rd, Liotta LA, Veenstra TD, Conrads TP. **An investigation into the human serum "interactome"**, *Electrophoresis* 2004; **25**: 1289–1298.
94. Petricoin EF, Liotta LA. **SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer**, *Curr. Opin. Biotechnol.* 2004; **15**: 24–30.
95. Liu T, Qian WJ, Mottaz HM, Gritsenko MA, Norbeck AD, Moore RJ, Purvine SO, Camp II DG, Smith RD. **Evaluation of Multiprotein Immunoaffinity Subtraction for Plasma Proteomics and Candidate Biomarker Discovery Using Mass Spectrometry**. *Mol. Cell. Proteomics* 2006; **5**: 2167–2174.
96. Zheng X, Baker H, Hancock WS. **Analysis of the low molecular weight serum peptidome using ultrafiltration and a hybrid ion trap-Fourier transform mass spectrometer**, *J. Chromatogr. A* 2006; **1120**: 173–184.
97. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. **Characterization of the Low Molecular Weight Human Serum Proteome**, *Mol. Cell. Proteomics* 2003; **2**: 1096–1103.
98. Vowinkel T, Mori M, Krieglstein CF, Russell J, Saijo F, Bharwani S, Turnage RH, Davidson WS, Tso P, Granger DN, Kalogeris TJ. **Apolipoprotein A-IV inhibits experimental colitis**, *J. Clin. Invest.* 2004; **114**: 260-269.
99. Broedl UC, Schachinger V, Lingenhel A, Lehrke M, Stark R, Seibold F, Göke B, Kronenberg F, Parhofer KG, Konrad-Zerna A. **Apolipoprotein A-IV is an independent predictor of disease activity in patients with inflammatory bowel disease**; *Inflamm. Bowel Dis.* 2007; **13**: 391-397.
100. Jofre-Monseny L, Loboda A, Wagner AE, Huebbe P, Boesch-Saadatmandi C, Jozkowicz A, Minihane AM, Dulak J, Rimbach G. **Effects of apoE genotype on macrophage inflammation and heme oxygenase-1 expression**, *Biochem. Biophys. Res. Commun.* 2007; **357**: 319–324.
101. Tziakas DN, Chalikias GK, Antonoglou CO, Veletza S, Tentes IK, Kortsaris AX, Hatseras DI, Kaski JC. **Apolipoprotein E genotype and circulating interleukin-10 levels in patients with stable and unstable coronary artery disease**, *J. Am. Coll. Cardiol.* 2006; **48**: 2471–2481.

102. Villanueva J, Nazarian A, Lawlor K,Yi SS, Robbins RJ, Tempst P. **A sequence-specific exopeptidase activity test (SSEAT) for 'functional' biomarker discovery**, *MCP in press*
103. Dahan S, Roth-Walter F, Arnaboldi P, Agarwal S, Mayer L. **Epithelia: lymphocyte interactions in the gut**, *Immunol. Rev*. 2007; **215**: 243-253.
104. Yuan CS, Saso Y, Lazarides E, Borchardt RT and Robins MJ. **Recent advances in S-adenosyl-L-homocysteine hydrolase inhibitors and their potential clinical applications**, *Exp. Opin. Ther. Patents* 1999; **9:** 1197–1206.
105. Stockand JD, Zeltwanger S, Bao HF, BecchettI A, Worrell RT, Eaton DC. ***S*-adenosyl-L-homocysteine hydrolase is necessary for aldosterone-induced activity of epithelial Na^+ channels** *Am. J. Physiol. Cell. Physiol* 2001; **281**: C773–C785.
106. Sakanoue Y, Horai T, Okamoto T, Hatada T, Shoji Y, Fujita S, Kusunoki M, Utsunomiya J. **Lipocortin-present perforating and lipocortin-absent nonperforating Crohn's disease**, *Am. J. Surg*. 1992; **164**: 341-344.
107. Gerke V, Moss SE. **Annexins: from structure to function**, *Physiol. Rev.* 2002; **82**: 331–371.
108. Flower RJ, Rothwell NJ. **Lipocortin-1: cellular mechanisms and clinical significance**, *Trends Pharmacol. Sci.* 1994; **15**: 71–76.
109. Yang YH, Toh ML, Clyne CD, Leech M, Aeberli D, Xue J, Dacumos A, Sharma L, Morand EF. **Annexin 1 Negatively Regulates IL-6 Expression via Effects on p38 MAPK and MAPK Phosphatase-1_1,** *The Journal of Immunology* 2006; **177**: 8148–8153.
110. Hsieh SY, Shih TC, Yeh CY, Lin CJ, Chou YY, Lee YS. **Comparative proteomic studies on the pathogenesis of human ulcerative colitis**, *Proteomics* 2006; **6**: 5322–5331
111. Jacob C, Yang PC, Darmoul D, Amadesi S, Saito T, Cottrell GS, Coelho AM, Singh P, Grady EF, Perdue M, Bunnett NW. **Mast cell tryptase controls paracellular permeability of the intestine. Role of protease-activated receptor 2 and beta-arrestins**, *J. Biol. Chem.* 2005; **280**: 31936-31948.
112. Petrof EO, Ciancio MJ, Chang EB. **Role and regulation of intestinal epithelial heat shock proteins in health and disease**, *Chin. J. Dig. Dis.* 2004; **5**: 45-50.

Publications

- "**Serum protein profiling in patients with inflammatory bowel diseases using selective solid-phase bulk extraction, MALDI-TOF MS and chemometric data analysis**"
  P. Nanni, D. Parisi, G. Roda, M. Casale, A. Belluzzi, E. Roda, L. Mayer, A. Roda
  Rapid Commununication in Mass Spectrometry 2007; **21**: 4142–4148.

- "**A new oral formulation for the release of sodium butyrate in the ileo-cecal region and colon**"
  A. Roda, P. Simoni, M. Magliulo, P. Nanni, M. Baraldini, G. Roda, E. Roda
  World Journal of Gastroenterology - 2007; **13**: 1079-84.

- **"Analysis and classification of bacteria by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry and chemometric approach"**
  D. Parisi, M. Magliulo, P. Nanni, M. Casale, M. Forina, A. Roda
  Analytical and Bioanalytical Chemistry - submitted

- **"A new approach for the study of absorption and endocrine function in the rat intestine"**
  M. Montagnani, M. Tsivian, F. Neri, L. Puviani, I. Mantovani, P. Nanni, A. Marangoni, M. Pariali, P. Simoni, R. Fato, C. Bergamini, S. Leoni, G. Lenaz, B. Nardo, R. Aldini
  Manuscript