UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXV ciclo

Bioinformatics Approaches to MALDI-TOF Mass Spectrometry Data Analysis

Massimiliano Borsani - mat.R08747

Ph.D. Thesis

Scientific tutor: Prof. Giulio Pavesi

Anno accademico: 2011-2012

SSD: BIO/10 (Biochemistry)

Thesis performed at BEACON lab - Bioinformatics, Evolution And COmparative geNomics lab - Department of Biosciences, University of Milano, in collaboration with Departments of Informatics (DISCo) and Experimental Medicine (DIMS), University of Milano-Bicocca.

Cover image: the Ultraflex III^{TM} MALDI-TOF/TOF Mass Spectrometer by Bruker Daltonics.

To my wife, Rosy,
for her patience and support

Contents

List of Figures

List of Tables

Abstract

Despite the increasing performance of Mass spectrometry (MS) and others analytical tools, only few biomarkers have been validated and proved to be robust and clinically relevant; indeed a large numbers of proteomic biomarkers have been described, but they are not yet clinical implemented [1]. MALDI-TOF MS seems one of the more powerful tool for biomarkers discovery [2, 3], and shows interesting clinical properties, for instance the possibility to directly search in peripheral fluids for proteins related to an altered physiological state: samples (urine, plasma, serum, etc.) can be collected easily and cheaply by non-invasive, or very low-invasive, methods [4]. The combination of some biomarkers is actually considered more informative than a single biomarker [5, 6], and the improvement in the bioinformatics analysis of MS data could probably help this investigation, decreasing costs and time necessary for each discovery [7].

It is possible to approach the problems related to the analysis of (MALDI-TOF) MS data in two ways, either trying to increase the number of available samples or by reducing the complexity of the problem [8]: in the first case, we developed an approach to compare small datasets from different sources (i.e. hospitals), based on mu-

tual information and mass spectra alignment, that showed significant performance increase compare to the competing ones tested.

In the latter case, we developed novel methods and approaches to compare MALDI-TOF MS profiles of normal and Renal Cell Carcinoma (RCC) patients, with the goal of isolating the more interesting subset of small proteins and peptides from the whole analysed peptidome. MS-based profiling is in fact able to detect differently expressed proteins or peptides during physiological and pathological processes. Every MALDI-TOF MS *spectrum*, that reports the relative abundance of sample analytes, could be considered as a snapshot of samples peptidome in a definite mass range. The relationship between mass/charge ratio, or m/z, and concentration of detected peptides can be represented by networks. Tumor case and control subjects show different peptidome profiles, due to differences in biomolecular and/or biochemical features of cancer cells: they will show some changes in the networks that describe them. We use graphs to create networks representation of data and to evaluate networks properties. We explore the networks properties comparing cases versus controls datasets, and subdividing cases in the different histological subtypes of RCC, clear cell RCC (ccRCC) and not-ccRCC, using different methods both for networks creation and analysis, and for results evaluation. We identify, for each datasets (controls, ccRCC and not-ccRCC) some interesting mass ranges within which we believe biomarkers signals should be searched.

In conclusion, we have developed a set of methods which we believe improve the current computational approaches for the analysis of mass spectrometry data. These results have been published or presented at workshops and conferences.

Part I

Ph.D. project

Chapter 1

State of the Art

1.1 Mass Spectrometry and Biomarkers discovery

Assessing differences between normal and pathogenic processes is a mainstream topic of biomedical research, particularly in cancer and neurodegenerative treatments, diseases involving several economics and social aspects [7]. Often these differences can be identified through biomarkers [9]. A biomarker is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention" [10].

Only a small number of proteins have been validated as cancer biomarkers in this last years of intensive analysis [8, 11, 12]. The performance improvement of the analytical tools and software involved could probably increase this pool, but, on the other hand, the also

increasing specificity and sensitivity of both analytical methods and hardware raise also the magnitude of biological complexity involved, implying greater difficulty in the discovering of new molecular biomarkers [13, 14]. The use of a single biomarker is now widely recognized to be inadequate and multivariate predictive models combining existing tumor markers improve cancer detection; therefore interest in the search of multiple biomarkers is growing and required [5,6,15]. This view agrees with the new Systems-oriented paradigm of life sciences [13, 14, 16].

Mass spectrometry (MS), an analytical method measuring molecular masses, played an increasing important role in clinical diagnostic during the latter half of the twentieth century [2,3]. Mass spectrometry is also an analytical techniques widely used in different biological studies, because is one of the simplest and most powerful way to identify and characterize biological molecules [17, 18]. To this end proteomics has become an interesting field in the post genomic area and offers the opportunity of large-scale protein analysis in tissues and body fluids. Proteomic pattern diagnostics enables to characterize proteins and functional protein networks as well as their dynamic alteration during physiological and pathological processes and protein profiling with mass spectrometry is a valid approach in the discovery of disease biomarkers [19].

Mass spectrometry data are usually visualized using a plot called *spectrum* (see picture 1.1). *Spectra* from complex biological mixtures are composed by several peaks, sometimes distorted by overlaps [20]. Many chemicals and physicals factors increase spectra complexity,

so also data pre-processing is a crucial step in sample analysis [21]. Computers are mandatory to handle such complexity: bioinformatics, a science promoted to understand the complexity of biological sequences [22], and, generally, the increasing amount of datasets produced by life sciences [23], helps to understand biological problems [24]. Data from projects for biomarkers identification can be processed in different ways, using different models, producing very different results: despite its importance, data modelling started to gain attention only recently [25].

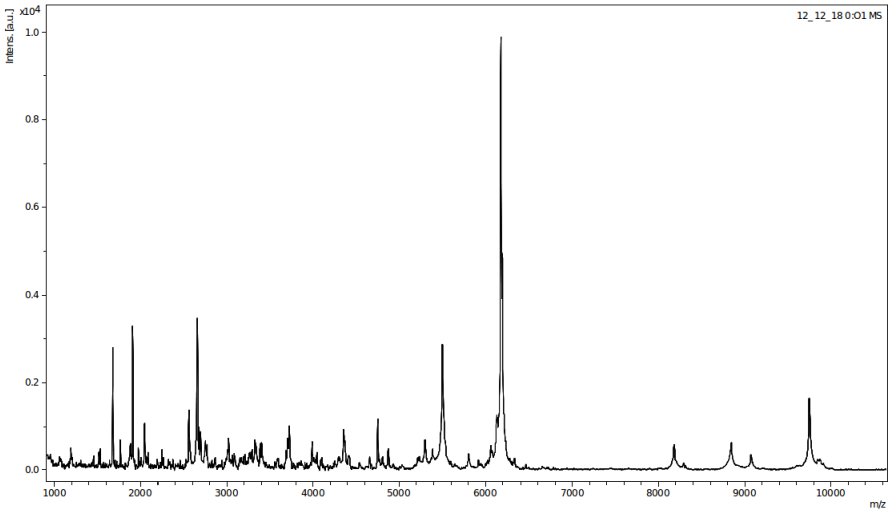


Figure 1.1: A MALDI-TOF *spectrum*. *Mass-to-charge* ratio (m/z) on x-axis, signals intensity on y-axis.

As bioinformaticians, we try to approach Mass Spectrometry-

related problems, developing new methods for data alignment and for biomarkers detection. An MS data alignment approach is discussed in sections 2.1, 3.1, 4.1. All the others sections describe some different methods for MS data analysis.

The first part ("Part I") of the thesis is divided into four chapters: the first chapter describes MALDI-TOF Mass Spectrometry and problems associated to, the guidelines and the key points of our approaches. The second chapter collects the descriptions of datasets and of each single method used in the project. The third and the fourth chapters show respectively the main results and the conclusions. Chapters are basically divided into two parts: one, smaller, relative to a method for the alignment of MS datasets from different labs; the other one, thicker, relative to analysis of MS data.
Thesis "Part II" collects the published papers on the topics discussed, "Part III" papers on different topics, and supplementary materials.

1.2 MALDI-TOF Mass Spectrometry

A mass spectrometer measures the elemental composition of a sample, elucidating the molecules masses and structures. The Matrix-assisted laser desorption/ionization (MALDI; [26]) analysis and characterization of peptides and proteins has been the fastest expanding area, by far, that has resulted from methods for introducing non-volatile compounds into the mass spectrometer [17]. The Time-Of-Flight (TOF; [27,28]) mass analyzer is one of the means for measuring ion masses with MALDI [29]. All the MS-related data analysed in this PhD project were produced using MALDI-TOF Mass Spectrometry: a brief summary of the features of this MS technique could be useful.

1.2.1 MALDI-TOF Mass Spectrometry description

There are different MS technologies, but few of them have a high-resolution capability of the MALDI-TOF [28]). The prominence of MALDI-TOF MS was clearly showed in the the latter half of the twentieth century, in different contexts, like clinical diagnosis [2, 3], post-translational modifications [18, 30, 31], specific and nonspecific proteins interactions [32, 33], and in many other different biological fields [34–36]. Last but not least, the importance of MALDI was also underlined by the 2002 Nobel Prize in Chemistry, co-awarded to John B. Fenn and Koichi Tanaka "for their development of soft desorption ionisation methods for mass spectrometric analyses of biological macromolecules" [www.nobelprize.org].

Despite the impressive range of MALDI-TOF MS applications, also biological-oriented (analysis of peptides, proteins, oligonucleotides, and oligosaccharides, synthetic polymers, etc.), the nature of the MALDI remain poorly understood [37].

In a MALDI experiment, analytes are linked to a matrix (an organic acid), then the resulting mixture, coupled to a metal target support, are fired with a laser beam. The excess of energy accumulated by the affected molecules allows the formation of ions. These ions are accelerated in an uniform electromagnetic field and the time of flight provides information about the accelerated ions obtained. Since the electromagnetic field applies a constant kinetic energy ($2eV$), the same for all ions, the flight times (t) is proportional to the square root of the mass/charge ratio (m/z) [38]:

$$t = \sqrt{\frac{m}{z}\frac{1}{2eV}}D \tag{1.1}$$

Ions with the same m/z ratio have identical kinetic energy, and hit the detector at same time. However, ions from the same analyte could have some additional kinetic energy ($eV + \Delta U_0$), due to the ions plume described below, and/or an initial spacial distribution ($D \neq 0$), so collisions against the detector do not happen exactly at the same time, reducing the resolution.

Moreover, it is possible to overcome the ion energy spread using a reflecting magnetic field, called reflectron [39], based on principles of ion optics: the time of flight of ion packets quitting a decelerating

field, whose potential grows exponentially, does not depend on the initial velocity [40]. This allows to delete the contribution of the ΔU_0, increasing the resolution.

There are several theories about mechanisms of ion production, and probably there is no a single cause for MALDI ionization dynamics [37, 41]; different theories concern both matrix or analyte molecules. Usually MALDI ionization is described as a *plume*, a very rapid, even explosive, solid-to-gas phase transition. An additional pool of phenomena probably act together (see [42–44]).

Although the physics phenomenon mechanism is not clear, MALDI shows extraordinary robustness, high speed and relative immunity to contaminants, bio-chemical buffers, and common additives [45].

1.2.2 Sources of variability

The key issue of this kind of analysis is to understand and to manage the variability of the datasets analysed [8]. There are two main different reasons for variability:

- biological variability;
- intra subject variability,

both affected by biological and experimental noise, that highly increase the complexity of the landscape observed [46]. Understanding

MALDI-TOF MS strength and weakness helps comprehend our approaches and ideas [47].

A MALDI-TOF Mass Spectrometer works recording the number of ion collisions against a detector. Data are collected dividing flight time in small bins: ions detected in the same bin are considered related, with high probability, to the same molecule. The relative concentration of the chemical compounds detected is calculated comparing the ratio between the number of hits in the same bin and the total hits.

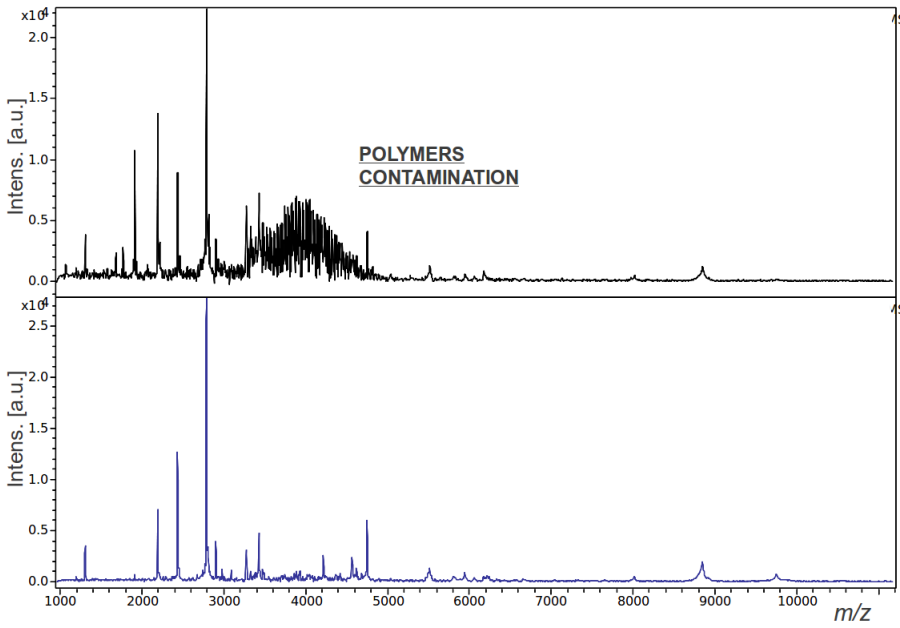


Figure 1.2: A spectrum distorted by chemical background noise, due to the disturbance produced by polymers contamination.

Several issues and shortcomings are related to this procedure, like: chemical and baseline noises, mass-dependent sensitivity, chemical adducts, fragmentation of big proteins, reproducibility, calibration, and ion suppression effects [8]. Some ***chemical background noise*** is due to the disturbance produced by the matrix molecules, but this usually happens only for low m/z values; some is due to molecules derived from sample preparation, like trypsin, keratin and polymers from disposable material: see figure 1.2.2. An additional source of noise concerns the ***anomalous baseline level***, a distortion due to low mass molecules, some of the same causing chemical noise (see figure 1.2.2).

A great issue in MALDI-TOF data analysis is that very large proteins could give more than one signal: the protein is so large that spontaneous ***protein fragmentations*** happen, also in mild conditions. Original signal is split in different signals, one per fragment, related to his new m/z, so the original information (the m/z of the precursor protein) must be reconstructed from the different signals. This principle is used for peptide mass fingerprint (PMF), a technique useful for protein identification [48]: unfortunately, unlike PMF, that employs restriction enzyme digestions, in a MS run the polypeptide chain of large proteins break in random-like way.
Signals shifts could also happen due to chemical adduct ions (salt, solvent, or matrix ions) that could be carried by large (unbroken) proteins.

Like many experimental technologies, MALDI gives some variability due to different *apparata* involved (lasers, quality of the ma-

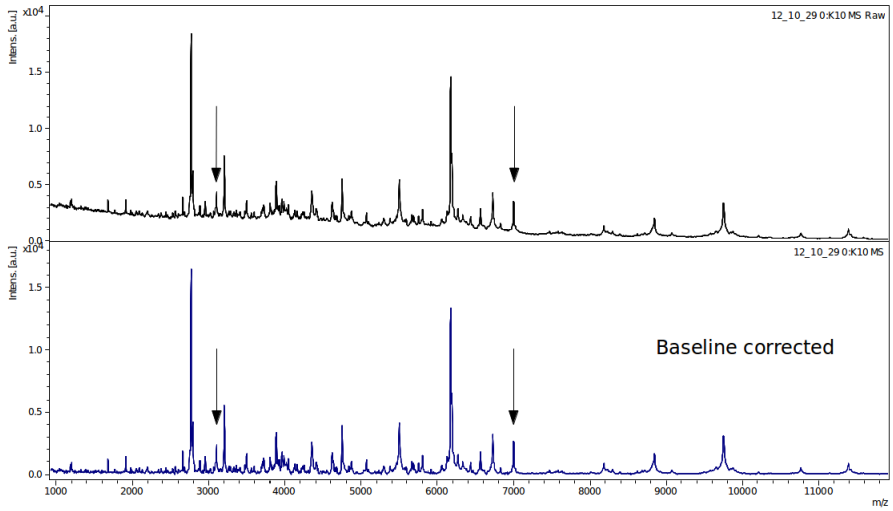


Figure 1.3: Comparison of the same spectrum, before and after baseline correction.
The two arrows identify the same peaks: to be note that the peak on the left, initially more intense, become less intense than the right one after baseline correction.

trix, sample preparation). ***Reproducibility*** could be improved using commercial kits and standardized procedures, like standardized sample collecting. We employed a solid-phase extraction technique based on an off-line fractionation of the proteome present in biological fluids using magnetic beads with activated surface, before MS analysis, i.e. ClinProt^{TM} by Bruker Daltonics (Germany): every step of sample preparation and proteins extraction is based on kit protocols, and every aspect of the procedure was developed by the same company, from sample purification to mass spectrometer. This procedures support

more consistent results than using heterogeneous methods and equipments. Also mass spectrometer ***calibration*** yields similar problems, and, again, the use of standardized calibration methods helps to increase the reproducibility of the experiments.

In MALDI-TOF profiling technique, proteins and peptides concentration is a relative measures. If the signal intensity of a protein (peptide) is too strong compared to other analytes, some ***minor signals suppressions*** could happen: in other words, the signal intensity is not always linear and the suppression is non-homogeneous for different peptides but homogeneous for the same peptide [49].

In addition, some different variables can influence the outcome of MALDI ***mass spectrum analysers***: time scales, acceleration fields, sample temperatures, incident angles of the laser beam, laser wavelengths, pulse energies, and pulse widths: see, for instance [41].

1.2.3 Data overview

Each mass spectrum is composed by the intensity values of thousand of different masses or, to be more accurate, m/z ratios: each m/z on the x-axis is associated with a relative intensity value on y-axis [50].

From the mathematical point of view, each m/z can be represented as a single point in a high-dimensional geometry space. Data mining in a such multi-dimensional space could be performed easily only if the size of analysed sample is large enough. This is not true

in our case: sample from patients are relative few compared to the number of m/z involved. This is called the ***high-dimensionality-small-sample (HDSS) problem***, and it is the main issue of the current research on protein mass spectra classification [8]. Main goals of MS data analysis are to reduce the complexity of the high-dimensional geometry space, extracting only meaningful information, and to increase the amount of available data, including data from different sources [51]. In the first case, for instance, it is possible to detect the changes between normal and case proteome profiles; in the latter case, identification of the same peptides in different spectra allows the alignment of data from different sources. Unfortunately, direct comparisons are not so quite functional: also the same model of Mass Spectrometer built by the same company could give slightly different results on the same sample [52].

In this Ph.D. project we try to focus on both this problems, trying to align data from different sources ("Mass Spectrometry data alignment") and to isolate interesting portions of mass spectra, useful for biomarkers discovery ("Mass Spectrometry data analysis").

1.3 Mass Spectrometry data alignment

The first problem we faced was to provide a new method for MS data alignment, in order to integrate three small Alzheimer's disease experimental data we hold. We used feature extraction methods based on Mutual Information, and tested then using classification methods. This approach was published (see chapter 5).

1.3.1 Alzheimer's disease

Alzheimer's disease (AD) is a form of dementia that relates to the progressive loss of cognitive abilities. The disease rarely affects young people, and is responsible for one out of three cases of dementia in the elderly. It is estimated that the number of cases of AD in the years 2000 was 25 million, with a increasing tendency (estimates for 2050: 114 million) [53].

The disease is characterized by an accumulation of amyloid-beta peptide in the brain [54]. This seems caused by improper cleavage of the amyloid precursor protein (APP) [55]. APP is toxic for nerve cells, both in vivo and in vitro, so to induce inflammation and oxidative damage [56]. The toxicity is caused in different ways by different polymeric forms dell'amyloid-β peptide (Aβ), which causes various types of damage, like microglial infiltration (see for example [57, 58]).

Mass spectrometry offers various possibilities for the study of AD, and publications are substantial (for example, see [59–61]). One of the main problems is the availability of only small datasets [62, 63]: our work originates from this fact.

1.3.2 Experimental issues

The main issues associated to the alignment of MS spectra are due to the resolution of mass spectrometer, that usually does not permit to distinguish molecules with very similar weights. For example, the resolution of the MALDI-TOF spectrometers involved in this project is ± 8 Dalton (linear mode, no reflectron): signals (peptides) which differ for less than 8 Da are indistinguishable, and collapse to the same peak. This problems also affects data produced by similar spectrometers (even the same model from the same company) and also the same spectrometer over day-to-day variations.

Experimental variations, even minimal ones, determine a change in the spectra profiles and the inability to precisely align two similar experiments, due to a slight variation in the m/z ratios of the peptides analysed. The key point is therefore to detect a set of common attributes (commonalities) useful to compare relative abundance of the same peptide (protein) in the different spectra. This procedure is also compatible with experimental noise and the splitting of the signal of the same protein in multiple peaks, as happens in fingerprinting (the use of molecular weight information to identify proteins in sequence databases; [64–67]) or, accidentally, for protein fragmentations (see above, §1.2.2).

1.3.3 Our approach

There are two ways to achieve integration between different datasets [68]. If we know exactly the question we want to answer and if we know what information are available, the best idea is to design and

set up a database [68]: the whole information can then be analysed in an efficient way, thanks to the fact that the data are already sorted and easily retrievable. For example, it is possible to design and construct a database composed of different Affymetrix microarrays from different subjects and different tissues: we could query the database asking for levels of expression of a given gene, or to compare the levels of expression within the same tissue [69]. The second way concerns data on which we have no great certainties: the purpose is to analyse the data to search for some kind of correlation or relationship, to identify differences or similarities [68], like we did in this project.

Reducing the dimensionality of the raw input variable space is an important step in biomarkers identification, essential in data exploring and analysis. We are interested in methods that reveal or enhance the class structure of the data and rank the useful ones, to help define biomarkers: ***feature selection*** methods, that keep only useful features and discards others [70]. Feature selection methods can be classified into different main groups, based on the statistical approach adopted to reduce dimensionality: particularly, Mutual Information (MI) and Area Under ROC Curve (AUC), can be classified as Individual Variable Selection methods [8].

The ROC (Receiver Operating Characteristic) function describes the results of a classification model, usually represented in a contingency table (see table 1.1), that is a 2x2 table that list the results of a prediction (actual values versus predicted ones), listing the true/false positives and the true/false negatives. The AUC is a useful global way to quantify the accuracy of a test [71]. For several distribu-

tions, the AUC is a first-rate indicator of biomarkers discriminative power [72].

	Condition Positive	Condition Negative
Test outcome Positive	True positive	False positive
Test outcome Negative	False negative	True negative

Table 1.1: A contingency table

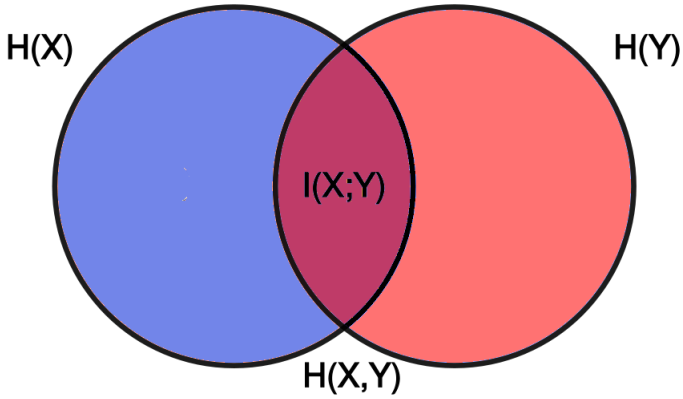


Figure 1.4: Use of set theory to visualize Mutual Information as quantity of information shared by X and Y.
The bigger the superposition, the higher the quantity of information of X could be deduced by Y, and vice versa; if there is no join, variables are independent.

Mutual Information (***MI***) is a dependency measure between two random variables. Is also defined as an entropy-based criterion between a predictive and a class variable [73], whose predictive capability for biomarkers has been shown [74].
MI quantify the dependency between two random variables (r.v.) X

and Y. It is possible to describe MI using set theory: the bigger the superposition, the higher the quantity of information of X could be deduced by Y, and vice versa; if there is no join, the variables are independent (see fig.1.3.3). It is also possible to describe MI as the amount of information $\mathcal{I}$ shared by X and Y. Formally:

$$\mathcal{I}(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1.2}$$

where $H(X)$ is the entropy of X, $H(Y)$ is the entropy of Y and $H(X,Y)$ is the entropy of the joint variable (X,Y). If $X \in \{x_1, x_2,..., x_k\}$ and $Y \in \{y_1, y_2,..., y_k\}$ are two discrete random variables, than we have:

$$H(X) = -\sum_{i=1}^{k} P_X(x_i) \log_2 P_X(x_i) \tag{1.3}$$

$$H(Y) = -\sum_{i=1}^{k} P_Y(y_i) \log_2 P_Y(y_i) \tag{1.4}$$

$$H(X,Y) = -\sum_{i,j}^{k} P_{XY}(x_i, y_i) \log_2 P_{XY}(x_i, y_i) \tag{1.5}$$

where P_X, P_Y are the distributions of X and Y, and P_{XY} the joint distribution between discrete random variables. Equations 1.3, 1.4 and 1.5 expressed the uncertainty contained in X, Y and (X,Y) respectively [73].

1.4 Mass Spectrometry data analysis

Data obtained by statistical methods are usually directly compared to highlight significant differences between case and control datasets. As already seen in other "omics" frameworks, biological variability, experimental noise and other factors prevent this kind of analysis to reach satisfactory results [75]: it is then necessary to reduce the complexity of the data observed (§1.2.2, §1.2.3), or, if biological knowledge allows it, to do a pre-selection of relevant data [76,77].

We tried to investigate the relationships between the mass spectra collected, attempting to establish a network of links between peptides, hoping to detect changes between profiles, like the increase or decrease of genetic products levels, absence or presence of peptides, etc. This network describes interactions between actors (peptides), like a social network describe interactions between people, and can be described and analyzed using graph theory (§1.4.2). Our work can be summarized, in a very general way, as the creation of networks starting from the spectra recorded from different subjects, the study and evaluation of the properties of networks so created and the comparison of the detected properties through statistical tests.

Since a large part of the knowledge acquired regarding the analysis of networks have been developed in the analysis of social networks, there are some references to social networks.

1.4.1 Renal Cell Carcinoma

Renal Cell Carcinoma (RCC) is the most common kidney-related tumor in adults, and accounts for about 3-4% of the total number of malignancies [78]. The clear cell variant (ccRCC) is the most frequent histological subtype of this tumor: it is responsible for approximately 75% of cases [79]. The incidence of RCC has increased steadily in the past years, but recently it seems more stable, probably as a consequence of an increasing use of imaging procedures [80].

RCC is generally asymptomatic, and at the time of diagnosis, about 30-50% of the patients already have local or distal metastases [81]. Moreover, RCC is one of the most radiation- and chemotherapy-resistant tumors [82]. The diagnosis of RCC is often confirmed by imaging studies, like X-ray and computed-tomography, but sometimes benign lesions could be hardly distinguished from malignant ones, for example, in presence of several cystic renal lesions or peculiar solid masses [83].

There are currently no biomarkers available for RCC early detection [1] (some attempts for biomarkers detection in [84–86]): nothing for an efficient prognosis, nor for monitoring recurrence after surgical treatment, nor for optimal predictive therapeutic approach [87].

ccRCC is not the only RCC subtype: some of our datasets were also taken from patients with papillary RCC (pRCC), clear cell papillary RCC, and also benign tumors (Oncocytoma, Angiomyolipoma, Cyst). pRCC is the second most frequent subtype of RCC (13-

15%) [88]. As virtually every RCC, pRCC is diagnosed accidentally, because is asymptomatic. It is the most bilateral renal tumor [89]. For details on how these different subtypes were used, see section 2.2.1.

Regardless of the subtype of RCC, patients usually undergo surgery, with a partial or total nephrectomy [90]. The 5-year survival rate is 60-70% but, if metastases appear, it decreases consistently: it is plausible that an early diagnosis may result in a significant increase in survival [91].

1.4.2 Oriented bipartite graphs

A network, as generally understood, is a collection of elements and relationships, and is usual represented with graphs [92].

A graph $G = (V, E)$ consist of a set of nodes or *vertices* V and a set of *edges* E, i.e. links between vertices $(x, y) \in V \times V$. The number of the vertices $S_V = |V|$ is defined *order*, the number of the edges $S_E = |E|$ is defined *size* of the graph.
A graph $A = (V_A, E_A)$ is a *subgraph* of $G = (V_G, E_G)$ if $V_A \subseteq V_G$ and $E_A \subseteq E_G$. Edges can be directed, like a graphical link between two point could be direct drawing a row: an ordered set of edges defines a *directed graph*. An ***oriented graph*** is a subtype of directed graph, the orientation of an undirected graph (no self-loops, no multiple adjacencies, and no 2-cycles; see [93]).

In our case the use of graphs is useful to establish network re-

lationships between the m/z signals (peptides) detected in a condition (for example, detected from ccRCC patients). We compare each signal with the others: this splits the same signals in two distinct subsets. This condition describes a particular graph topology, called ***bipartite***. Formally, a graph $G = (V, E)$ is bipartite if

$$G = (V, E); V_1 \subset V; V_2 \subset V; V_1 \cap V_2 = \emptyset; V_1 \cup V_2 = V$$

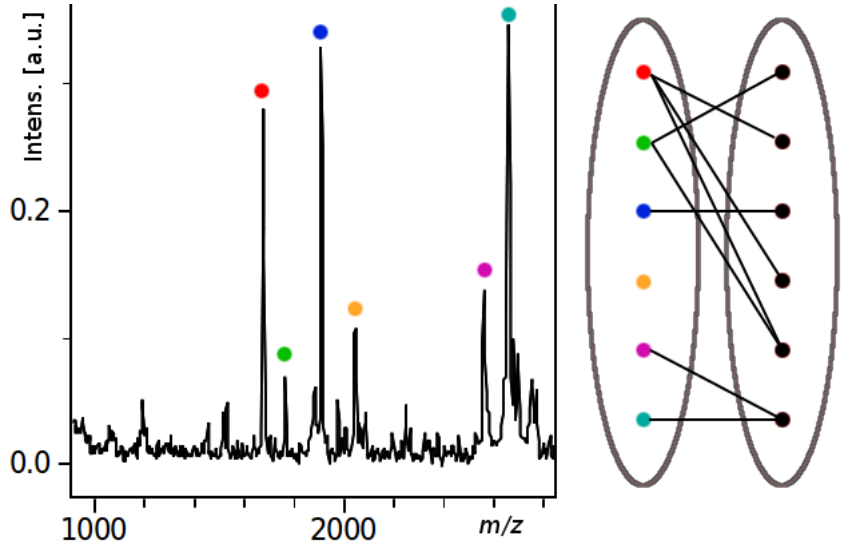


Figure 1.5: Representation of signals of a MALDI-TOF spectrum as an oriented bipartite graph.
Each signal is represented by a vertex, and colors allow to associate signals and vertices. Signals are sorted by m/z: this property is preserved also in the graph. This picture is actually simplified: the graph shows only some of the bigger signals: we use instead every single peak.

It should be emphasized that our graphs show an additional prop-

erty: the m/z signals are an ordered list of peaks, so they can be represented as an ordered set of vertices (see fig.1.5).

1.4.3 Random graphs

In order to consider stochastic components in our analysis we will also make use of random graphs (RGs). A random graph is defined as

$$RG = (\mathcal{G}, Pr) \tag{1.6}$$

where $\mathcal{G}$ is the set of graphs with similar properties and Pr is a probability measure [94]. The different models of generation of RGs rely on these two aspects, $\mathcal{G}$ and Pr (see [92, 95]).

There are two closely related basic models of RGs. One is the Erdos-Rényi model [96], given by considering the whole set of $\mathcal{G}$, graphs whose size is given (i.e. $|V| = n$) and each of the $\binom{n}{2}$ possible edges exists independently with a probability p. The other model $G(n, m)$, which we used in our analysis, is Gilbert's one [97], based on random choices between the given collection of graphs $\mathcal{G}$, having n nodes and m edges.

Our procedure is based on the generation of graphs obtained from real data. These graphs are the template $T = (n, m)$ for the construction of RGs $G = (n, m)$ that actually describe random relationships between the n vertices (peptides) considered.

1.4.4 Neighborhoods

Graphs can be studied observing their local or global properties [92]. To study the local properties of a graph is necessary to split it into subgraphs. A neighborhood is a subgraph composed by a constant number N of vertices.

In order to define a subgraph is necessary to identify the set of vertices from which it is composed: randomly selecting two vertices of an oriented bipartite graph as "centers" and a value k as "radius", we select the neighboring vertices for each of the two centers. Therefore, k define the size of the neighborhood and, for each pair of vertices selected, all the neighboring vertices included: $N = 2(2k+1)$.

Using the neighborhoods we can compare different subgraphs, analyzing different portions of the graph. We can fix a neighborhood and compare it with all the others, identifying the neighborhood with more edges, or those that show major variations in the comparison of data from cases against control data, and so on: we use neighborhoods to sample graph properties.

1.4.5 Graph Density

Global density is a overall indicator describing how nodes are more or less intensely connected to each other [92]. Formally, for a graph $G = (V, E)$:

$$den(G) = \frac{S_G}{|V_{G_1} \times V_{G_2}|} \tag{1.7}$$

where $S_G = |E|$ (graph size), V_{G_1} and V_{G_2} are the two subsets of nodes of a bipartite graph G ($V_{G_1} \subset V$; $V_{G_2} \subset V$; $V_1 \cap V_2 = \emptyset$; $V_1 \cup V_2 = V$).

Global density is a generic metric that provides only a particular property of the structure of a graph: it is in fact easy to imagine a great number of alternative graphs sharing the same global density. The composition of a network can be studied more effectively using methods for the characterization of network cohesion. This is an approach that helps to investigate the relationships between objects of a network and answering more interesting questions, like, for instance [92]:

- do friends of a member of a social network tend to be friends of one another?

that, changing to a biological context, sounds very similar to

- do proteins that work together also work with another protein?

It is also possible to reverse the question:

- do peptides showing a particular behaviour in normal condition show the same behaviour in different conditions (e.g. ccRCC)?

Formally, considering two connected vertices, v_A and v_B,

- is a subset of vertices V_F, all connected with v_A, also connected to v_B?

There are many different approaches to network cohesion estimation, but we chose local density, because of its strong local perspective,

hard bonded to the neighborhood features.

Local density is an indicator of cohesion of a network, and is defined, for a bipartite subgraph $A = (V_A, E_A)$, as

$$den(A) = \frac{S_A}{|V_{A_1} \times V_{A_2}|} \tag{1.8}$$

where A is a subgraph of G ($V_A \subset V$; $E_A \subset E$), $S_A = |A|$ (size of A), $V_{A_1} \subset V_A$; $V_{A_2} \subset V_A$; $V_{A_1} \cap V_{A_2} = \emptyset$; $V_{A_1} \cup V_{A_2} = V_A$.

1.4.6 Hypothesis testing, the Neyman-Pearson framework

The hypothesis testing is a valid method for comparing the distributions of values obtained from neighborhoods sampling of controls and patients data. Hypothesis testing is also one of the most important yet most confusing parts of statistical inference. This is due to several reasons, the main one being that the hypothesis testing is explained in a "hybrid form" [98] that combines the formulation of Fisher with the subsequent formulation of Neyman-Pearson [99].

Fisher's test Fisher's approach is based on a test that he called *Null hypothesis*

$$\mathrm{H}_0 : \mu = \mu_0 \tag{1.9}$$

and a "distance" measure that evaluates the values $\mathbf{X} = (X_1, X_2, ..., X_n)$ observed for H_0, a test statistic

$$\tau(\mathbf{X}) \tag{1.10}$$

It is the measurement of this distance that allows to assess the null H_0: if this distance is sufficiently "small", then H_0 is considered valid. That assessment is performed using a pivotal function and the distribution of values $\tau(\mathbf{X}, \mu)$.

The performance of this assessment is calculated by:

$$\mathbb{P}(\tau(\mathbf{X}) \geq \tau(\mathbf{x}); H_0 \text{is valid}) = p \tag{1.11}$$

where $\tau(\mathbf{x})$ is the observed value of the statistic $\tau(\mathbf{X})$. The p-value defines the worst possible case for a Null hypothesis: low values describe low probability events, so either the observation of such a value x is a rare event or the null hypothesis is not valid. The smaller the p-value, the less plausible is H_0 [99].

Neyman and Pearson test Neyman and Pearson revealed the limitations of Fisher's approach: how does the modeler choose $\tau(\mathbf{X})$ [100]? In particular, they questioned [99–101]:

- the ability to construct different valid statistical tests without being able to distinguish the most appropriate one;
- the use of p-value as a measure for the assessment of the null hypothesis derived from the sample.

The key point of the approach of Neyman-Pearson is the addition of an alternative hypothesis that transforms testing in a choice among two mutually exclusive hypothesis [99]. Formally:

$$H_0 : \theta \in \Theta_0 \text{ against } H_1 : \theta \in \Theta_1 := \Theta - \Theta_0 \tag{1.12}$$

where the parameter space is $\Theta = \Theta_0 \cup \Theta_1$.
The main purpose of the statistical test becomes the formulation of a decision rule that allows for each observed value $\mathbf{x} := (x_1, x_2, ..., x_n)$ (i.e. *realization* of sample $\mathbf{X} := (X_1, X_2, ..., X_n)$) to accept or reject H_0 on the basis of a statistical test $\tau(X)$. This effectively splits the sample space $\aleph$ into two complementary sets S_1 and S_0, such that $S_0 \cup S_1 = \aleph$ and $S_0 \cap S_1 = \emptyset$ [99]. To compare the Fisher's Eq.1.9:

$$H_0 : \mu = \mu_0 \tag{1.13}$$

$$H_1 : \mu \neq \mu_0 \tag{1.14}$$

Accept or reject H_0 on the basis of a statistical test can lead to two types of errors: refuse H_0 when this is actually valid - type I error, or α - or accept H_0 when it is in fact false - type II error, or β. This situation can be summarized in the table 1.2. The calculation of α and β also allows you to define the statistical ***significance*** $(1 - \alpha)$ and ***power*** $(1 - \beta)$ of a test. Since the type I error usually

	Accept H_0	Reject H_0
H_0 valid	correct decision	type I error (α)
H_0 invalid	type II error (β)	correct rejection

Table 1.2: Possible errors in hypothesis testing

has worse consequences (e.g. a person who is diagnosed with a tumor that doesn't have), clinical tests are designed to minimize α rather than β. Since our goal is the identification of biomarkers signals, we generally minimize β, so to avoid missing signals of interest.

The Neyman-Pearson hypothesis tests is actually a comparison between two different distributions: α and β result from the errors in interpreting values found in the overlapping portion of the two distributions. This suggests that the there is a trade-off between type I and II errors: the decrease of α increases β and vice versa.

1.4.7 Robustness

Robustness is an essential property of biological system [102, 103] and can therefore be considered as a decisive factor for selecting a credible model or pinpointing the weaknesses of a failed model [104], as we have done for some of our approaches. This is particularly important in areas such as medicine and drug discovery where robustness analyses are the logical next step to face with many uncertainties, arising for example from the experimental design or even from technical (or biological) variabilities (§1.2.2) [104].

In its general form, the word *robustness* refers to the ability of a process to cope well with uncertainty, the different ways in which the performance of the process is evaluated. The framework used for modelling uncertainty leads to many alternative definitions of the word itself. Hampel defines the word robustness within a general statistical context [105]. The definition can be summarized by consider-

ing the *robustness as the stability theory of statistical procedures*. He systematically investigates the effects of deviations from modelling assumptions on known procedures and, if necessary, develops new, better procedures.

Recently, the relationship between robustness and multiple criteria decision analysis has been observed by a number of researchers [106, 107]. For instance, Kouvelis and Yu studied the robustness in the context of discrete optimization [108]. They provide theoretical results and algorithms for determining the solution that exhibits the best worst case deviation (or percentage deviation) from optimality, among all feasible decisions over all realizable input data scenarios. Related ideas can be found in the Robust Bayesian literature [109], where the robustness analysis has been developed mainly to cope with the arbitrariness affecting the choice of a prior distribution. The high grade of that arbitrariness make the Bayesian methods difficult to be acceptable as standard practice, therefore the key idea behind Robust Bayesianis the need to base inferences only on the actual assessment by the experts, specifying a class of *priors* compatible with their opinions and studying the influence of changes in the prior on the values of the quantity of interest [110, 111].

Chapter 2

Aim of the Project

The aims of the project are:

- to develop a novel method for MS data alignment;
- to provide novel and original methods for the analysis of MS data, in order to identify suitable biomarkers signals.

Both goals were examined separately in the following sections.

2.1 Mass Spectrometry data alignment

2.1.1 Summary

In the first part of the project we focused on Alzheimer's disease data. We collected samples from 77 subjects in three different hospitals and medical institutions. The poor number of samples gathered per hospitals (41, 21 and 15 subjects involved) made any significant data analysis hard to perform, so the idea was to integrate MS

data produced by different sources, using similar equipment in different labs: a "data integration/fusion" approach, useful to improve the performance of data analysis, thus biomarkers signals discovery. This Mass Spectrometry Data Alignment (MSDA) approach could be useful to merge different datasets from different labs, also in different context.

Briefly, we developed a theoretical method for the alignment of MALDI-TOF MS data from Alzheimer patients and controls. This approach is founded on feature construction and extraction methods (FSCM) and on a measure for the stochastic dependence of random variables (Mutual Information). We tested this approach using a machine learning environment (`RapidMiner`, [112]), and compared it to other approaches with satisfactory results.

2.1.2 Materials: samples from Alzheimer disease patients and controls

Samples were collected after receiving informed consent from all the subjects participating in the study from three different hospitals using a standardized protocol. A cohort of 6 control subjects and 9 AD patients was recruited from the University of Florence - School of Medicine network (Florence, Italy), 23 controls and 18 AD patients from San Gerardo Hospital (Monza, Italy), and a total of 6 controls and 15 AD patients from the Center for Aging Brain and Dementia (Brescia, Italy). Plasma was obtained from blood collected in EDTA. A cohort summary is available in table 2.1.

Table 2.1: Cohort description

Location	CASE	CONTROLS	Sums
Monza	18	23	41
Florence	9	6	15
Brescia	15	6	21
Totals	42	35	77

Plasma pre-fractionation Sample purification was performed in duplicate at room temperature with ClinProt^{TM} MB-HIC8 (Magnetic Beads based Hydrophobic Interaction Chromatography) kit. All processes were automatically executed by using a ClinProt^{TM} Robot as previously described [84].

MALDI-TOF MS and Data Processing The plasma protein profiles were obtained by an MALDI-TOF Reflex IV^{TM} mass spectrometer (Bruker Daltonics, Germany). The instrument was externally calibrated using a mixture of standard peptides/proteins. Mass spectra were acquired in positive linear mode in the m/z range of 1,000-10,000 Da; accumulation of signals from different sample spot positions resulted in a total averaging spectrum. The spot was pre-irradiated with higher laser power to improve the spectra quality before each acquisition cycle. Multiple spectra comparison was performed using ClinProTools^{TM} 2.1 software (Bruker Daltonics). First, each raw spectrum was normalized and all spectra were then recalibrated (realignment) using prominent internal m/z values. Subsequently, baseline subtraction and peak detection were achieved before peak area calculation. The software calculates the mean spectrum

for each subject's data set, and then, selects the spectrum that is most similar to the average one to be used for further evaluations. ClinProTools^{TM} automatically provided a list of peaks sorted according to the statistical relevance to differentiate between classes with their corresponding p-value.

2.1.3 Methods: our approach

Our method for Mass Spectrometry Data Alignment (MSDA) from different lab is based on the Features Extraction and Construction Method (FSCM), a process of dimensionality reduction for the selection of a set of relevant features in a dataset, useful to build a model for the evaluation of datasets of similar origin [113]. Features extraction and construction method consists of

- features construction mechanism
- relevance mechanism

Common attributes are usually investigated with statistic methods that search for dependencies between variables. We choose Mutual Information (MI; see §1.3.3 for a formal description) to measure the commonalities shared by signals/peptides. In brief, MI quantify the dependencies of two distributions of values (X, Y).

We apply the Mutual Information to quantify the dependencies between signals, merging the peptide signals with higher values of shared MI, to maximize the performance of a classification task. MI is not the only method to quantify commonalities: however, at the

time the publication, there was no evidence that somebody used MI for the alignment of MS spectra.

Problem description

The data for each laboratory k can be defined using three objects:

1. the subpopulation of peptides $\mathcal{P}^{(k)}$ useful to identify biomarker signals within the whole peptides population detected by the k lab spectrometer. There is an intensity value, a random variable ${I_p}^{(k)}$ associated with each peptide $p \in \mathcal{P}^{(k)}$, distributed accordingly to $f_{I_p^{(k)}}(i_p^{(k)})$. For simplicity:

$$f_{p,k}(i) \equiv f_{I_p^{(k)}}(i_p^{(k)}) \tag{2.1}$$

$$f_{p,k}(I) \equiv f_{I_p^{(k)}}(I_p^{(k)}) \tag{2.2}$$

2. the random variable $M_p^{(k)}$, that is the m/z for each peptide p, distributed accordingly to $f_{M_p^{(k)}}(m_p^{(k)})$;

3. a Bernoulli random variable $D^{(k)}$ expressing the case-control group membership.

Our aim is to highlight and evaluate the relationships between features (disease class and intensity). Basically, it corresponds with the evaluation of the joint distribution that, write with the aid of (2.1) and (2.2) is:

$$f_{p,k}(i,d) \equiv f_{I_p^{(k)},D^{(k)}}(i_p^{(k)}, d^{(k)}) \tag{2.3}$$

$$f_{p,k}(I,D) \equiv f_{I_p^{(k)},D^{(k)}}(I_p^{(k)}, D^{(k)}) \tag{2.4}$$

Features construction mechanism

As described above (see section 1.3.2), the MALDI-TOF mass spectrometer cannot distinguish signals originated by peptides of proteins with very similar weights. In our case, the resolution limit is ± 8 Da. From the formal point of view, considering two labs, A and B ($k = A, k = B$), for each pair of peptides p_x e p_y, satisfying

$$|M_{p_x}^A - M_{p_y}^B| \leq 8 \tag{2.5}$$

Considering the last equation, we define the dependence Z_p^k between intensity value and disease class using:

$$\begin{aligned} Z_{p_x}^{(A)} &= lg \frac{f_{p_x,A}(I, D^{(A)})}{f_{p_x,A}(I) \cdot f_{D^{(A)}}(D^{(A)})}, \\ Z_{p_y}^{(B)} &= lg \frac{f_{p_y,B}(I, D^{(B)})}{f_{p_y,B}(I) \cdot f_{D^{(B)}}(D^{(B)})}, \\ Z_{p_x,p_y}^{(A,B)} &= lg \frac{f_{p_x,A}(I, D^{(A)})}{f_{p_x,A}(I) \cdot f_{D^{(A)}}(D^{(A)})} + lg \frac{f_{p_y,B}(I, D^{(B)})}{f_{p_y,B}(I) \cdot f_{D^{(B)}}(D^{(B)})} \end{aligned} \tag{2.6}$$

Relevance mechanism

The relevance method is implemented as the sum of the Mutual Information shared by $I_{p_x}^{(A)}$ with $D^{(A)}$ and $I_{p_y}^{(B)}$ with $D^{(B)}$. The MI $\mathcal{I} = \mathcal{I}(I_{p_x}^{(A)}, D^{(A)}) + \mathcal{I}(I_{p_y}^{(B)}, D^{(B)})$ is calculated using the expected value $E[Z_{p_x,p_y}^{(A,B)}]$ accordingly to the following equation (see 2.6):

$$\begin{aligned} E[Z_{p_x,p_y}^{(A,B)}] = E\left[lg \frac{f_{p_x,A}(I, D^{(A)})}{f_{p_x,A}(I) \cdot f_{D^{(A)}}(D^{(A)})}\right] + \\ E\left[lg \frac{f_{p_y,B}(I, D^{(B)})}{f_{p_y,B}(I) \cdot f_{D^{(B)}}(D^{(B)})}\right] \end{aligned} \tag{2.7}$$

We want to identify those peptides that provide the highest values (argmax) of MI:

$$\underset{(p_x,p_y)\in p^{(A)}\times p^{(B)}}{\arg\max}(\mathcal{I}(I^{(A)}_{p_x}, D^{(A)}) + \mathcal{I}(I^{(B)}_{p_y}, D^{(B)})) \tag{2.8}$$

From the computational perspective, the calculation of MI of $\mathcal{I}(I^{(A)}_{p_x}, D^{(A)})$ and $\mathcal{I}(I^{(B)}_{p_y}, D^{(B)})$ in Eq.2.8 is made discretizing and tallying, for each peptide, the samples from distribution of intensities $f_{p_x,A}(i)$ (or $f_{p_y,B}(i)$), the class disease $f_{D^{(A)}}(d^{(A)})$ (or $f_{D^{(B)}}(d^{(B)})$), and the join distribution $f_{p_x,A}(i,d)$ (or $f_{p_y,B}(i,d)$). However this leads to troubles if the datasets from all the three laboratories are involved, and, generally, if we want to involve an greater number of labs.

Aligning more than two labs From the formal viewpoint, it is possible to formulate again the problem in (2.8) using graphs theory (graphs theory is discussed in §1.4.2), formulated via ***Maximum Weight Bipartite Matching*** (***MWBM***) [114]. Roughly, we can symbolize each signal as a vertex of a bipartite weighted graph: among all the possible weighted graphs, the one with the greatest MI is the one that maximize the sum of the weights. Without going too far into the formalism [115], we can rewrite the equation 2.5 and consider our m/z data as observation to estimate

$$\mathcal{R}_i = \{(M^A_{p_x}, M^B_{p_y}) : |M^A_{p_x} - M^B_{p_y}| \leq 8\} \tag{2.9}$$

that is

$$\tilde{\mathcal{R}}_i = \{(m^A_{p_x}, m^B_{p_y}) : |m^A_{p_x} - m^B_{p_y}| \leq 8\} \tag{2.10}$$

Consider now the bipartite graph $G = (V_1 \cup V_2, E)$:

$$V_1 = \{m_{p_x}^{(A)} | \exists\, p_y, j : (m_{p_x}^{(A)} - m_{p_y}^{(B)}) \in \tilde{\mathcal{R}}_j\}, \tag{2.11}$$

$$V_2 = \{m_{p_y}^{(B)} | \exists\, p_x, j : (m_{p_x}^{(A)} - m_{p_y}^{(B)}) \in \tilde{\mathcal{R}}_j\}, \tag{2.12}$$

$$E = \bigcup \tilde{\mathcal{R}}_i \tag{2.13}$$

We can now estimate the weights for the two peptides involved:

$$\begin{aligned} w(m_{p_x}^{(A)}, m_{p_y}^{(B)}) = \sum_{t,d} \tilde{f}_{p_x,A}(t,d) \log \frac{\tilde{f}_{p_x,A}(t,d)}{\tilde{f}_{p_x,A}(t) \cdot \tilde{f}_{D^{(A)}}(d^{(A)})} + \\ \sum_{t,d} \tilde{f}_{p_y,B}(t,d) \log \frac{\tilde{f}_{p_y,B}(t,d)}{\tilde{f}_{p_y,B}(t) \cdot \tilde{f}_{D^{(B)}}(d^{(B)})} \end{aligned} \tag{2.14}$$

and so on, to assess the MWBM in a general form, that could embroil more labs.

2.1.4 Methods: competitive approaches

To understand the usefulness of our method we decide to test it against possible competitors. We choose two simple but effective tests, that show the progress of our method in a easy way. The methods we choose are:

- Equal Mass Fusion test (EM);
- t-Test Fusion test (TT).

The EM unify features from different labs whenever the associated mass values are equal. It's a simple approach that postulate that there's a low level of noise and misalignment in the spectra, so

that it is better to use the data without further processing.

The TT is based on a statistical approach, the t-Test. The t-Test is a statistical test, which tries to understand if the difference of means of datasets is due to chance. Further information on t-Test in Section 2.2.3. For all pair of features whose mass difference ranges in an interval of ± 8 Da, we compare the means from two different samples by a statistical t-Test. Then, we unified these pairs of features with the maximum value of significance.

2.1.5 Evaluation tool

The purpose of our work is to propose a method that optimizes the alignment of mass spectrometry data. In the previous section we have listed out others methods; what follows is a description of the machine learning environment used for the comparison. The results can be found in section 3.1.

We employed `RapidMiner`, a flexible and powerful machine learning environment. The interface is intuitive enough to allow simple editing also for users without a solid foundation in computer science. `RapidMiner` is based on knowledge discovery processes (KD processes): every process is viewed through complex nested tree. Every tree object is called operator, and each operator could incorporate a number of operations and parameters, a "not-so-black" boxes that allow you to manage the flow of input and output of data in a very simple way. All processes are described using XML mark-up language. The figure 2.1 shows a snapshot of the operators described in table 2.2, that we use to implement our approach.

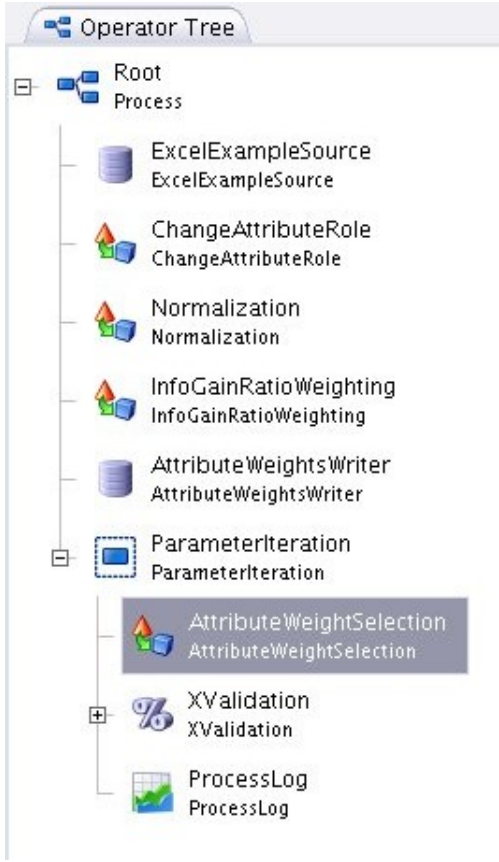


Figure 2.1: A snapshot of the operator tree of `RapidMiner`. Operators are described in table 2.2

Table 2.2: Summary of the knowledge discovery process we implemented in `RapidMiner`

Sequence	Operator	Description
1	Data Source	Read input file. The input files are the aligned data produced with the methods discussed.
2	Normalization	Normalize data signal intensities in $[-1, 1]$.
3	Information Gain	Computation of MI.
4	Parameter Iteration	Performs an iterative cycle of operations testing all the parameters set in.
4.1	Attribute Weight Selection	See 3 - Information Gain.
4.2	Cross Validation	It starts a cross-validation (training/testing sets) sub-process: input data set $\mathbf{S}$ is split up into subsets $\{S_1, S_2, ..., S_n\}$; the forthcoming operators are applied n times using, for each iteration i, the set S_i as test set and $S \backslash S_i$ as training set.
4.3	Model Applier	It applies the model delivered by SVM (see 4.2)
4.4	Binomial Classification Performance	Computes the performance of classification providing the Area Under ROC Curve (AUC; see 1.3.3)

2.2 Mass Spectrometry data analysis

The following sections will report a variety of approaches and methods for the analysis of data from MALDI-TOF mass spectrometry we employed. The basic concepts used in these approaches are described in section 1.4, and are briefly summarized below:

- the data originated from mass spectrometric analysis obtained from patients affected by Renal Cell Carcinoma (ccRCC, not-ccRCC), and from healthy controls (section 1.4.1);

- analyses were not directly performed on the data, but from the properties of a graph created using the same datasets (section 1.4.2);

- graphs were sampled and studied by calculating density, employing neighborhoods (sections 1.4.4, 1.4.5);

- density and neighborhoods produced sets of values, which are weighted with different methods, and provided a range of numerical results. These results were used to make a series of hypothesis tests (section 1.4.6) that compare the different datasets (e.g., cases versus controls);

The results of the tests of hypotheses were used to define regions of interest within the spectra analyzed: these regions contained the most interesting signals, on which focus the attention for the recognition of biomarkers signals.

2.2.1 Materials: ccRCC, not-ccRCC patients and controls datasets

Cohort description

The cohort is composed by 187 people screened in three different clinics:

- San Gerardo Hospital (Monza, Italy),
- "Ospedale Maggiore Policlinico" Foundation (Milano, Italy),
- Desio Hospital (Desio, Italy),

and consist of 85 controls (58 men, 27 women) and 102 cases (64 men, 38 women). Mean age for controls was 45 with a range of 30-68 years, while for patients it was 64 with a range of 33-88 years. Patients have been divided into groups according to their pathologies: clear cell RCC (n = 79) and other different histological subtypes i.e., non-ccRCC (n = 23). A bird-eye view of the cohort is summarized in table 2.3, made accordingly to 2002 TNM (tumor-node-metastasis) system.

Samples: Urine collection and handling procedure

The samples consisted in urines collected from patients the day before surgery (ccRCC and not-ccRCC); controls samples was collected from healthy volunteers. All subjects had signed an informed consent prior to sample donation. Study protocols and procedures were approved by the local ethic committee and analysis was carried out in agreement with the Declaration of Helsinki.

Second morning midstream urine were collected in sterile urine tubes (Anicrins.r.l., Italy). Within one hour the sample was centrifuged at 4°C for 10 min at 1000 g to remove cell debris and casts. Supernatant was immediately transferred into 2 mL tubes and stored at -80°C until further use. A tube per each sample was thawed once for the automated peptide isolation procedure.

Mass Spectrometry techniques

Peptidome separation with magnetic beads The proteome/peptidome were extracted using ClinProt™, a technology providing a wide range of functionalities for excellent and sophisticated peptide and protein separation and preparation directly from biological fluids. In our case, we used Weak Cation ion eXchange Magnetic Beads (WCX MB), as previously described in [85, 116].
The profiling kit was employed to purify all samples and checked before use with a standard light microscope (Dialux EB-20, Leitz, Germany) in order to evaluate dispersion and potential aggregation within the suspension. An auto-mated extraction procedure was achieved using a 96-channel Hamilton STARplus® pipetting robot (Hamilton, Bonaduz, Switzerland) for a greater sample throughput. As concerning the WCX protocol, the binding, wash and desorption steps of the beads were based on the manufacturer instructions and optimized for the implementation on the pipetting robot. Briefly, 10 μL of WCX MB were used for the analysis of 40 μL of urine sample, mixed intensively with 10 μL of a binding buffer supplied with the kit and incubated for few minutes at RT in a 96-well plate. After the removal of supernatant,the WCX beads were washed three times with

45 μL of a recommended washing solution and eluates with 15 μL of a 130 mM ammonium hydroxide solution. Thus obtained eluates were immediately stabilized with a 30 μL of a 3% TFA solution and then used for the MALDI-TOF analysis. Pipetting was automatized using a 96-channel Hamilton STARplus® robot.

MALDI-TOF peptide profiling

Aiming for the MALDI-TOF acquisition of urinary spectra profiles for all studied patients, a MALDI-spotting procedure was automatically obtained by robot. To this purpose, 4 μL of the WCX eluates peptide fraction were mixed with 15 μL CHCA matrix solution (0.3 g/L in ethanol/acetone 2:1). Then, 1 μL of this mixture was spotted in quadruplicate directly onto a MALDI AnchorChip 600/384 target plate (Bruker Daltonics, Germany) with the pipetting robot. The target plates were air-dried and immediately kept in an environment controlled storage chamber (RT, 5% oxygen, 95%nitrogen) until transfer into the MALDI-TOF/TOF mass spectrometer.
Fractionated samples were analyzed using an UltraFlexII™ MALDI TOF/TOF MS instrument (Bruker Daltonics, Germany) and mass spectra were automatically acquired in positive linear mode (LM). The acquisition was performed in a m/z range of 1 to 12 kDa and the external calibration was achieved using a mixture of peptide/protein standards, ProtMix I and PepMix II (Bruker Daltonics, Germany). Analyses were performed using AutoXecute tool (v. 3.0.100.0) of FlexControl software v. 3.0 (Bruker Daltonics, Germany). For each MALDI spot, spectra were recorded from six different spot positions (200 shots per position) and summed up (1200 satisfactory shots).

	Number of patients
Patients	102
Male - Female	64 - 38
SD Age (mean, at diagnosis)	64 ± 12.4
STAGING	
Primary tumor (T)	
pT1 - pT2 - pT3 - pT4	57 - 14 - 17 - 1
unknown	3
Regional Lymph nodes (N)	
NX - N0 - N1	59 - 27 - 1
unknown	5
GRADE	
G1 - G2 - G3 - G4	6 - 62 - 17 - 1
unknown	6
HISTOLOGY	
Clear cell RCC (ccRCC)	79
Papillary RCC	7
Clear cell & papillary RCC	1
Chromophobe	2
Adenocarcinoma	1
Renal neoplasm	1
Mucinous tubular and spindle cell RCC	1
Oncocytoma (benign)	6
Angiomyolipoma (benign)	3
Cyst (benign)	1
TUMOR TYPE	
Malignant - Benign	92 - 10

Table 2.3: Patients' clinical characteristics according to the 2002 TNM (tumor-node metastasis) system.

2.2.2 Divergence Analysis with Random Graphs

In the first approach we focused on the identification of spectral regions that showed divergence in the comparison between case and controls. We supposed that divergent regions were those containing signals that justify the biological differences between the two states, healthy and sick.

A first goal was to demonstrate, using statistic tests, that the properties obtained from the graphs constructed from data (controls and cases) deviate from a uniform reference model with similar properties, but generated randomly: this showed that our system was able to distinguish between real and random data. We then compared controls and cases, hoping that the ability to differentiate let highlight regions showing the greatest divergence, i.e. the regions probably holding biomarkers signals.

The whole analysis was based on hypothesis testing with Random Graphs (RGs): if, on the one hand, RGs were useful for the creation of graphs with properties similar to those obtained from the data, but randomly assembled, on the other hand they were also used to perturb the "real" graphs to create a populations of similar graphs.

This is the summary of the main steps of the proposed method (some terms are better explained in the next section):

1. Use of divergence to track edges of bipartite graphs $R^{(obs)}$, that represent real data (case and controls). Tools: Kullback-Leibler

(KL) divergence, bipartite graphs.

2. Density calculation for $R^{(obs)}$ graphs:

   - computation of global density, Tool: global density;
   - use of neighborhoods for calculating distribution of values of local density. Tools: neighborhoods, local density.

3. Construction of random graphs (RGs) for comparison between RGs and

   - cases data, using the global density of the graph $R^{(case)}$: Uniform Reference model (URfM). Tool: global density;
   - controls data, using the local density values distribution of the graph $R^{(control)}$: Uniform Random model (URnM). Tools: local density, perturbation probability.

4. Comparison between values of local density of the neighborhoods between:

   - URfM versus Cases;
   - URnM versus Controls;
   - Controls versus Cases;

   The distributions of local density values are used for hypothesis testing: we carry out an hypothesis test for each neighborhood. Tool: hypothesis tests.

5. Retrieval of the mass ranges that identify the more interesting neighborhoods.

Parameters involved in this method, such as the KL threshold δ, the parameters related to the perturbation and the size of the neighborhoods, will be detailed in the results chapter (section 3.2.1).

Divergence representation

Peptides detected by MS can be represented using a graph: each vertex stand for a specific m/z signal; edges are tracked between vertices which divergences in the intensity values $\mathcal{I}$ exceed a threshold. More formally, for each group of subjects, for instance, patients, signal intensity $\mathcal{I}^{(case)}$ can be expressed through a product $\mathcal{I}_{m_1}^{(case)} \times \mathcal{I}_{m_2}^{(case)} \times ... \times \mathcal{I}_{m_n}^{(case)} \times$ of spaces $\mathcal{I}_{m_i}^{(case)}$, $1 \leq i \leq n$, given by all potential intensity values of m_i. We also assumed that each $\mathcal{I}_{m_i}^{(case)}$ was associated with a distribution function $f_{\mathcal{I}_{m_i}}^{(case)}$.

We called *template* $R^{(case)}$ a bipartite graph $(V_1 \cup V_2, E)$ with $V_1 = V_2 = \{m_1, m_2, ..., m_n\}$, and $(m_i, m_j) \in E$ if a measure of divergence $D(\tilde{f}_{I_{m_i}}^{(case,1)}, \tilde{f}_{I_{m_j}}^{(case,2)})$ between the empirical distribution $\tilde{f}_{I_{m_i}}^{(case,1)}$ and $\tilde{f}_{I_{m_j}}^{(case,2)}$ exceeds threshold δ. The template $R^{(case)}$ was calculated sampling $I_{m_i}^{(case,1)}$ and $I_{m_j}^{(case,2)}$ from each pair $(\mathcal{I}_{m_i}^{(case)}, \mathcal{I}_{m_j}^{(case)})$. We chose **Kullback-Leibler entropy divergence** as measure of divergence [117]:

$$\tilde{D}(f_{I_{m_i}} || f_{I_{m_j}}) = \sum_i f_{I_{m_i}}(i) \log \frac{f_{I_{m_i}}(i)}{f_{I_{m_j}}(i)} \tag{2.15}$$

The template $R^{(case)}$ has been obtained by sampling $I_{m_i}^{(case)}$ and $I_{m_j}^{(case)}$ from $(\mathcal{I}_{m_i}^{(case)}, \mathcal{I}_{m_j}^{(case)})$: this template described pattern of di-

vergences *inside* a population (i.e.,case). Generally, we wanted to associate observed m/z ratios with the value of divergence for their respective observed intensities $\mathcal{I}$: given an observed group of $\mathcal{I}^{(obs)}$ from which the template $R^{(obs)}$ has been obtained through some sampling mechanism, we called *local divergence* the vector

$$D = (d_{(m_i,m_j)})_{(m_i,m_j)\in V_1\cup V_2} \tag{2.16}$$

where $d_{(m_i,m_j)} = \tilde{D}(f_{I_{m_i}}||f_{I_{m_j}})$. We also called *amount of divergence*

$$K = \sum_{(m_i,m_j)\in E} d_{(m_i,m_j)} \tag{2.17}$$

Graphs like $R^{(obs)}$ were created by adding an edge to E if $d_{(m_i,m_j)} \geq \delta$ (or removing one if $d_{(m_i,m_j)} < \delta$), where δ was the divergence threshold.

Random models

We employed two different types of random models:

- an "uniform reference model", based on Gilbert Model for Random Graphs: see §1.4.3;
- an "uniform random model", random graphs built starting from an observed graph $R^{(obs)}$.

The essential difference was that the first preserves only global properties, while the latter preserves also the amount of divergence, a local property, even with a certain (low) probability that this could vary substantially.

Uniform Reference model (URfM) We chose as uniform reference model a Random oriented bipartite graph defined as follows: URfM(v, e) is a random oriented bipartite graph taking values from $(\mathcal{G}, \Pr)$ where $\mathcal{G}$ is the set of all oriented bipartite graphs with v vertexes and e edges and Pr is a uniform probability measure on $\mathcal{G}$ assigning to each graph $G_i \in \mathcal{G}$ the same probability value $p = \Pr(\{\mathcal{G} = G_i\})$.

This graph will be used for comparison with the cases datasets: it would be incorrect to use a graph $R^{(case)}$, which is considered in some sense already perturbed by the pathology (RCC), to build an Uniform Random model, which is in fact a perturbed graph with a given probability, as happens for $R^{(control)}$.

Uniform Random model (URnM) We wished to construct a random graph able to preserve some property of an observed template $R^{(obs)} = (V_1 \cup V_2, E^{(obs)})$ (see §2.2.2). The key idea was to apply enough distortion (i.e., perturbation of $R^{(obs)}$) in such a way that the probability for the model to fail (i.e., failing to preserve that property) is small.

Among the different graph properties cited in literature, here we were interested in preserving the original cohesion of $R^{(obs)}$ [92]: we tried to preserve the density of $R^{(obs)}$ by maintaining its order $S_V^{(rnd)} = S_V^{(obs)}$ (i.e., the number of its vertices; see §1.4.2). A density-preserving random graph $R^{(rnd)} = (V_1 \cup V_2, E^{(rnd)})$ was created starting from a template $R^{(obs)}$ adding a random quantity of noise $\epsilon_{(m_i,m_j)}$

such that

$$\mathbb{E}(\sum_{(m_i,m_j)\in V_1\times V_2} \epsilon_{(m_i,m_j)}) = 0 \tag{2.18}$$

$$\mathrm{Var}(\sum_{(m_i,m_j)\in V_1\times V_2} \epsilon_{(m_i,m_j)}) = \sigma^2_{tot} \tag{2.19}$$

So we obtained $R^{(rnd)}$ from $R^{(obs)}$ by adding an edge to E if $d_{(m_i,m_j)} + \epsilon_{(m_i,m_j)} \geq \delta$ or removing one edge if $d_{(m_i,m_j)} + \epsilon_{(m_i,m_j)} < \delta$.

Another key point was that the size $S_E^{(rnd)} = |E^{(rnd)}|$ of the random graph should have to deviate from $S_E^{(obs)} = |E^{(obs)}|$ with low probability. We used a "drop off" function $f_{S_E}(c)$:

$$\Pr(X \geq cS_E) \leq f_{S_E}(c) \tag{2.20}$$

Summarizing, by locally adding up random quantities $\epsilon_{(m_i,m_j)}$ we obtained a random graph $R^{(rnd)}$ which globally preserved (on average) the amount of divergence K of $R^{(obs)}$ (see previous paragraph). Also, it deviated from the original number of edges with low probability.

Hypothesis testing

We applied three different hypothesis tests (Neyman-Pearson), each involving different sets of data. In particular:

1. "Uniform Reference model versus case", also called random versus cases test, based on graph $R^{(obs)}$, built from case datasets. The global properties of the graph obtained ($R^{(obs)} = R^{(case)}$) are used to generate the random graph URfM(v, e). The aim is

to show that the graph created experimental data differs from random one:

$$\mathrm{H}_0 : \mathrm{URfM}(v, e) \text{ and } \mathrm{H}_1 : R^{(case)} \tag{2.21}$$

2. "Uniform Random model versus control" based on control datasets. The observed graph $R^{(obs)} = R^{(control)}$ is used to made the URnM graph. Again, the aim is to show that the graph created from the experimental data differs from random one:

$$\mathrm{H}_0 : \mathrm{URnM}(v, e) \text{ and } \mathrm{H}_1 : R^{(control)} \tag{2.22}$$

3. "Control versus case": a slightly different test, based on comparisons between two different $R^{(obs)}$, $R^{(obs)} = R^{(controls)}$ and $R^{(obs)} = R^{(case)}$. We chose $R^{(controls)}$ as null hypothesis (H_0) because it is more reasonable to think about the healthy state as the reference state, so the hypothesis test is

$$\mathrm{H}_0 : R^{(control)} \text{ and } \mathrm{H}_1 : R^{(case)} \tag{2.23}$$

The variability space of interest (the sampling from $\mathcal{I}^{(case)}$ in order to obtain $R^{(case)}$), is the statistic $\eta(I^{(case,1)}, I^{(case,2)})$ or $\eta(G)$, where η is some suitable graph property compactly summarizing the structural relation (i.e., given through the random model G) endowing the set $I^{(case,1)} \cup I^{(case,2)}$.

Test and numeric evaluation

In order to evaluate the tests proposed in the last section, it is generally enough to ensure that:

1. the probability of rejecting the null when valid (type I error) is small:

$$p(\text{rejecting } H_0 | H_0 \text{ is true}) = \alpha \tag{2.24}$$

   for instance, with $\alpha = 0.05$ or $\alpha = 0.01$, and

2. a test which minimizes the probability of type II error is chosen

$$p(\text{not rejecting } H_0 | H_1 \text{ is true}) = \beta \tag{2.25}$$

Following our experimental design let us assume we reject the null hypothesis when $\eta(G) > CV$, η being a statistic (i.e., graph properties) of the random graph G and CV a fixed constant. For example, by taking the density $den(G)$ and equations 2.24, 2.25:

$$p(\eta(G) > CV | H_0 \text{ is true}) = \alpha \tag{2.26}$$

$$p(\eta(G) \leq CV | H_1 \text{ is true}) = \beta \text{ minimized} \tag{2.27}$$

Density distribution $\eta(G)$ and power computation

Distributions for random graph properties are notoriously difficult to obtain, even for the simple characteristics of equations 2.26, 2.27. This problem could be solved using *Monte Carlo* method [118]. In order to employ this method we needed a model that represents the population or the phenomena of interest, and to generate sampling

realizations (random numbers). Generated data can be then studied as if they were real observations [92].

In our case, we replaced the N sample realizations of ($G^{\tilde{(1)}}$, $G^{\tilde{(2)}}$, ... , $G^{\tilde{(n)}}$) with the random realization ($g^{\tilde{(1)}}$, $g^{\tilde{(2)}}$,..., $g^{\tilde{(n)}}$) which satisfies the properties of the sample. Then we proceeded to estimate the $g^{\tilde{(n)}}$ and view them as observations from the sample distributed as $den(G)$. An intuitive way to proceed was to approximate the distribution of $den(G)$ using the histogram of the estimates ($g^{\tilde{(1)}}$, $g^{\tilde{(2)}}$,..., $g^{\tilde{(n)}}$). This way, as in Eq. 2.26 the null hypothesis was rejected when $den(G) > T_{N(1-\alpha)}$. The type II error (Eq.2.25) could be finally derived using the significance level α and the corresponding *Critical Value* $CV \approx T_{N(1-\alpha)}$. In our case we will have Type II errors a number of estimated times

$$\hat{\beta} = \frac{1}{M} \sum_{r=1}^{M} I\{X_{(N),r} \leq T_{N(1-\alpha)}\} \tag{2.28}$$

with $X_{(N),r}$ the r-th sample and I the indicator function for the event $\{X_{(N),r} \leq T_{N(1-\alpha)}\}$. Finally, the power of the test was given by $1-\tilde{\beta}$. We briefly reviewed the procedure in Algorithm 1.

Algorithm 1 Power computation

INPUT: CV
OUTPUT: p
$E \Leftarrow 0$
for $i = 1$ to M **do**
  $G \Leftarrow$ sampling (H_1); {sampling G from H_1}
  $V \Leftarrow$ MaxDegree(G); {compute max vertexes degree for G}
  if $V < CV$ **then**
    $E \Leftarrow E + 1$; {an error occurred: H_0 not rejected}
  end if
end for
$\tilde{\beta} \Leftarrow E/M$; {Type II Error Estimation}
$p \Leftarrow 1 - \hat{\beta}$

2.2.3 Analysis of Correlation Structures

A number of criticisms can be raised about the method of divergence just described, particularly in relation to the use of the KL divergence and the use of Random Graphs for the simulation of biological and experimental noise required to test the robustness of the system (see §4.2.1). Consequently, we have archived the use of RGs and the KL divergence to retain only those points which we consider to be less critical in our process, such as the use of graphs for the representation of relations and for the composition of a network R^{obs}, of global and local density, neighborhoods and in particular the evaluation of test results.

The analysis of Correlation Structures in Renal Cell Carcinoma datasets that we performed was therefore managed using a good part of the concepts already introduced and some notions that we will introduce below, accounting also the importance of constrained classification: we constrained Type I error (α) the ensure the best classification error for for the most important class, a common practice useful to improve the ability to discriminate between case and controls [15].

This section summarizes what we will see regarding the analysis of correlation structures:

1. Use of a bipartite graph $R^{(obs)}$: the edges between vertices are no longer tracked using the KL divergence, but the correlation $p^{subj}_{m_i,m_j}$. Tools: bipartite graphs, Pearson correlation.

2. Subdivision of the graph obtained in a low number of adjacent and non-overlapping regions: de facto, a division of *spectra* into also-adjacent-and-non-overlapping regions. Tool: regions size ($k_{(regions)}$ "radius").

3. Evaluation of regions properties: estimation of the distribution of the values of local density through neighborhoods sampling. Tools: neighborhood, local density.

These steps are performed for each individual datasets: controls, ccRCC, not-ccRCC. Results are there compared:

4. Tests of hypotheses between different datasets: the values distributions previously sampled are compared by t-test. We try to isolate regions that show a different behaviour in the comparison between two different datasets (rejection of the null hypothesis). Tools: hypothesis tests, t-test.

Correlation structures

In the divergence analysis we represented the divergence structure through a graph whose edges were traced by calculating the difference between the different m/z signals (peptides) in the spectrum. The divergence was calculated using the Kullback-Leibler divergence entropy. This time we represented graphs through correlations (or anti-correlations) between the recorded signals in the spectra. This new representation (*template*) was called *correlation structure*.

The formal aspect was almost identical to what we saw in the analysis of divergence (see previous section): definitions for $\mathcal{I}^{(subj)}$ and the

templates $R^{(subj)}$ were identical. The only difference concerns the definition of the edges: if a measure of *correlation* $|Corr(\tilde{f}^{(subj,1)}_{I_{m_i}}, \tilde{f}^{(subj,2)}_{I_{m_j}})|$ between empirical distribution $\tilde{f}^{(subj,1)}_{I_{m_i}}$ and $\tilde{f}^{(subj,2)}_{I_{m_j}}$ exceeds a threshold δ, than $(m_i, m'_j) \in E$. The template $R^{(subj)}$ was calculated sampling $I^{(subj,1)}_{m_i}$ and $I^{(subj,2)}_{m_j}$ from each pair $(\mathcal{I}^{(subj)}_{m_i}, \mathcal{I}^{(subj)}_{m_j})$. We employed the **Pearson correlation** as measures of correlation [119]:

$$p^{subj}_{m_i,m_j} = \frac{\sum_{k=1}^{n}(I^{(subj)}_{m_{i,k}} - \overline{I^{(subj)}_{m_i}})(I^{(subj)}_{m_{j,k}} - \overline{I^{(subj)}_{m_j}})}{\sqrt{\sum_{k=1}^{n}(I^{(subj)}_{m_{i,k}} - \overline{I^{(subj)}_{m_i}})^2}\sqrt{(I^{(subj)}_{m_{j,k}} - \overline{I^{(subj)}_{m_j}})^2}} \geq \delta \quad (2.29)$$

Regions

Correlation structures were constructed splitting spectra, then graphs, into an arbitrary number of distinct regions S, contiguous and not overlapped. Each region was characterized by the distribution of local density values $D = \{d_1, d_2, ..., d_n\}$, sampled using neighborhoods.

In divergence analysis, we used the graph properties (the distribution of local density values generated by perturbation) as *test statistics*: every single neighborhood was assessed via test powers, one test per neighborhood. Instead, in correlation structure analysis, we used regions properties (the distribution of local density D) as test statistics.

Student's t-distribution and t-test

In the analysis of divergence the *test statistics* $\eta(G)$ was the density distribution obtained perturbing the R^{obs}. In this analysis of

Correlation Structures we provided the distribution of density values by sampling each region with neighborhoods. We can then employ a more classic statistic test based on Student's t-distribution with $n-1$ degree of freedom, published by WS Gosset during his working period at the Guinness brewery in Dublin, Ireland [120].

To *estimate* the mean of a normal population μ using the sample mean $\bar{X}$, usually the standard deviation of statistical population σ, more accurately $\sigma/\sqrt{n}$ (95% confidence), is unknown, but could be estimated using the standard deviation of the sample, s (sample size: n). The estimation of t-distribution is similar to that of the standard or Z distribution

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{2.30}$$

with the use, as mentioned, of s in place of σ:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} \tag{2.31}$$

Compared to the normal distribution, the Student's t has a greater dispersion, and consists of a family of distributions that vary with the sample size n; if the sample is very large the distribution becomes approximately equivalent [121].

A classic two-sample, paired ***t-Test*** was applied, rejecting the null hypothesis if $|t|$ is greater than the quantile of Student's t-distribution with $n-1$ degrees of freedom:

$$|t| > t_{1-\alpha/2}(n-1) \tag{2.32}$$

Hypothesis testing

We conducted three different hypothesis tests (Neyman-Pearson), each involving different set of data:

1. "Control versus ccRCC" (CVR):

$$\begin{aligned} H_0 &: \mu_S^{ctrl} = \mu_S^{ccRCC} \\ H_1 &: \mu_S^{ctrl} \neq \mu_S^{ccRCC} \end{aligned} \tag{2.33}$$

2. "Control versus not-ccRCC" (CVNR):

$$\begin{aligned} H_0 &: \mu_S^{ctrl} = \mu_S^{not-ccRCC} \\ H_1 &: \mu_S^{ctrl} \neq \mu_S^{not-ccRCC} \end{aligned} \tag{2.34}$$

3. "ccRCC versus not-ccRCC" (RVNR):

$$\begin{aligned} H_0 &: \mu_S^{ccRCC} = \mu_S^{not-ccRCC} \\ H_1 &: \mu_S^{ccRCC} \neq \mu_S^{not-ccRCC} \end{aligned} \tag{2.35}$$

Compared to the analysis of divergence, to be note the disappearance of the test against the random sets and the differentiation of cases in the histological subgroups ccRCC and not-ccRCC. All tests involve the same procedure:

- the representation of R^{obs} by sampling the distribution functions ($f_{\mathcal{I}_{m_i}}^{(subj)}$, $f_{\mathcal{I}_{m_j}}^{(subj)}$), using $\{I_{m_{i,1}}^{obs}, I_{m_{i,2}}^{obs}, ..., I_{m_{i,n}}^{obs}\}$ and $\{I_{m_{j,1}}^{obs}, I_{m_{j,2}}^{obs}, ..., I_{m_{j,n}}^{obs}\}$ (see above, "Correlation structures");
- the collection of observations on the local density of the region S: $D_S^{obs} = \{den(M_1^{obs}), den(M_2^{obs}), ..., den(M_n^{obs})\}$. These

collections describe the t-distribution useful to perform the different tests.

2.2.4 Characterization of Distinguishing Regions

The main flaw of the analysis of correlation structures is that it does not provide a tool for the evaluation of the validity of the decisions made. In addition, the Pearson correlation has some flaws: for example, it miss the correct estimation of nonlinear relationships [122]. We tested then if it was possible to improve the method by varying the metric used to trace edges between vertices, preserving all that remains.

Here we provide a summary of the Distinguishing Regions approach: it is almost identical to the previous one (Analysis of Correlation Structures). To note the replacement of the correlation with the Mutual Information and the use of Fisher's exact test:

1. Use of a bipartite graph $R^{(obs)}$: the edges between vertices are no longer tracked using Pearson correlation, but MI. Tools: bipartite graphs, MI.

2. Subdivision of the graph (spectrum) obtained in a low number of adjacent and non-overlapping regions. Tool: regions size ($k_{regions}$ "radius").

3. Evaluation of regions properties: estimation of the distribution of the values of local density through neighborhoods sampling. Tools: neighborhood, local density.

These steps are performed for each individual datasets: controls, ccRCC, not-ccRCC. Results are there compared:

4. Tests of hypotheses between different datasets: the values distributions previously sampled are compared by t-test. We isolated regions rejecting the null hypothesis. Tools: hypothesis tests, t-test.

5. we tested new data, using samples not yet used during the analysis and the same parameters: with the results we filled the contingency table and calculate the Fisher's exact test.

Mutual Information

The characterization of Distinguishing Regions is based on Mutual Information (see also section 1.3.3), which replaces the Pearson correlation: the existence of an edge, that is a relationship between two vertices (peptides), is assessed on the basis of mutual dependence of the signal intensity observed. This is similar to that we discussed about *Divergence representation* (§2.2.2) and *Correlation structures* (§2.2.3): each group of subjects $\mathcal{I}^{(subj)}$ (controls, ccRCC or not-ccRCC; see §2.2.1) can be expressed through a product $\mathcal{I}_{m_1}^{(subj)} \times \mathcal{I}_{m_2}^{(subj)} \times ... \times \mathcal{I}_{m_n}^{(subj)}$ of spaces $\mathcal{I}_{m_i}^{(subj)}$, $i \in \{1, ..., n\}$, given by all potential intensity values whose m/z ratio is m_i. We also assume that each $\mathcal{I}_{m_i}^{(subj)}$ is associated to a distribution function $f_{\mathcal{I}_{m_i}}^{(subj)}$.

Mass spectra supply continuous data for which probability distributions $f_{\mathcal{I}_{m_i}}^{(subj)}$ are unknown and should be estimated. However, the concept of MI was initially developed for discrete data. The most

used technique useful to estimate MI from discrete distribution is the histogram estimation [123]. The calculation of mutual information is, therefore, based on the binning of data into M discrete intervals $a_k, k \in \{1, ..., M\}$. For any group g and m/z ratio m_i, our experimental data consist of j measurements of intensities $\{i^g_{m_{i,1}}, i^g_{m_{i,2}}, ..., i^g_{m_{i,j}}\}$. An indicator function $\Theta_{i^g_{m_{i,u}}, u \in a_k}, u \in \{1, ..., j\}$, can be employed to count the number of data points within each bin a_k. The probabilities are then estimated based on the relative frequencies of occurrence $\tilde{p}(a_k) = \frac{1}{n} \sum_{u=1}^{n} \Theta_{i^g_{m_{i,u}}, u \in a_k}$, where $\Theta_{i^g_{m_{i,u}}, u \in a_k} = 1$ if $i^g_{m_{i,u}} \in a_k$, else $= 0$.

Let $\{i^g_{m_{i,1}}, i^g_{m_{i,2}}, ..., i^g_{m_{i,j}}\}$ and $\{i^g_{m_{t,1}}, i^g_{m_{t,2}}, ..., i^g_{m_{t,j}}\}$ be two sets of observations obtained by sampling $f^g_{I_{m_i}}$ and $f^g_{I_{m_t}}$, for each $i, t \in \{1, ..., n\}$ and every group g. We call the *template* of g R^g a bipartite graph $(V_1 \cup V_2, E)$ with $V_1 = \{m_1, m_2, ..., m_n\}$, $V_2 = \{m'_1, m'_2, ..., m'_n\}$: if a measure of Mutual Information exceeds a threshold δ, than $(m_i, m'_t) \in E$:

$$\sum_{k=1}^{M} \sum_{l=1}^{M} \tilde{p}(a_k, b_l) \log_2 \frac{\tilde{p}(a_k, b_l)}{\tilde{p}(a_k)\tilde{p}(b_l)} \geq \delta \tag{2.36}$$

Fisher's exact test

Fisher's exact test is a statistical significance test useful for evaluating contingency tables (see table 2.4; [124]). It is usually applied for the study of small samples, which can not be evaluated with the distribution χ^2 [124, 125].

The Fisher's exact test allows to calculate the p-value

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!} \tag{2.37}$$

We use Fisher's exact test to verify that the property of the Regions (distinguish or not distinguish) and the test results are associated.

a	b	a+b
c	d	c+d
a+c	b+d	a+b+c+d = n

Table 2.4: A contingency table, as described by Fisher

2.2.5 Robust Conclusions in MS Analysis

The evaluation of the quality of the results obtained, the idea that led us to introduce the Fisher's exact test in the last section, prompted us to investigate the robustness of our approach. We have thus constructed a new approach based on the previous ones, because:

- retains all of the tools used in the last approach: bipartite graphs, neighborhoods, local/global density, regions, the use of Mutual Information for template construction, the CVR, CVNR, RVNR hypothesis tests;
- recovers some of the interesting items that we have lost by the wayside: Random Graphs (RGs; see §1.4.3), used to investigate the robustness of the decisions the method performed.

This approach is built from lessons learned from previous approaches, which can be summarized as follows:

1. Use of a bipartite graph $R^{(obs)}$ (template): the edges between vertices are tracked using Mutual Information. Tools: bipartite graphs, MI.
2. Subdivision of the graph (spectrum) in regions. Tool: regions size ($k_{(regions)}$ "radius").
3. Evaluation of regions properties: estimation of the distribution of the values of local density through neighborhoods sampling. Tools: neighborhood, local density.
4. Tests of hypotheses between different datasets (CVR, CVNR, RVNR). We isolated regions rejecting the null hypothesis. Tools: hypothesis tests, t-test.

Analysis of robustness:

5. creation of a pool of RGs, perturbed graphs based on templates $R^{(obs)}_{ctrl}$, $R^{(obs)}_{ccRCC}$, $R^{(obs)}_{not-ccRCC}$. Tools: RGs, global and local density.

6. method testing, employing samples not yet used during the analysis, and the same parameters: with the results we filled the contingency table and calculate the Fisher's exact test p-values.

Robustness

Uncertainty characterizes many experimental processes and may change the analysis of the events being investigated. For this reason, robustness analysis needs to be considered in an appropriate manner [126–128].

Robustness (see also section 1.4.7), here, is specifically defined as the persistence of statistical procedures (i.e., test of hypotheses) against graph property perturbations. Different graph perturbation approaches are employed in the literature to compute graph properties as the graph undergoes some perturbation or change (see for examples [129, 130]). This perturbation may represent new knowledge, reinterpretation of old data, or exploratory "what-if" type scenario. Differently, we verify (empirically) the persistence of conclusions (decisions) for the considered statistical procedures when the mechanism of perturbations (i.e.,the reference model of variability) is applied. In other words, we observe whether the statistical procedures (test of

hypotheses) still preserve their decisions even though a source of variability affects the observed (data) reference model.

We should also say that, similarly to *sensitivity analysis*, which consists of studying how a given solution changes when the reference model is perturbed, we investigate how the decisions of statistical tests are due to changes in the variability reference model parameters [131]. Solutions are given through statistical procedure decisions (i.e., test of hypotheses decisions), and perturbations over the data reference model (i.e., graphs) are provided from the considered variability model (i.e., RGs).

Reference model of variability

We define a model of variability through RGs instead of deal with parameters modifications of the reference method. This way we provide a perturbation mechanisms for our data reference model (i.e., template), thus permitting us to obtain information about the validity of the proposed conclusions (i.e., statistical test decisions) for a set of acceptable parameters.

In order to define a reference model of variability, we introduce stochastic elements in our analysis: starting from an observed template $R^{(obs)} = (V_1 \cup V_2, E)$ we wish to define a random graph able to preserve, within a defined range, a property of the template itself. We analyze the neighborhood cohesions of an observed graph (i.e., template) $R^{(obs)}$. Hence, we attempt to preserve the densities of $R^{(obs)}$ by preserving (on average) its size. Among the many methods for defin-

ing a RG from any observed graph while preserving some properties (see, for example, the problem of graph randomization in [130]), here we randomly modify (additions or deletions) the edges of RGs. We realize this perturbation in such a way that the expected number of modifications takes values inside specific ranges. The following definition formalizes this mechanism:

Definition 1 $(s, t, R^{(obs)})$**-Preserving Random Graph**
Let $R^{(obs)} = (V_1 \cup V_2, E)$ be a template. We consider the following experiment: for any $e \in V_1 \times V_2$, if $e \in E$ we delete e with probability p. Otherwise, if $e \notin E$, we add e to E with probability p. We say that this mechanism defines an $(s, t, R^{(obs)})$-preserving RG $\tau(s, t)$ if the expected number of edge additions and deletions in RG take values in $[s, t]$.

Property 1 Let $R^{(obs)} = (V_1 \cup V_2, E)$ be a template. We should obtain an $(s, t, R^{(obs)})$-preserving RG $\tau(s, t)$ by constraining the perturbation probability p in definition 1 in such a way that $\frac{s}{n^2} \leq p \leq \frac{t}{n^2}$, where $n^2 = |V_1 \times V_2|$.

Test of hypotheses can be formulated also when the perturbation mechanism in definition 1 is applied, using the Monte Carlo method for instance [118], defined two templates $R^{(ctrl)}$ and $R^{(ccRCC)}$ we generate, for any pair of regions R_1 and R_2, two Monte Carlo samples:

- n realizations $\{\tilde{\tau}_1^{(1)}, \tilde{\tau}_1^{(2)}, ..., \tilde{\tau}_1^{(n)}\}$ of $\tau_1(s, t)$;
- n realizations $\{\tilde{\tau}_2^{(1)}, \tilde{\tau}_2^{(2)}, ..., \tilde{\tau}_2^{(n)}\}$ of $\tau_2(s, t)$

where $\tau_1(s,t)$ is a (s,t,S_1)-Preserving Random Graph and $\tau_2(s,t)$ is a (s,t,S_2)-Preserving Random Graph.

Chapter 3

Main results

3.1 Mass Spectrometry data alignment

We aligned proteomics data from the three laboratories involved in our work (section 2.1.2), using the three methods (Mutual Information, Equal Mass and t-test based methods) mentioned in sections 2.1.3 and 2.1.4.
The performance on the datasets has been tested using `RapidMiner`: in particular the process described in the section 2.1.5.
Results are reported in the following pages, summarize by the performances of the SVM classifier employed (see section 2.1.5), and specifically using:

- the Area Under ROC Curve (**AUC**; see §1.3.3), that could be considered a good indicator of the diagnostic power of a method (here to be understood as the ability to detect biomarkers) [72].

- **Precision**, the fraction of objects truly relevant compared to

the number of objects classified as relevant:

$$Precision = \frac{true\,positive}{true\,positive + false\,positive}$$

- **Recall**, the fraction of identified relevant objects over the total number of objects that should have been identified as relevant:

$$Recall = \frac{true\,positive}{true\,positive + false\,negative}$$

The results are summarized in charts and tables, listed below. Some of the charts and tables discussed here are listed in Supplementary Material section (6.1). Specifically, we selected to show:

- the average values of AUC, Precision and Recall at different parameter k (number of features selected). We considered two different series of values, one from the alignments between pairs of labs (Monza and Florence, MF; Monza and Brescia, MB; Florence and Brescia, FB) and one for the alignment of all the labs involved (Monza and Florence and Brescia, MFB). See tables 3.1 (AUC), 3.2 (Precision and Recall, MF, MB, and FB), 3.3 (Precision and Recall, MFB), and related charts.

- the difference (percentage) between the various methods evaluated, obtained by comparing the mean values of AUC, Precision and Recall with the the parameter k. The comparison includes all types of alignment (MF, MB, FB, and MFB). See tables 3.4 (AUC), 6.1 (Precision and Recall), and relative charts.

- the difference (percentage) between the various methods evalu-

ated, obtained by comparing the mean values of AUC, Precision and Recall in the same labs. See tables 3.5 (AUC), 6.2 (Precision and Recall), and related charts.

AUC - Area Under ROC Curve						
k	Pairs of labs			All 3 labs (MFB)		
	EM	MI	TT	EM	MI	TT
2	0.774	0.805	0.796	0.730	0.660	0.467
3	0.722	0.827	0.751	0.632	0.720	0.660
4	0.839	0.809	0.782	0.515	0.824	0.641
5	0.681	0.881	0.722	0.499	0.759	0.749
6	0.621	0.854	0.754	0.507	0.538	0.573
7	0.832	0.877	0.766	0.541	0.776	0.467
8	0.797	0.854	0.844	0.506	0.755	0.652
9	0.612	0.836	0.784	0.477	0.563	0.629
10	0.744	0.821	0.775	0.525	0.547	0.653
11	0.766	0.851	0.766	0.452	0.691	0.605
12	0.670	0.873	0.769	0.578	0.547	0.660
Mean	0.733	0.845	0.773	0.542	0.671	0.614

Table 3.1: Mean values of AUC (§1.3.3) depending on the number k of features selected, both in the case of alignment between pairs of labs (Monza and Florence, MF; Monza and Brescia, MB; Florence and Brescia, FB), and between all the three labs involved (Monza, Florence and Brescia, MFB).
Data are represented in figure 3.1.

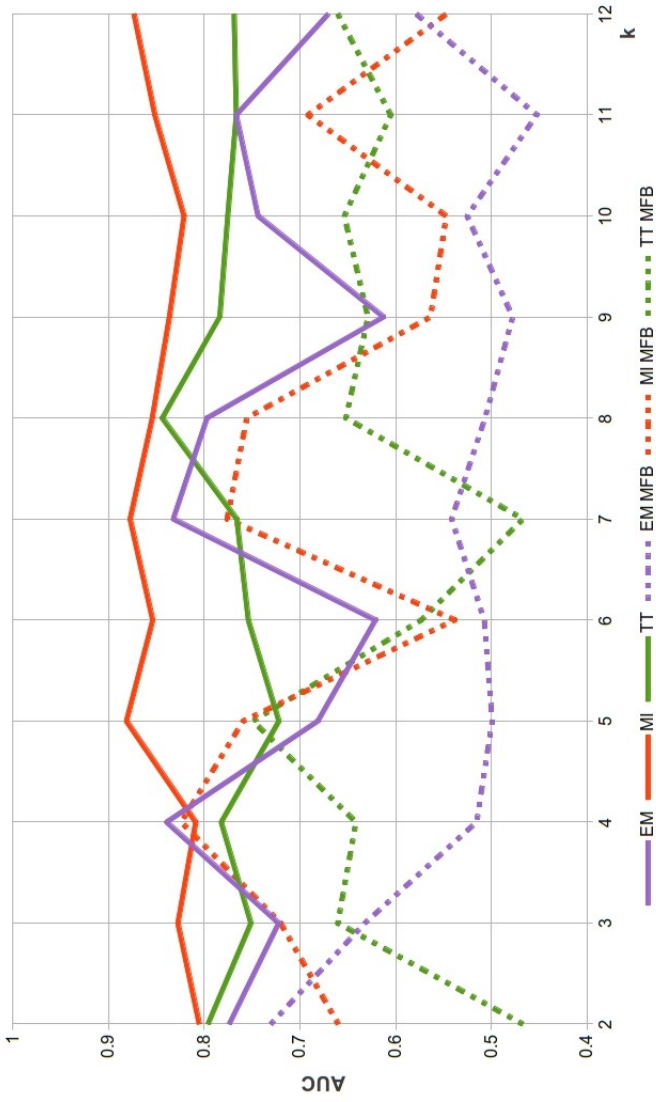


Figure 3.1: Mean values of AUC (§1.3.3) depending on the number k of features selected.
Results from the alignment between pairs of labs are draw in solid lines; dashed lines for results from the alignment of all labs involved (Monza, Florence and Brescia; MFB). Mutual Information shows virtually always better performance compared with other methods, and in both types of alignment analysed. This data are reported in table 3.1.

SVM performance - pairs of labs (MF, MB, FB)						
k	Precision			Recall		
	EM	MI	TT	EM	MI	TT
2	0.575	0.703	0.690	0.677	0.657	0.633
3	0.651	0.675	0.711	0.533	0.840	0.571
4	0.643	0.653	0.635	0.679	0.663	0.644
5	0.562	0.707	0.640	0.419	0.788	0.546
6	0.541	0.704	0.611	0.572	0.732	0.608
7	0.622	0.732	0.668	0.624	0.781	0.607
8	0.550	0.707	0.672	0.573	0.724	0.650
9	0.568	0.699	0.597	0.456	0.699	0.587
10	0.549	0.696	0.599	0.620	0.684	0.592
11	0.546	0.658	0.548	0.569	0.625	0.553
12	0.487	0.712	0.507	0.522	0.715	0.533
Means	0.572	0.695	0.625	0.568	0.719	0.593

Table 3.2: Mean values for Precision and Recall, depending on the number k of features selected, only for alignment between two labs (Monza and Florence; Monza and Brescia; Florence and Brescia). This data are represented in figure 3.1. For the three labs alignment, see table 3.3.

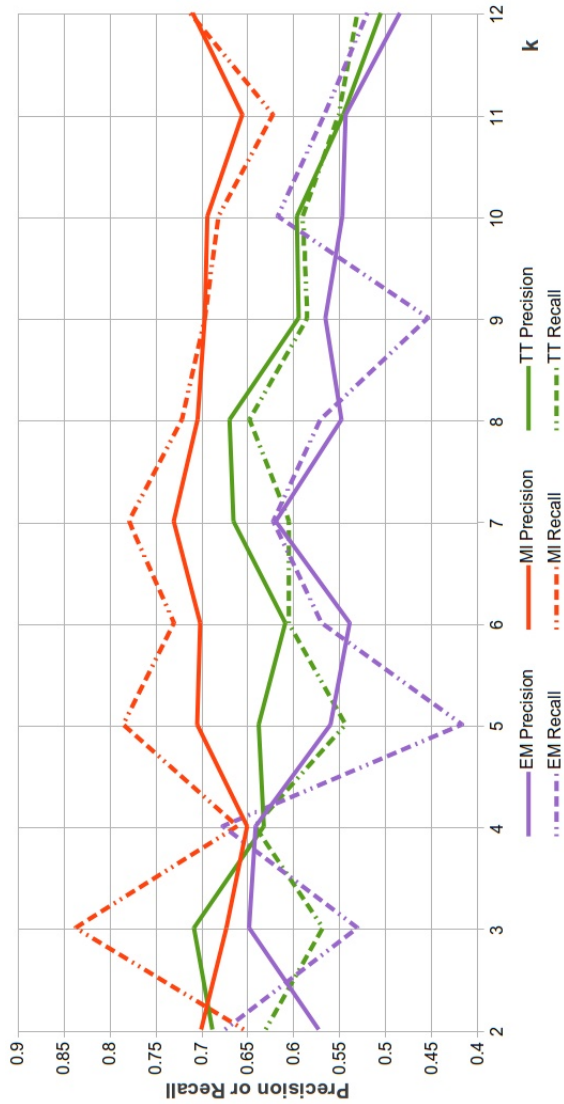


Figure 3.2: Mean values for Precision and Recall, depending on the number k of features selected, only for alignment between two labs (Monza and Florence; Monza and Brescia; Florence and Brescia). This data are reported in table 3.2. For the three labs alignment, see figure 3.1.

SVM performance - All 3 labs (MFB) involved						
k	Precision			Recall		
	EM	MI	TT	EM	MI	TT
2	0.580	0.545	0.356	0.721	0.630	0.400
3	0.514	0.598	0.539	0.536	0.645	0.530
4	0.330	0.735	0.528	0.288	0.661	0.561
5	0.432	0.619	0.554	0.473	0.536	0.661
6	0.410	0.444	0.389	0.376	0.497	0.403
7	0.472	0.731	0.451	0.409	0.503	0.473
8	0.556	0.617	0.518	0.600	0.530	0.652
9	0.481	0.487	0.533	0.688	0.530	0.685
10	0.517	0.526	0.502	0.506	0.467	0.594
11	0.374	0.657	0.527	0.370	0.473	0.533
12	0.461	0.476	0.495	0.630	0.342	0.533
Mean	0.466	0.585	0.490	0.509	0.529	0.548

Table 3.3: Mean values for Precision and Recall, depending on the number k of features selected, only for alignment between all the three labs (Monza and Florence and Brescia, MFB).
This data are represented in figure 3.1. For alignments between pairs of labs, see table 3.2.

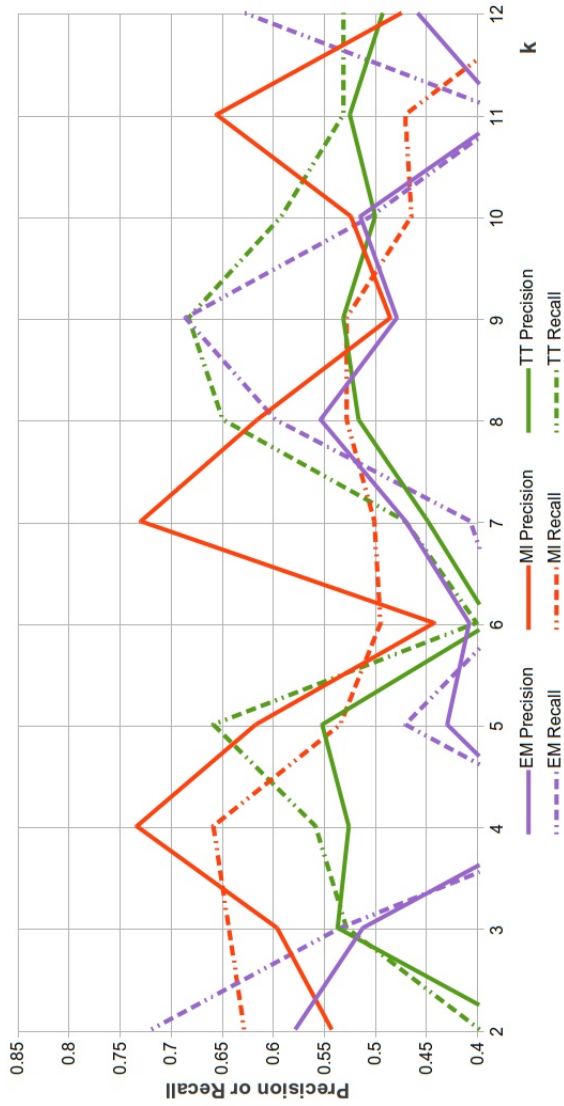


Figure 3.3: Mean values for Precision and Recall, depending on the number k of features selected, only for alignment between all the three labs (Monza and Florence and Brescia, MFB).
This data are reported in table 3.3. For alignments between pairs of labs, see figure 3.1.

AUC - All labs (MF, MB, FB and MFB)			
k	MI vs EM	MI vs TT	TT vs EM
2	0.75%	7.14%	-6.89%
3	12.62%	9.02%	3.96%
4	6.69%	8.08%	-1.52%
5	25.28%	14.35%	12.77%
6	23.53%	8.54%	16.39%
7	10.86%	18.90%	-9.92%
8	12.68%	4.07%	8.98%
9	24.67%	3.02%	22.33%
10	8.44%	1.01%	7.50%
11	15.22%	10.52%	5.25%
12	18.25%	6.32%	12.73%

Table 3.4: Performance comparison (percentage) between the AUC means of the various methods proposed, measured for each method by varying k.
Positive values indicates a better performance of the first method versus the second (for example, given A vs. B: A better than B), negative values the opposite (B better than A). Data drawn in picture 3.1.

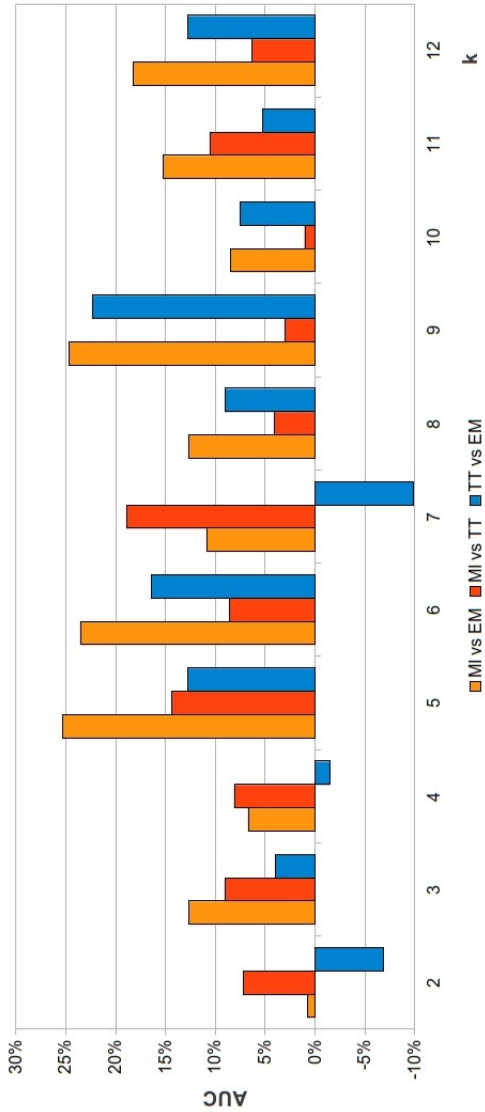


Figure 3.4: Performance comparison (percentage) between the AUC means of the various methods proposed, measured for each method by varying k.
MI (light and dark orange) performed always better than the competing methods, with peaks between 20-25%. See also table 3.4.

AUC - Labs comparison				
Methods	Monza & Florence	Monza & Brescia	Florence & Brescia	MFB
MI vs EM	3.17%	18.69%	14.93%	19.19%
MI vs TT	-6.34%	28.49%	-0.21%	8.46%
TT vs EM	8.94%	-13.70%	15.10%	11.73%

Table 3.5: Performance comparison (percentage) between the AUC means of different aligned datasets (different labs merging).
Positive values indicates a better performance of the first method versus the second (for example, given A vs. B: A better than B), negative values the opposite (B better than A). For a chart, see 3.1.

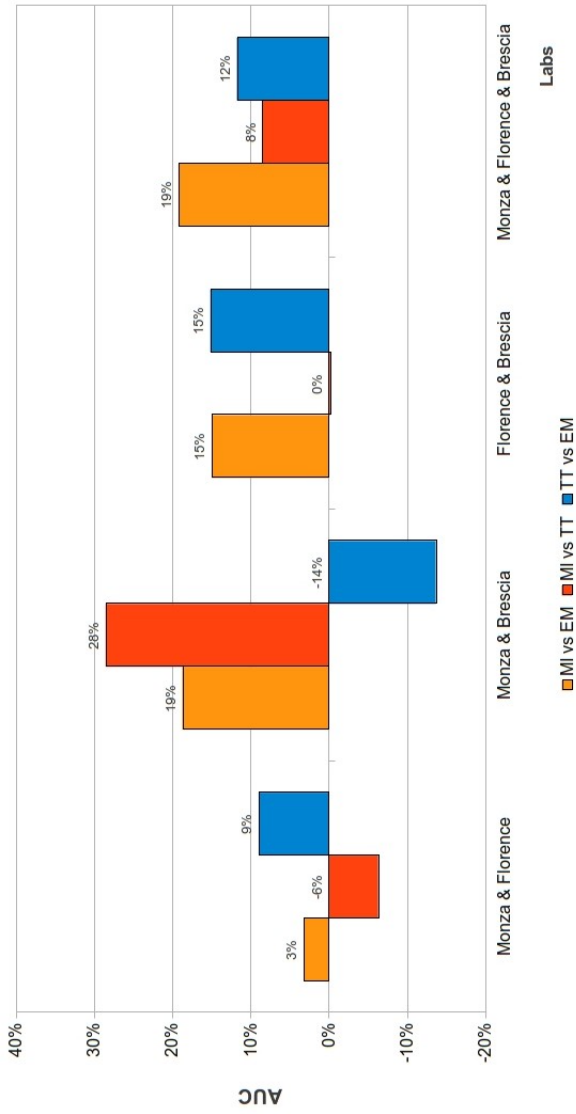


Figure 3.5: Performance comparison (percentage) between the AUC means of different aligned datasets (different labs merging).
MI (light and dark orange) performed better than the competing methods, notably in MFB, the most complex dataset. Data from table 3.5.

3.2 Mass Spectrometry data analysis

The following sections list the results obtained in Mass Spectrometry data analysis. All methods designed have been implemented using MATLAB® (MATrix LABoratory), a high-level language and interactive environment for numerical computation, visualization, and programming developed by MathWorks company [www.mathworks.com].

3.2.1 Divergence Analysis

Our purpose was to identify those part of spectrum (neighborhoods) that respond to the Neyman-Person hypothesis tests rejecting the null hypothesis.
As often happens in these cases, we have imposed a predefined level for type I error (usually $\alpha = 1\%$, otherwise reported), so to ensure tests with high statistical significance: consequently the final results highlight those neighborhoods that reject the null hypothesis with higher test powers (1 - β).

Parameters evaluation

We studied the free parameters of the algorithm:

1. the Kullback-Leibler divergence threshold δ, or KL threshold, that define, for each pair of nodes, the presence or absence of an edge;

2. the parameters related to perturbation: the number of perturbed graphs;

3. the size k of the neighborhoods.

This analyses have been repeated independently for each one of the three hypothesis test type designed:

1. $\mathrm{H}_0 : \mathrm{URfM}(v, e)$ and $\mathrm{H}_1 : R^{(case)}$ (random versus case);

2. $\mathrm{H}_0 : \mathrm{URnM}(v, e)$ and $\mathrm{H}_1 : R^{(control)}$ (random versus control);

3. $\mathrm{H}_0 : R^{(control)}$ and $\mathrm{H}_1 : R^{(case)}$ (control versus case).

The Kullback-Leibler (KL) threshold δ

As a first step we established a set of arbitrary KL thresholds, then we counted, for each threshold and for every neighborhood, the number of graphs with test powers greater than δ. This process was iterated several times with different KL thresholds and δ ($\delta > 0.25, \delta > 0.50, \delta > 0.75, \delta = 1$). These raw results helped us defining the range of best values useful to search for the optimal KL threshold: $[10, 20]$. We iterative replicated the procedure within the selected range, choosing the set of the most interesting results until we find the overall best KL threshold value (Fig.3.6). This steps was repeated for each one of the three type of hypothesis tests we performed.

The first runs of the KL threshold analysis were performed using a number of perturbations of several orders of magnitude (i.e. number of perturbed graphs), because at this point we has not yet

determined the effect of perturbation on results. Likewise, an arbitrary neighborhood size ($k = 2$) was selected. To be note that after establishing the values to be assigned to each parameter, all the tests were repeated, in order to confirm the rightness of the choices.

Number of Perturbations

Two different models of randomness have been employed: the Uniform Reference model (URfM) and Uniform Random model (URnM; see 2.2.2). The URfM is actually represented with a random graph consisting of v vertices and e edges. Since the number of vertices and edges are known, we could generate the entire population $\mathcal{P}$, or one or more graph $Pg \in \mathcal{P}$. The URnM was obtained instead by applying noise, with a normal distribution, to the $R^{(obs)}$: in both cases we could decide, at our discretion, the number of graphs Pg we wanted. To ensure that the results did not depend on chance, we repeated the test by generating multiple random graphs: as we will see shortly, if we exclude the fluctuations for small numbers of graphs, the number of random graphs does not affect the behavior of the test, demonstrating the non-dependence between the results and methods of perturbation.

For the analysis of the number of perturbed graphs Pg that must be computed we used the same principles of the last analysis, setting a type of test to be performed (case versus controls) and the KL threshold just calculated. We did the first analysis using arbitrary numbers of perturbed graphs: for example, $Pg = 100, Pg = 500$ and $Pg = 1000$. In none of our experiments there was significant

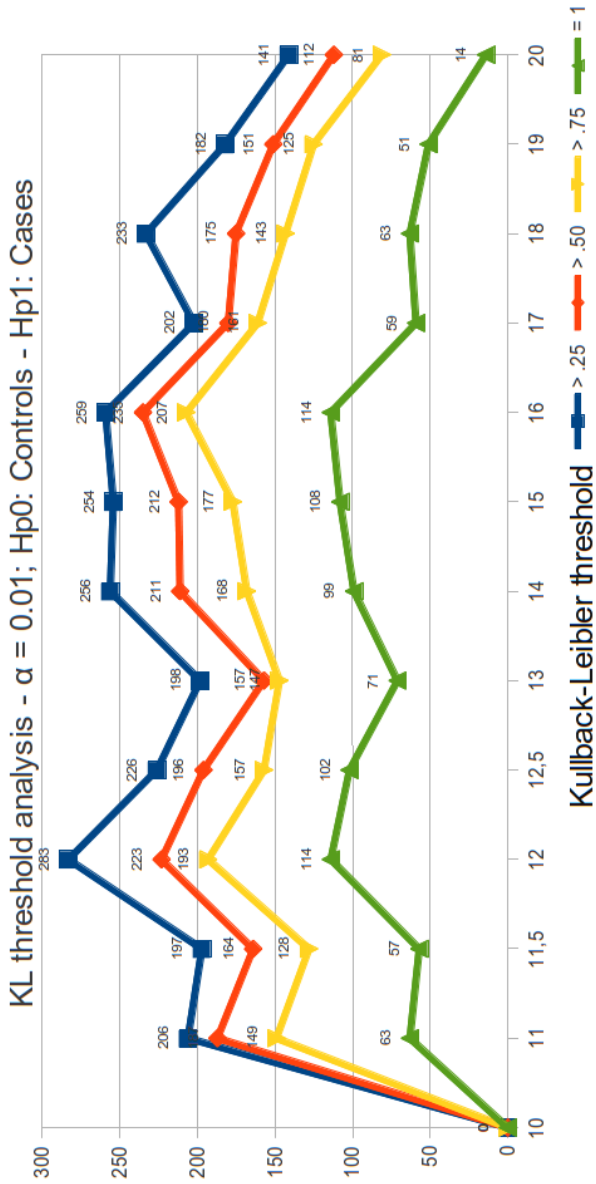


Figure 3.6: The Kullback-Leibler divergence threshold δ
Number of tests of hypotheses with test powers greater than (or equal to) arbitrary thresholds. H_0 = null hypothesis; H_1 = alternative hypothesis.

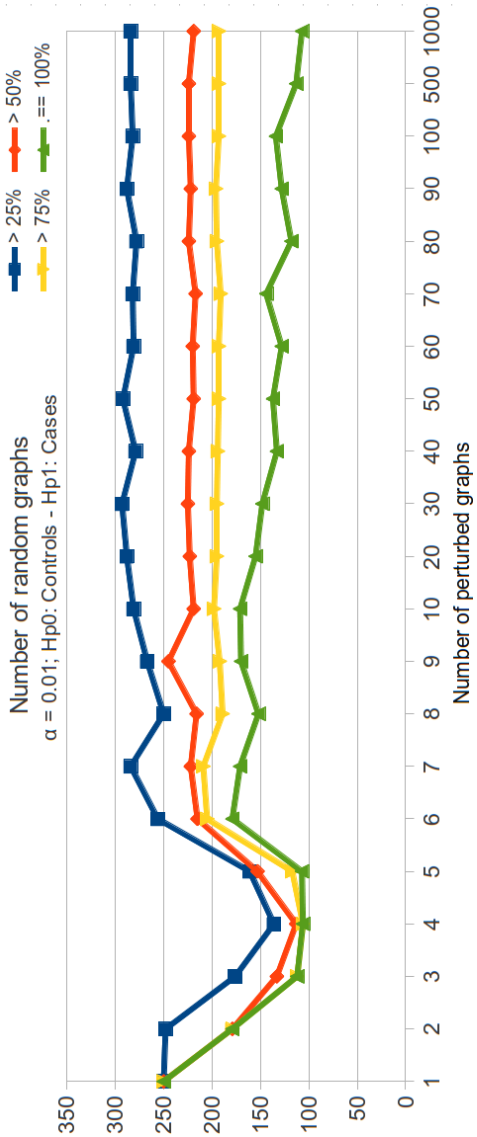


Figure 3.7: Number of perturbed graphs.
Performance of the same test (controls vs. case), varying the dimension of the population of perturbed graph. Y-axis: number of tests of hypothesis with test powers greater than the arbitrary thresholds; H_0 = null hypothesis; H_1 = alternative hypothesis.

variations between the results obtained from these Pg values (data not shown).

We then reduced search space in a smaller range, like $2 \leq Pg \leq 100$, without detecting, even in this case, dependencies between the number of perturbed graphs and results: for $Pg > 10$, the increase of the perturbation number does not affect results, (see Fig.3.7) except for what concerns the $\delta = 1$ threshold, which decreases very slowly but steadily with the increasing of the number of perturbed graphs.

All these tests were repeated several times, to verify that this behaviour was not due to chance, and that the perturbation method was adequate, achieving hard consistence (average deviation: 2.5%; data not shown). Supported by this observations, we set the number of perturbed graphs to the arbitrary value $Pg = 100$.

Neighborhood size (k)

The last parameter to be studied was related to the size of the neighborhoods. The analysis was similar to that performed for the study of KL threshold: we established an increasing set of arbitrary k size and counted how much tests had test power higher than a threshold.The value k is a "radius" that, centred on a vertex, defines the number of adjacent vertices in the neighborhood: all vertexes $[v^A_{n-k}, v^A_{n+k}]; [v^B_{n-k}, v^B_{n+k}]$. We called this window ***local density window***. Please, note that the population size of the possible local density windows decreases with the increasing of k: to avoid distortion of experimental results, is therefore necessary to weigh the absolute

values found with the population size (see Fig.3.8).

We finally selected the local density window that had the highest number of tests with high power level, in particular for more stringent thresholds ($> .75, = 1$).

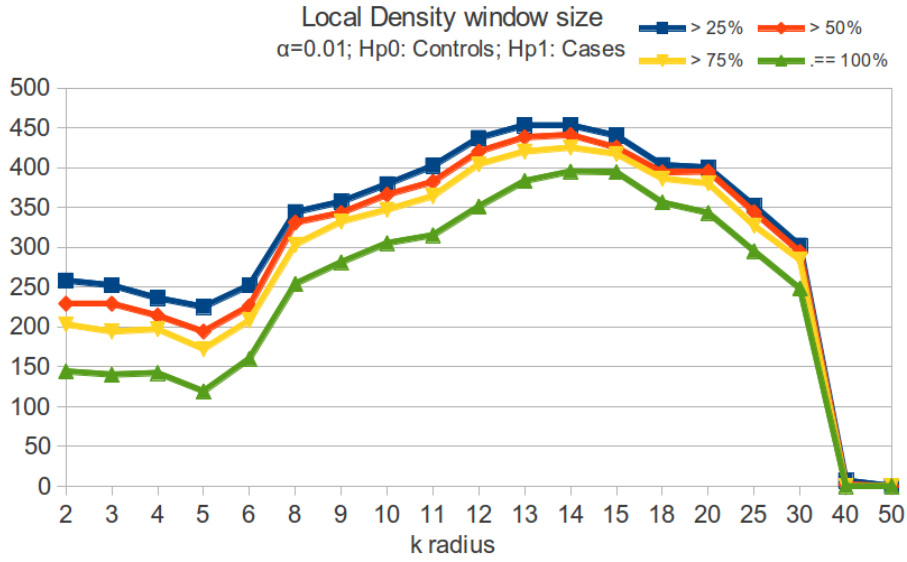


Figure 3.8: Local density window size
X-axis: Local density window radius k; Y-axis: number of hypothesis tests with test powers greater than the arbitrary thresholds; H_0 = null hypothesis; H_1 = alternative hypothesis.

Results for free parameters analysis

We did not find a set of free parameters suitable for all the tests, except for the number of perturbed graphs, setted to 100. Table 3.6 summarize the best results related to the three different hypothesis

tests.

Hypothesis test	KL threshold	Number of perturbed graphs	Local density window size
Random vs Case	7	100	13
Random vs Control	7	100	15
Control vs Case	12	100	14

Table 3.6: Best parameters for the three different tests

Heat maps and most interesting mass ranges

The next pages will show the results obtained with our method and with the parameters chosen. The results are represented using heat maps and mass ranges. Information are actually the same: results are represented by heat maps that show the hottest areas, ie tests with high power (all areas with low power were set to zero). These areas are then described by axes coordinates i and j, easy to identify and summarized in the following tables. These coordinates indicate the vertex taken as the center of the window of the local density.
The tables with mass ranges represent a discretization of the heat map information, because reported contiguous areas in which we obtain tests that rejects the null hypothesis with high test power.

In particular, the first two tests, random versus controls and random versus cases, show extremely large hot areas, to emphasize the ability of the method to distinguish an ordered set of data (healthy) compared to similar sets but perturbed (URnM), or a dataset more or less ordered (cases) compared to a system with similar general features (URfM). The main heat map, "controls versus cases", has a very different profile, because it identifies those areas of the spectrum that show a different behaviour in the two sets.

Here is the list of heat maps and tables displayed below:

1. random versus controls: Fig.3.9 (heat map) and Table 3.7;
2. random versus cases: Fig.3.10 (heat map) and Table 3.8;
3. controls versus cases: Fig.3.12 (heat map), Type I error (α) = 1%), Fig.3.11 ($\alpha = 5\%$) and Table 3.9;

areas	start & stop				masses start & stop (Da)			
	i		j		i		j	
a	25	53	24	96	2050.24	3464.28	2043.36	5924.40
b	16	77	96	109	1770.14	4543.79	5924.40	8858,15
c	55	107	37	43	3712.83	8192.67	2805.13	3164.61

Table 3.7: Random versus Controls test - Some coordinates for the best local density windows (best power test), and related mass range.

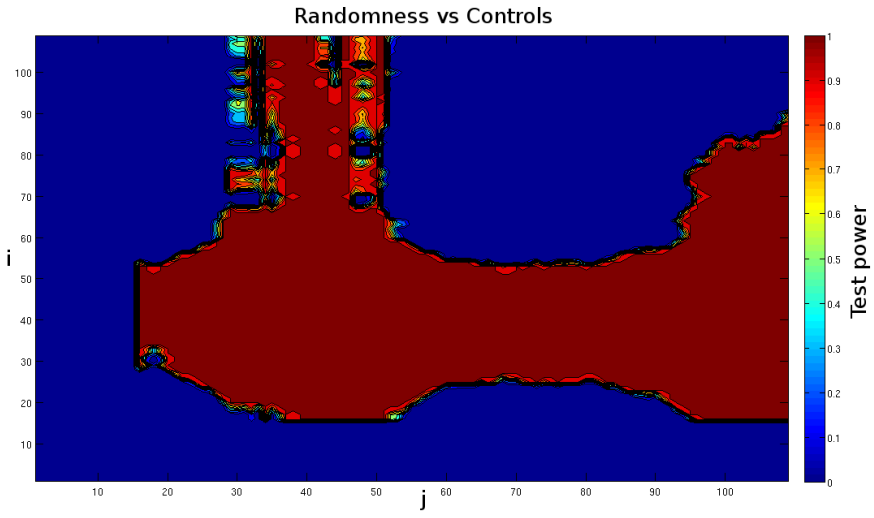


Figure 3.9: Random versus Controls test - Heat map representing the most interesting local density window coordinates

areas	start & stop				masses start & stop (Da)			
	i		j		i		j	
a	16	56	14	111	1728.8	3486.19	1666.92	8858.15
b	93	111	30	44	5237.85	8858.15	2382.26	3158.31

Table 3.8: Random versus Cases test - Some coordinates for the best local density windows (best power test), and related mass range.

areas	start & stop				masses start & stop (Da)			
	i		j		i		j	
a	23	58	97	110	1940.18	3743.34	5924.4	8858.15
b	38	41	38	46	2805.13	3017.81	2805.13	3214.45

Table 3.9: Controls versus Cases test - Some coordinates for the best local density windows (best power test), and related mass range.

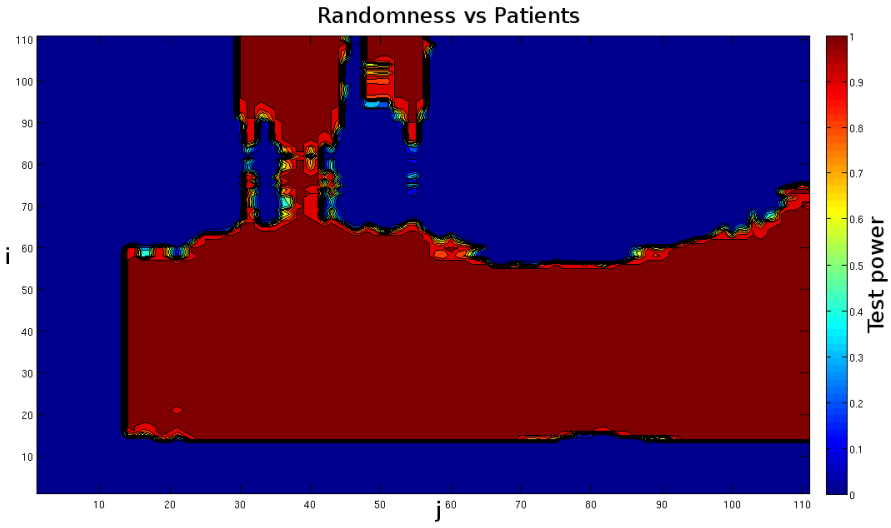


Figure 3.10: Random versus Cases test - Heat map representing the most interesting local density window coordinates

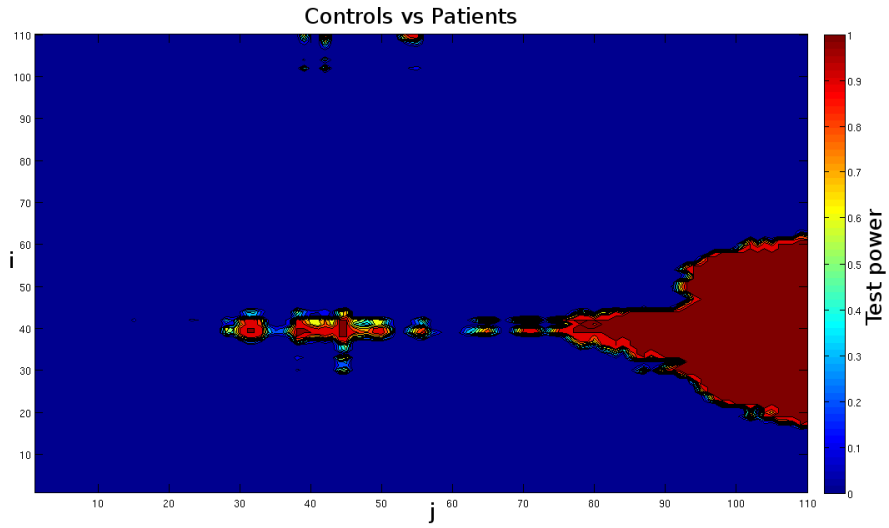


Figure 3.11: Controls versus Cases test - Heat map representing the most interesting local density window coordinates. $\alpha = 5\%$ (others heat map: $\alpha = 1\%$).

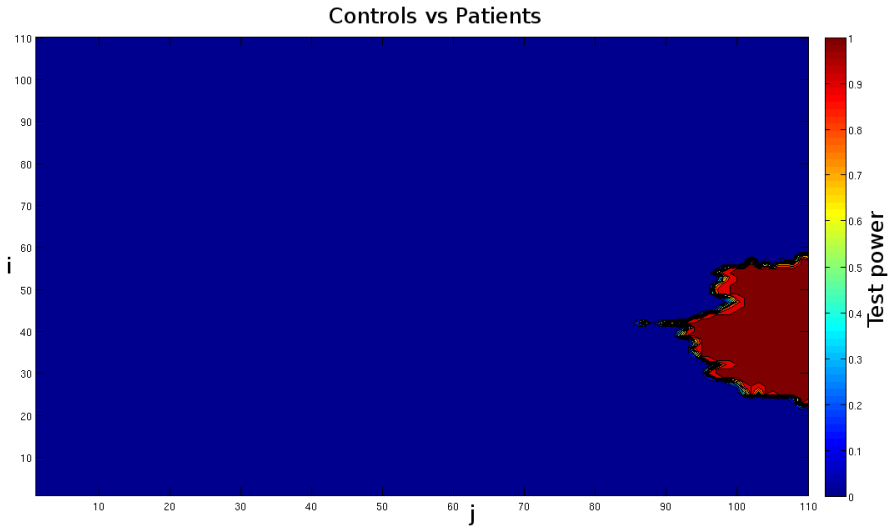


Figure 3.12: Controls versus Cases test - Heat map representing the most interesting local density window coordinates. $\alpha = 1\%$.

3.2.2 Analysis of Correlation Structures

As in divergence analysis, the purpose of our analysis of correlation structures was to detect the interesting regions of the RCC spectra. It should be noted that in this case the spectrum was actually divided into regions (see "Regions", §2.2.3): the number of regions $k_{regions}$ becomes then a first parameter to be tested. Instead, all parameters related to the perturbation of graphs disappeared.

Among the parameters to be studied we still kept the threshold δ and the neighborhoods k radius: the first no longer set the threshold of divergence that characterizes the presence/absence of an edge between two vertices, but, in a similar way, the threshold of the absolute value of *correlation* $p^{subj}_{m_i,m_j}$ (see Eq. 2.29) which, if exceeded, sets an edge between two vertices; the latter was identical, but it was limited by the size of the region sampled by neighborhoods: in practice, $max(k) \leq k_{region} - 1$.

We were no longer interested in representing directly the power of the tests (heat maps), but to select the windows of the spectrum that reject the null hypothesis, supported by powerful and reliable tests. This involved the study of the three parameters, δ, k and k_{region}.

Parameters evaluation

With the above concerns in mind, the targets of our experiments could be summarized as follows:

- the goal was to evaluate empirically $k_{regions}$, δ and k in order to detect the lowest number of correlation structure changes in

every datasets comparison. In other terms, for different pairs of $k_{regions}$, δ, k, we counted the number of significant and powerful tests that reject the null hypothesis, searching for a minimum (i.e. one rejection). The best result is shown in figure 3.13.

- In order to allow the comparison between different tests, we decided to use the same parameters for all the three different hypothesis tests (CVR, etc.). The benchmark test was selected by choosing the test that included the largest number of data: CVR.

- The number of regions used was constrained by the number of signals contained in the data (135): since the regions should not overlap, this has limited the $k_{regions}$ possible values.

- The range of values of the parameters analysed were:

  - $\delta = [0.8, 0.75]$; preliminary experiments have shown unsatisfactory results for $\delta > 0.8$.

  - $k_{region} = (2; 4; 7; 13; 22)$, that corresponds to splitting the spectrum into a number of regions respectively equal to 27, 15, 9, 5, 3;

  - k depends on k_{region} ($max(k) = k_{region} - 1$), so $k = [2, 21]$, consistently with $max(k)$. For $k_{region} = [2]$, $k = 2$; for $k_{region} = 4$, $k = [2, 3]$; for $k_{region} = 7$, $k = [2, 6]$, and so on.

- Using the values of δ and k_{region} obtained above, we finally recovered the mass-to-charge ratio bounds which identify regions where we have detected modification in a correlation structure

at high level of significance (95%) and test power (> 0.75). This mass ranges are listed in table 3.10.

	CVR $\delta = 0.75$, $k = 2$		CVNR $\delta = 0.75$, $k = 2$		RVNR $\delta = 0.75$, $k = 2$	
	From	To	From	To	From	To
m/z	1719.45	2084.34	1719.45	2084.34	4625.10	5374.00

Table 3.10: Mass Ranges (Da) of the best regions selected by our method, one region per test.
Parameters are the same, and was selected using *Ctrl VS ccRCC* datasets (see text).

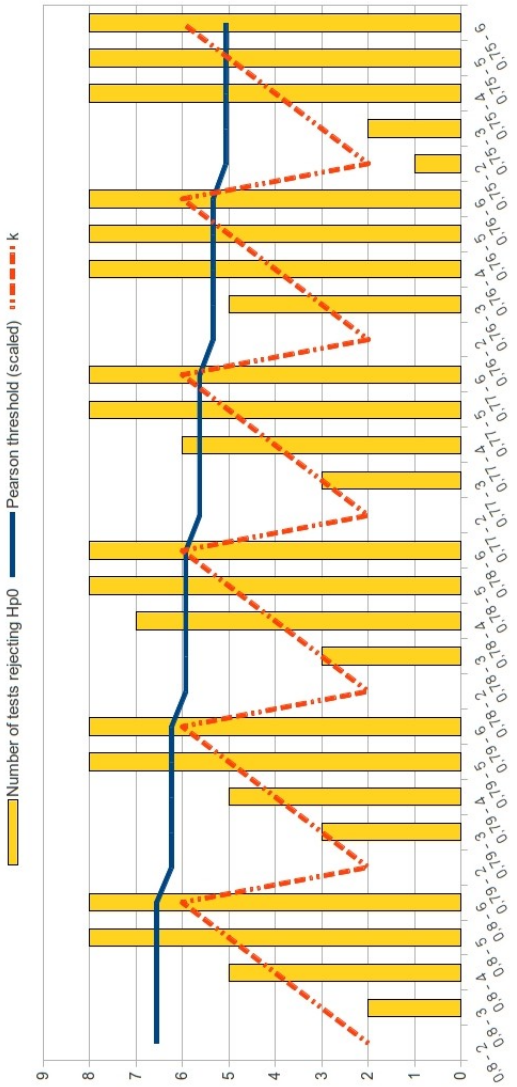


Figure 3.13: Number of rejected tests according to parameters δ and k; $k_{region} = 7$ (maximum number of rejected tests = 8; controls vs. ccRCC).
The only combination of parameter with only one test rejected is the fifth from the right: $\delta = 0.75$, $k = 2$.

3.2.3 Characterization of Distinguishing Regions

The results of the Characterization of Distinguishing Regions were obtained with the same method reported in the previous section, §3.2.2. The only differences are:

- the replacement of Pearson correlation with Mutual Information;
- the addition of the Fisher's exact test, which involves a simple computation of the distinguishing/not-distinguishing regions and calculation of the relative p-value, described in §2.2.4.

The three parameters $k_{regions}$, δ and k are also involved (§3.2.2): $k_{regions}$ and k are the same ($k_{regions} = 7$, $k = [2, 6]$), only δ changes: $\delta = [0.01, 0.00001]$.
We report only the final results of the procedure: the table of interesting mass ranges (table 3.11) and related Fisher's exact tests (table 3.2.3).

	CVR		CVNR		RVNR	
	From	To	From	To	From	To
m/z	2644.49	3214.26	1719.45	2084.34	1719.45	2084.34
m/z	3270.53	4018.88	4050.39	4540.10	1832.33	2217.20

Table 3.11: Mass Ranges (Da) of the two best regions selected by our method, two regions per test.
Parameters are always the same, and was selected using CVR (Controls vs. RCC datasets) datasets: $\delta = 0.005$; $k = 2$.

	CVR		CVNR		RVNR	
Region	H_0	H_1	H_0	H_1	H_0	H_1
DR	2	17	2	17	0	19
NDR	11	8	13	6	11	8
p-value	0.005		0.0006		0.0001	

Table 3.12: Fisher's exact test and p-values
DR: Distinguishing Region; NDR: Not-distinguishing region.

3.2.4 Robust Conclusions in MS Analysis

By introducing the reference model of variability, we provided a perturbation mechanisms for the data reference model (i.e. template). This way, we can also interpret robustness as the persistence of statistical conclusions (i.e., test of hypotheses decisions) against template property perturbations. We verified empirically the persistence of these conclusions when the perturbation mechanism is applied to the RCC data. We observed if the statistical procedures (test of hypotheses) still preserve their decisions even when a source of variability affects the observed templates (R^{obs}).

We have provided a set of arbitrary values to all the variables involved, the same of §3.2.3: $k_{regions} = 7$, $k = [2, 6]$, $\delta = [0.01, 0.00001]$. For each combination δ and k, we considered the number of significant tests rejecting the null hypothesis. For each class we evaluated (empirically) the threshold δ and ray k detecting a low number of dependence structure modifications from control to case groups. By using these values (i.e. δ and k), we detected the m/z bounds identifying modified regions over the spectra at a specific level of significance (usually 5%). This was iterated for each class of tests. We appointed regions rejecting the null hypothesis as *distinguishing regions* (DRs; not-distinguishing: NDRs).

Therefore we classified regions in distinguishing/not-distinguishing reflecting the test decisions. The aim of our analysis was to check the robustness of these decisions or, equivalently, the robustness of the distinguishing/not-distinguishing capabilities. We verified then

if, after the application of the variability model (see §2.2.5), the distinguish/not-distinguish capability was preserved: after a perturbation we still obtain new distinguishing/not-distinguishing regions, but the question of interest was to asses, for each region, the distinguishing capability before and after the perturbation. We tested this results by using the Fisher's exact test.

We have selected the parameters that allowed us to see a number of DRs = 3 ($\delta = 0.001$; $k = 4$).

	CVR		CVNR		RVNR	
	From	To	From	To	From	To
	1719.45	2084.34	1719.45	2084.34	1719.45	2084.34
m/z	2644.49	3214.26	2092.18	2563.79	2644.49	3214.26
	3270.53	4018.88	4050.39	4540.10	3270.53	4018.88

Table 3.13: Mass Ranges (Da) of distinguishing regions (DRs) selected by our method, three regions per test.
Parameters are always the same, and was selected using CVR (Controls vs. RCC datasets) datasets.

CVR								
Perturbation prob.	0.05		0.1		0.2		0.3	
After perturbation	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1
Before perturbation								
DRs	0	12	0	12	1	11	1	11
NDRs	19	1	13	7	12	8	12	8

Table 3.14: Fisher's exact test for CVR class
DRs: Distinguishing Regions; NDRs: Not-distinguishing regions.

CVNR								
Perturbation prob.	0.05		0.1		0.2		0.3	
After perturbation	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1
Before perturbation								
DRs	0	12	0	12	0	12	0	12
NDRs	20	0	17	3	17	3	12	8

Table 3.15: Fisher's exact test for CVNR class
DRs: Distinguishing Regions; NDRs: Not-distinguishing regions.

RVNR								
Perturbation prob.	0.05		0.1		0.2		0.3	
After perturbation	H_0	H_1	H_0	H_1	H_0	H_1	H_0	H_1
Before perturbation								
DRs	0	12	0	12	0	12	0	12
NDRs	19	1	15	5	12	8	15	5

Table 3.16: Fisher's exact test for RVNR class
DRs: Distinguishing Regions; NDRs: Not-distinguishing regions.

	p-values		
Perturbation prob.	CVR	CVNR	RVNR
0.05	5.75×10^{-8}	4.43×10^{-9}	5.75×10^{-8}
0.1	2.23×10^{-4}	1.99×10^{-6}	2.74×10^{-5}
0.2	4.57×10^{-3}	1.99×10^{-6}	5.58×10^{-4}
0.3	4.57×10^{-3}	5.58×10^{-4}	2.74×10^{-5}

Table 3.17: Fisher's exact test p-values

Chapter 4

Conclusions and Perspectives

4.1 Mass Spectrometry data alignment

The aim of the work was to provide a method for alignment of mass spectrometry data: data alignment from Alzheimer's patients allowed the creation of larger starting datasets, useful, eg, for data mining or other type of analysis.

We started with a theoretical frame and then materially implement the solutions in a machine learning environment (`RapidMiner`, see §2.1.5): this allowed us to compare our method against two competitive methods (§2.1.4) using a neutral context (the same machine learning process, with input data obtained according to the different methods).

As reported in the graphs listed in section 3.1 and 6.1, our method showed better performance in almost all tests, often with peaks of 20-25 %, and for all the different performance indexes considered (AUC, Precision and Recall; §3.1). Results were largely positive both regarding the comparison (alignment of pairs of laboratories), and the generalization to more laboratories (in our case, extension to all three labs).
Overall, our method provided a sensible fusion criterion between MS signals of different laboratories while also providing a generalization by Maximum Weight Bipartite Matching and maximization of shared information.

Our conclusions should be further strengthened by additional experimental data.

4.2 Mass Spectrometry data analysis

4.2.1 Divergence Analysis

The ultimate purpose of our divergence analysis was to compare datasets of cases and controls, to identify interesting regions in the spectra. The changes in these regions indicated a change in the peptidome, a mark showing that something changed upstream (proteome, transcriptome, genome).

The system had allowed us to identify the mass ranges of interest on which direct the work of mass spectrometrists. In analogy with wet labs, the comparison between observed and random graphs es-

tablished a kind of "blank" able to highlight the areas in which the method worked well: it was no coincidence that the areas highlighted by the comparison between the case and control systems were totally covered by those highlighted in the comparisons random versus cases or controls. The range of signals extracted in controls vs. cases comparison were (Table 3.9):

a) From 1940-3743 for controls;

b) From 5924-8858 for cases.

How to interpret the interval a), which indicated a controls mass range with unexpected behaviour? In "inverse" analogy with "random vs. case" and "random vs. controls" tests, we could consider this range of signals as untrustworthy for any signal case-related. Fortunately, a) and b) do not overlap, so with these data, the problem does not exist.

This first approach showed some defects, such as:

- use of the Kullback-Leibler divergence, a measure which is not metric (not define distances), is not symmetric, and is always positive [132]: in short, is confusing when compared to measures more understandable as the Pearson correlation. Its meaning would also be more difficult to interpret if we should develop a bioinformatics tool for a wider audience;

- the system is entirely based on the use of Random Graphs that simulate the noise biological and experimental measured and the robustness of the system. Is yet to be proved that the

robustness of a system can be verified using this method and, generally, a system completely designed with random objects, also not-so-random ones (see URnM above), does not seems so compelling.

4.2.2 Analysis of Correlation Structures

The analysis of correlation structures was based on the construction and the comparison between the signal correlation structures of control, ccRCC and not-ccRCC data. This allowed us to extract a region of interest (a range of signals) for each category of possible tests (see section 2.2.3). First of all, we must consider the greater clarity of the final result: for each test, each one very significant and powerful, we isolated a single contiguous range of signals in a region which showed different behavior in the two in different clinical states considered (for instance, controls and ccRCC).

A very interesting aspect is the ability of the method to operate not only the comparisons against controls, but also on histological sub-divisions, ccRCC and not-ccRCC, which in fact singles out different regions.

The decision to divide the spectrum, and therefore the graph, into regions created the conditions for calculating a distribution of values and for the elimination of Random Graphs. We understand, however, that the choice of dividing the spectrum in different regions was in itself arbitrary, because it established strict boundaries within a population of peptides ordered by weight (more precisely, for m/z):

this could be counterproductive, because we wanted to identify proteins linked by functional relationships, usually uncorrelated to mass. However, we think that the fact that studying a number of different sizes for the regions can overcome at least in part this problem. Finally, it is certainly necessary to identify the list of peptides in the ranges indicated, a list which we hope will be provided in the future by our mass spectroscopists.

The weak point of this method was the lack of a criterion for the evaluation of the results, which we introduced in the following methods: we could not verify in any way the goodness of the choices made by the algorithm. The Fisher's exact test allowed us to overcome this limitation.

4.2.3 Characterization of Distinguishing Regions

The analysis of correlation structures did not provide a tool for the evaluation of test results: the selected regions were not described by any index like, for instance, a p-value, widely used in other contexts (see, for instance, [133–135]). More importantly, the hypothesis tests presented in the previous methods were entirely designed and analyzed using always the entire datasets, which can induce statistical bias [136]. We decided to verify the presence of statistical bias by checking if there was an association between the properties of each region (discriminating/not discriminating) and the result of hypothesis testing, using Fisher's exact test. Fisher's exact test is probably one the best statistical tool available for small samples [125].

The results of the Characterization of Distinguishing Regions were substantially similar to that seen in the previous paragraph, regarding the analysis of correlation structures: however, we tried to identify two distinguishing regions instead of one. As already mentioned, we also examined each type of test with Fisher's exact test, always finding a significant association ($\alpha = 5\%$) between decision and region's property. The regions were in fact different from those identified by the previous method, with the exception of the test CVNR (see table 3.11).

4.2.4 Conclusions in MS data Analysis

This method represents the synthesis of everything we did before: we decided then to enclose in this section not only the conclusions concerning the "Robust Conclusions in MS Analysis" approach, but also more general conclusion about the analysis of MS data.

The robustness of a biological system is mainly defined as a property of a biological function [102,103]. For this reason robustness here relates to the determination of the effect of certain perturbations on the expression levels (i.e., spectra signals) of protein dependencies. Specifically, we referred to robustness as the persistence of our data model behavior (i.e., template behavior) against perturbations, as reflected in the deviations from proteomic signal dependencies. In the broadest sense of the word, robustness studies need to determine

how a process copes with uncertainties: data values models and parameters used in methods could be ill-determined, so the role of the modeler is to provide information about the validity of the proposed solutions for different sets of acceptable values for the reference *model* and the reference *method* [106,107]. Based on these ideas, we focused on the following three key points:

- *Reference model for the observed data* - Many conditions are best described by *relational* models in which instances of multiple types are related to each other in complex ways. Graphs provide a canonical representation for such relational data and their employment to reassess traditional data seems to be promising in order to better understand, summarize and visualize relationships amongst very large number of observations [137–140]. The rich literature on social network analysis gives probably the main tools for working with this aim (for example, [141, 142]).

  In mass spectrometry analysis, when it comes to analyzing peaks with different intensity in the MS spectra, comparisons are generally performed between proteins (peptides) profiles of different groups - or between statistics summarizing the peak property of a group [143]. Actually, different signals in the m/z spectra can be related to each other, and this property in turn may change from group to group. For this reason, following the idea to introduce relational information, we represented each group of subjects (controls, ccRCC, not-ccRCC) through graphs providing our reference model for the observed samples.

In this representation vertices are m/z ratios and edges express dependencies (i.e., mutual information) between signal intensities with specific m/z values.

- *Reference method to provide decisions* - The theoretical framework was employed mainly to define a reference methods for our analysis, i.e., a standard test of hypotheses approach over graph properties. Using this approach we obtained "differentially expressed" spectra regions between case and control groups of subjects, even if the signal identity was not yet ensured. As a matter of fact, *one of the major advantages of this strategy is that no pre-knowledge of the identity of signals selected for the pattern is needed* to allow their use as biomarkers [144]. It should be noted that the identification of peptide or protein signals in a profile is not straightforward: such efforts are tedious because of the requirement of specific separation or enrichment strategies. In addition, a high MS/MS data quality is needed for identification of endogenous species, i.e. large coverage of fragment ions. For these reasons, it is useful to first determine the diagnostic power of candidate markers before performing identification studies and further investigations into their biological role in disease mechanisms.

- *Reference model of variability* - We formulated through random graphs the *reference model of variability*, that can be helpful for a variety of purposes in the statistical practice: in our case, we derived a simple statistical property (property 1, see section 2.2.5), supporting the modeler to draw reliable conclusions about noised data. The modeler first should employ the con-

sidered variability model - i.e., the (s,t)-preserving Random Graph (defined in section 2.2.5), with a set of acceptable parameters s,t. Then he should check whether for these parameters a perturbation probability satisfying property 1 give rise to test decisions, which still maintains the conclusions previously obtained (i.e., before the perturbation mechanism was applied). In other words, due to property 1, a set of acceptable changes (e.g., at most deleted/added edges) give rise to the associated probabilities for the simulation of the effect of an uncertainty over the represented data. Throughout this process, we obtained robust conclusions for all the class of tests applied in this study, by considering as acceptable those template modifications for which, on average, at most 10% of the possible edge have been modified (i.e. added or deleted) from the original (observed) representation.

Our case-study concerned robust conclusions for differentially expressed mass spectrometry regions. Spectra regions are sequences of m/z values which provide information about the mass of biological molecules. Inside these regions we verified differentially expressed properties (i.e., signal dependencies cohesion) between control and case (ccRCC, not-ccRCC) groups. We point out that we defined as graph regions those subgraphs which reflect the signal dependencies occurring over mass spectrometry regions. On the other hand, we called neighborhoods these subgraphs which characterize statistical units of analysis. We gave these definitions simply for the different use we made of these structures.

Many questions still need to be addressed in future analysis. First of all, defining on average acceptable number of modifications (section 2.2.5) should also require info about the related proteins (peptides) involved in the added and/or deleted dependencies. Here we employed the reference method mainly to obtain spectra distinguishing regions for their biological evidence, but a better understanding of the molecular interactions would give a greater biological significance to the edges addition or removal. Also,the number of acceptable modifications is clearly arbitrary and strictly dependent on the modeler's opinion: this lead to having to many alternative results in the robustness analysis. For instance, for some large value assignment to the parameter t (see section 2.2.5), eg probability = 50% of additions and deletions, can even results in puzzling conclusions. De facto, we give the modeler the chance to choose both a perturbation inducing association (e.g., p = 0.1) and a perturbation inducing independence (e.g., for p = 0.2) in the Fisher's exact test.

Further analysis should concern the use of some parameters which we have arbitrary defined as constant values (for examples, the size $k_{regions}$ of regions, or the "radius" k of a neighborhood). The selection of these arbitrary values can have different effects on the model accuracy. For instance, in order to adapt the method to different data sets, it should be important to allow one to choose them in a more principled way.
Probably the most critical parameter is the hard threshold δ, useful to represent dependencies (i.e., edges) in the template graphs. It would be possible to overcome this hard threshold using weighted graphs, but the calculation of the weight needs more biological in-

formation: first, even partial recognition of the peptides associated with the signals, and then the creation of weighted networks using, for example, interactions reported in literature.

Finally from a biological and clinical prospective, since the proteins that show changes in expression level as a consequence of a disease have great potential as new biomarkers (in diagnosis, prognosis and as potential therapeutic targets), we need to conclusively fix both the classification predictive power of the RCC distinguishing regions and their biological identity aimed to explore the structure and function of these potential biomarkers.

References

[1] M. Pietrowska and P. Widlak. MALDI-MS-Based Profiling of Serum Proteome: Detection of Changes Related to Progression of Cancer and Response to Anticancer Treatment. *Int J Proteomics*, 2012:926427, 2012.

[2] R. J. Elin. Instrumentation in clinical chemistry. *Science*, 210(4467):286–289, Oct 1980.

[3] D. H. Chace. Mass spectrometry in the clinical laboratory. *Chem. Rev.*, 101(2):445–477, Feb 2001.

[4] C. Boccardi, S. Rocchiccioli, A. Cecchettini, A. Mercatanti, and L. Citti. An automated plasma protein fractionation design: high-throughput perspectives for proteomic analysis. *BMC Res Notes*, 5(1):612, 2012.

[5] K. A. Landers, M. J. Burger, M. A. Tebay, D. M. Purdie, B. Scells, H. Samaratunga, M. F. Lavin, and R. A. Gardiner. Use of multiple biomarkers for a molecular diagnosis of prostate cancer. *Int. J. Cancer*, 114(6):950–956, May 2005.

[6] D. M. Good, V. Thongboonkerd, J. Novak, J. L. Bascands, J. P. Schanstra, J. J. Coon, A. Dominiczak, and H. Mischak. Body fluid proteomics for biomarker discovery: lessons from the past hold the key to success in the future. *J. Proteome Res.*, 6(12):4549–4555, Dec 2007.

[7] J. R. Langabeer and Y. A. Ozcan. The economics of cancer care: longitudinal changes in provider efficiency. *Health Care Manag Sci*, 12(2):192–200, Jun 2009.

[8] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller. Processing and classification of protein mass spectra. *Mass Spectrom Rev*, 25(3):409–449, 2006.

[9] J. N. Andersen, S. Sathyanarayanan, A. Di Bacco, A. Chi, T. Zhang, A. H. Chen, B. Dolinski, M. Kraus, B. Roberts, W. Arthur, R. A. Klinghoffer, D. Gargano, L. Li, I. Feldman, B. Lynch, J. Rush, R. C. Hendrickson, P. Blume-Jensen, and C. P. Paweletz. Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors. *Sci Transl Med*, 2(43):43ra55, Aug 2010.

[10] AJ Atkinson, WA Colburn, VG DeGruttola, DL Demets, and GJ Downing. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.*, 69:89–95, 2001.

[11] J. Villanueva, D. R. Shaffer, J. Philip, C. A. Chaparro, H. Erdjument-Bromage, A. B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E. C. Holland,

C. Cordon-Cardo, H. I. Scher, and P. Tempst. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.*, 116(1):271–284, Jan 2006.

[12] J. Villanueva, D. R. Shaffer, J. Philip, C. A. Chaparro, H. Erdjument-Bromage, A. B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E. C. Holland, C. Cordon-Cardo, H. I. Scher, and P. Tempst. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.*, 116(1):271–284, Jan 2006.

[13] H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, Nov 2002.

[14] H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–1664, Mar 2002.

[15] M. Latterich, M. Abramovitz, and B. Leyland-Jones. Proteomics: new technologies and clinical applications. *Eur. J. Cancer*, 44(18):2737–2741, Dec 2008.

[16] L. Alberghina, D. Gaglio, C. Gelfi, R.M. Moresco, G. Mauri, P. Bertolazzi, C. Messa, M.C. Gilardi, F. Chiaradonna, and M. Vanoni. Cancer cell growth and survival as a system-level property sustained by enhanced glycolysis and mitochondrial metabolic remodeling. *Front. Physio.*, 3(362), 2012.

[17] F.W. McLafferty. Mass spectrometry across the sciences. *Proc. Natl. Acad. Sci.*, 105:18088–89, 2008.

[18] A. Gruhler, J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann, and O. N. Jensen. Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell Proteomics*, 4(3):310–327, Mar 2005.

[19] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, Mar 2003.

[20] D. Fenyo and R. C. Beavis. Informatics development: challenges and solutions for MALDI mass spectrometry. *Mass Spectrom Rev*, 27(1):1–19, 2008.

[21] A. Cruz-Marcelo, R. Guerra, M. Vannucci, Y. Li, C. C. Lau, and T. K. Man. Comparison of algorithms for preprocessing of SELDI-TOF mass spectrometry data. *Bioinformatics*, 24(19):2129–2136, Oct 2008.

[22] R Staden. Sequence data handling by computer. *Nucleic acids research*, 4(11):4037–51, 1977.

[23] Bilofsky et al. The GenBank genetic sequence databank. *Nucleic acids research*, 14(1):1–4, 1986.

[24] M. Hilario and A. Kalousis. Approaches to dimensionality reduction in proteomic biomarker studies. *Brief. Bioinformatics*, 9(2):102–118, Mar 2008.

[25] Schewe KD and Thalheim B. Conceptual modelling of web information systems. *Data Knowl Eng*, 54, 2005.

[26] F Hillenkamp, M Karas, RC Beavis, and BT Chait. Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. *Anal. Chem.*, 63, 1991.

[27] W.E. Stephens. A pulsed mass spectrometer with time dispersion. *Bull Am Phys Soc*, 21(2):22, 1946.

[28] N. Mirsaleh-Kohan, W. D. Robertson, and R. N. Compton. Electron ionization time-of-flight mass spectrometry: historical review and current applications. *Mass Spectrom Rev*, 27(3):237–285, 2008.

[29] R. J. Cotter. The New Time-of-Flight Mass Spectrometry. *Anal. Chem.*, 71(13):445A–51A, Jul 1999.

[30] N. Dephoure, C. Zhou, J. Villen, S. A. Beausoleil, C. E. Bakalarski, S. J. Elledge, and S. P. Gygi. A quantitative atlas of mitotic phosphorylation. *Proc. Natl. Acad. Sci. U.S.A.*, 105(31):10762–10767, Aug 2008.

[31] P. A. Grimsrud, D. L. Swaney, C. D. Wenger, N. A. Beauchene, and J. J. Coon. Phosphoproteomics for the masses. *ACS Chem. Biol.*, 5(1):105–119, Jan 2010.

[32] A. J. Tackett, J. A. DeGrasse, M. D. Sekedat, M. Oeffinger, M. P. Rout, and B. T. Chait. I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions. *J. Proteome Res.*, 4(5):1752–1756, 2005.

[33] O. Rinner, L. N. Mueller, M. Hubalek, M. Muller, M. Gstaiger, and R. Aebersold. An integrated mass spectrometric and com-

putational framework for the analysis of protein interaction networks. *Nat. Biotechnol.*, 25(3):345–352, Mar 2007.

[34] M. Sugaya, R. Saito, Y. Matsumura, K. Harada, and A. Katoh. Facile detection of specific RNA-polypeptide interactions by MALDI-TOF mass spectrometry. *J. Pept. Sci.*, 14(8):978–983, Aug 2008.

[35] P. Ross, L. Hall, I. Smirnov, and L. Haff. High level multiplex genotyping by MALDI-TOF mass spectrometry. *Nat. Biotechnol.*, 16(13):1347–1351, Dec 1998.

[36] W. Pusch and M. Kostrzewa. Application of MALDI-TOF mass spectrometry in screening and diagnostic research. *Curr. Pharm. Des.*, 11(20):2577–2591, 2005.

[37] R. Zenobi and R. Knochenmuss. Ion formation in MALDI mass spectrometry. *Mass Spectrometry Reviews*, 17(5):337–366, 1998.

[38] A. G. Marshall and C. L. Hendrickson. High-resolution mass spectrometers. *Annu Rev Anal Chem (Palo Alto Calif)*, 1:579–599, 2008.

[39] S.G. Alikhanov. A new impulse technique for ion mass measurement. *Sov. Phys. JETP 4*, 452, 1957.

[40] B.A. Mamyrin. Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*, 206(3):251 – 266, 2001.

[41] K. Dreisewerd. The desorption process in MALDI. *Chem. Rev.*, 103(2):395–426, Feb 2003.

[42] A Vertes, G Irinyi, and R Gijbels. Hydrodynamic model of matrix-assisted laser desorption mass spectrometry. *AnalChem*, 65:2389, 1993.

[43] LV Zhigilei, PBS Kodali, and BJ Garrison. Molecular dynamics model for laser ablation and desorption of organic solids. *J Phys Chem B*, 101:2028, 1997.

[44] LV Zhigilei and BJ Garrison. Velocity distributions of analyte molecules in matrix-assisted laser desorption from computer simulations. *Rapid Comm Mass Spectrom*, 12:1273, 1998.

[45] B. T. Chait. Mass Spectrometry in the Postgenomic Era. *J. Proteome Res.*, 4(5):1752–1756, 2005.

[46] D. W. Mahoney, T. M. Therneau, C. J. Heppelmann, L. Higgins, L. M. Benson, R. M. Zenka, P. Jagtap, G. L. Nelsestuen, H. R. Bergen, and A. L. Oberg. Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from iTRAQ-labeled peptides. *J. Proteome Res.*, 10(9):4325–4333, Sep 2011.

[47] Daniel C. Liebler. *Introduction to Proteomics: Tools for the New Biology*. Humana Press, 2002.

[48] M. Blueggel, D. Chamrad, and H. E. Meyer. Bioinformatics in proteomics. *Curr Pharm Biotechnol*, 5(1):79–88, Feb 2004.

[49] R. Kratzer, C. Eckerskorn, M. Karas, and F. Lottspeich. Suppression effects in enzymatic peptide ladder sequencing using ultraviolet - matrix assisted laser desorption/ionization - mass spectrometry. *Electrophoresis*, 19(11):1910–1919, Aug 1998.

[50] J. Pierson, J. L. Norris, H. R. Aerni, P. Svenningsson, R. M. Caprioli, and P. E. Andren. Molecular profiling of experimental Parkinson's disease: direct analysis of peptides and proteins on brain tissue sections by MALDI mass spectrometry. *J. Proteome Res.*, 3(2):289–295, 2004.

[51] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, Oct 2007.

[52] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, Jul 2005.

[53] L. E. Hebert, P. A. Scherr, J. L. Bienias, D. A. Bennett, and D. A. Evans. Alzheimer disease in the US population: prevalence estimates using the 2000 census. *Arch. Neurol.*, 60(8):1119–1122, Aug 2003.

[54] R. J. O'Brien and P. C. Wong. Amyloid precursor protein processing and Alzheimer's disease. *Annu. Rev. Neurosci.*, 34:185–204, 2011.

[55] G. Thinakaran and E. H. Koo. Amyloid precursor protein trafficking, processing, and function. *J. Biol. Chem.*, 283(44):29615–29619, Oct 2008.

[56] B. A. Yankner, L. R. Dawes, S. Fisher, L. Villa-Komaroff, M. L. Oster-Granite, and R. L. Neve. Neurotoxicity of a fragment of the amyloid precursor associated with Alzheimer's disease. *Science*, 245(4916):417–420, Jul 1989.

[57] G. Krabbe, A. Halle, V. Matyash, J. L. Rinnenthal, G. D. Eom, U. Bernhardt, K. R. Miller, S. Prokop, H. Kettenmann, and F. L. Heppner. Functional impairment of microglia coincides with Beta-amyloid deposition in mice with Alzheimer-like pathology. *PLoS ONE*, 8(4):e60921, 2013.

[58] A. Bobba, G. Amadoro, V. Petragallo, P. Calissano, and A. Atlante. Dissecting the molecular mechanism by which NH2htau and Abeta1-42 peptides impair mitochondrial ANT-1 in Alzheimer disease. *Biochim. Biophys. Acta*, Apr 2013.

[59] S. M. Cologna, X. S. Jiang, P. S. Backlund, C. V. Cluzeau, M. K. Dail, N. M. Yanjanin, S. Siebel, C. L. Toth, H. S. Jun, C. A. Wassif, A. L. Yergey, and F. D. Porter. Quantitative proteomic analysis of Niemann-Pick disease, type C1 cerebellum identifies protein biomarkers and provides pathological insight. *PLoS ONE*, 7(10):e47845, 2012.

[60] T. J. Craddock, J. A. Tuszynski, D. Chopra, N. Casey, L. E. Goldstein, S. R. Hameroff, and R. E. Tanzi. The zinc dyshomeostasis hypothesis of Alzheimer's disease. *PLoS ONE*, 7(3):e33552, 2012.

[61] C. Czech, P. Berndt, K. Busch, O. Schmitz, J. Wiemer, V. Most, H. Hampel, J. Kastler, and H. Senn. Metabolite pro-

filing of Alzheimer's disease cerebrospinal fluid. *PLoS ONE*, 7(2):e31501, 2012.

[62] A. S. Arefin, L. Mathieson, D. Johnstone, R. Berretta, and P. Moscato. Unveiling clusters of RNA transcript pairs associated with markers of Alzheimer's disease progression. *PLoS ONE*, 7(9):e45535, 2012.

[63] E. M. Castano, C. L. Maarouf, T. Wu, M. C. Leal, C. M. Whiteside, L. F. Lue, T. A. Kokjohn, M. N. Sabbagh, T. G. Beach, and A. E. Roher. Alzheimer disease periventricular white matter lesions exhibit specific proteomic profile alterations. *Neurochem. Int.*, 62(2):145–156, Jan 2013.

[64] D. J. Pappin, P. Hojrup, and A. J. Bleasby. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.*, 3(6):327–332, Jun 1993.

[65] W. J. Henzel, T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley, and C. Watanabe. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U.S.A.*, 90(11):5011–5015, Jun 1993.

[66] M. Mann, P. Hojrup, and P. Roepstorff. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.*, 22(6):338–345, Jun 1993.

[67] P. James, M. Quadroni, E. Carafoli, and G. Gonnet. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.*, 195(1):58–64, Aug 1993.

[68] Olga Brazhnik and John Jones. Anatomy of data integration. *Journal of Biomedical Informatics*, 40:252–269, 2007.

[69] Massimiliano Borsani. HuNTED: una banca dati di espressione di tessuti normali e tumorali (ENG: HuNTED: a gene expression database of normal and tumor tissues). Master's thesis, University of Milano, Feb 2009.

[70] Kari Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[71] MH Zweig and G Campbell. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin. Chem.*, 39(4):561–577, Apr 1993.

[72] MS Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95, 2000.

[73] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.

[74] Hilario M, Kalousis A, Müller M, and Pellegrini C. Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, 3(9):1716 – 1719, 2003.

[75] R. H. Christenson and S. H. Duh. Methodological and analytic considerations for blood biomarkers. *Prog Cardiovasc Dis*, 55(1):25–33, 2012.

[76] M. Dettling and P. Buhlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–1069, Jun 2003.

[77] R. Diaz-Uriarte and S. Alvarez de Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

[78] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun. Cancer statistics, 2007. *CA Cancer J Clin*, 57(1):43–66, 2007.

[79] S. George and R. M. Bukowski. Biomarkers in clear cell renal cell carcinoma. *Expert Rev Anticancer Ther*, 7(12):1737–1747, Dec 2007.

[80] C. Bosetti, C. Bianchi, E. Negri, and C. La Vecchia. Estimates of the incidence and prevalence of renal cell carcinoma in Italy in 2002 and projections for the years 2007 and 2012. *Tumori*, 95(2):142–145, 2009.

[81] S. Skates and O. Iliopoulos. Molecular markers for early detection of renal carcinoma: investigative approach. *Clin. Cancer Res.*, 10(18 Pt 2):6296S–301S, Sep 2004.

[82] Robert J. Motzer and Paul Russo. Systemic therapy for renal cell carcinoma. *The Journal of Urology*, 163(2):408 – 417, 2000.

[83] Ting Shi, Fan Dong, Louis S. Liou, Zhong-Hui Duan, Andrew C. Novick, and Joseph A. DiDonato. Differential protein profiling in renal-cell carcinoma. *Molecular Carcinogenesis*, 40(1):47–61, 2004.

[84] N. Bosso, C. Chinello, S. C. Picozzi, E. Gianazza, V. Mainini, C. Galbusera, F. Raimondo, R. Perego, S. Casellato, F. Rocco, S. Ferrero, S. Bosari, P. Mocarelli, M. G. Kienle, and F. Magni. Human urine biomarkers of renal cell carcinoma evaluated by ClinProt. *Proteomics Clin Appl*, 2(7-8):1036–1046, Jul 2008.

[85] C. Chinello, E. Gianazza, I. Zoppis, V. Mainini, C. Galbusera, S. Picozzi, F. Rocco, G. Galasso, S. Bosari, S. Ferrero, R. Perego, F. Raimondo, C. Bianchi, M. Pitto, S. Signorini, P. Brambilla, P. Mocarelli, M. Galli Kienle, and F. Magni. Serum biomarkers of renal cell carcinoma assessed using a protein profiling approach based on ClinProt technique. *Urology*, 75(4):842–847, Apr 2010.

[86] A. He, J. Bai, C. Huang, J. Yang, W. Zhang, J. Wang, Y. Yang, P. Zhang, and F. Zhou. Detection of serum tumor markers in multiple myeloma using the CLINPROT system. *Int. J. Hematol.*, 95(6):668–674, Jun 2012.

[87] B. J. Drucker. Renal cell carcinoma: current status and future prospects. *Cancer Treat. Rev.*, 31(7):536–545, Nov 2005.

[88] V. E. Reuter. The pathology of renal epithelial neoplasms. *Semin. Oncol.*, 33(5):534–543, Oct 2006.

[89] R. Vikram, C. S. Ng, P. Tamboli, N. M. Tannir, E. Jonasch, S. F. Matin, C. G. Wood, and C. M. Sandler. Papillary renal cell carcinoma: radiologic-pathologic correlation and spectrum of disease. *Radiographics*, 29(3):741–754, 2009.

[90] Brian I Rini, Steven C Campbell, and Bernard Escudier. Renal cell carcinoma. *The Lancet*, 373(9669):1119 – 1132, 2009.

[91] J. S. Lam, J. T. Leppert, R. A. Figlin, and A. S. Belldegrun. Role of molecular markers in the diagnosis and therapy of renal cell carcinoma. *Urology*, 66(5 Suppl):1–9, Nov 2005.

[92] Eric D. Kolaczyk. *Statistical Analysis of Network Data*. Springer Series in Statistics. Springer, 2009.

[93] Reinhard Diestel. *Graph Theory*, chapter 1. Graduate Texts in Mathematics. Springer, IV edition, Feb 2010.

[94] B. Bollobás. *Random Graphs*. Academic Press, London, 1985.

[95] David J. Marchette. *Random Graphs for Statistical Pattern Recognition*. Wiley-Interscience, 2004.

[96] P. Erdos and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–7, 1957.

[97] E. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30:1141–4, 1959.

[98] G. Gigerenzer and D. J. Murray. *Cognition as intuitive statistics*. Hillsdale, 1987.

[99] Aris Spanos. *Probability Theory and Statistical Inference.* Econometric Modeling with Observational Data. Cambridge University Press, 1999.

[100] J. Neyman and E.S. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 1928.

[101] J. Neyman and E.S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

[102] H. Kitano. Biological robustness. *Nat. Rev. Genet.*, 5(11):826–837, Nov 2004.

[103] H. Kitano. Towards a theory of biological robustness. *Mol. Syst. Biol.*, 3:137, 2007.

[104] Mineo Morohashi, Amanda E. Winn, Mark T. Borisuk, Hamid Bolouri, John Doyle, and Hiroaki Kitano. Robustness as a measure of plausibility in models of biochemical networks. *Theor. Biol.*, 216(1):19–30, 2002.

[105] F.R. Hampel. *Robust statistics: a brief introduction and overview.* Seminar für Statistik. Eidgenössische Technische Hochschule, 2001.

[106] Philippe Vincke. Robust solutions and methods in decision aid. *Journal of Multi-Criteria Decision Analysis*, 8:181–187, 1999.

[107] P. Perny, O. Spanjaard, and L.X. Storme. A decision-theoretic approach to robust optimization in multivalued graphs. *Annals of Operations Research*, 147(1):317–341, 2006.

[108] Panos Kouvelis and Gang Yu. *Robust Discrete Optimization and Its Applications*. Nonconvex Optimization and Its Applications (closed). Springer, 1996.

[109] J. Martín, C.J. Pérez, and P. Müller. Bayesian robustness for decision making problems: Applications in medical contexts. *Int. J. Approx. Reasoning*, 50:315–323, 2009.

[110] O.J. Berger. Robust bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328, 1990.

[111] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Chapman and Hall/CRC, second edition edition, 2003.

[112] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[113] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh. *Feature extraction: foundations and applications*, volume 207. Springer, 2006.

[114] D.B. West. *Introduction to Graph Theory*. Prentice Hall, 1999.

[115] JA McHugh. *Algorithmic Graph Theory*. Prentice Hall, 1990.

[116] E. Gianazza, C. Chinello, V. Mainini, M. Cazzaniga, V. Squeo, G. Albo, S. Signorini, S. S. Di Pierro, S. Ferrero, S. Nicolardi, Y. E. van der Burgt, A. M. Deelder, and F. Magni. Alterations of the serum peptidome in renal cell carcinoma discriminating benign and malignant kidney tumors. *J Proteomics*, 76 Spec No.:125–140, Dec 2012.

[117] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[118] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo Method*. Wiley Series in Probability and Statistics. Wiley, 2011.

[119] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, Feb. 1988.

[120] Student (William Sealy Gosset). The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908.

[121] T.H. Wonnacott and R.J. Wonnacott. *Introductory Statistics*. Wiley, 2009.

[122] Jeremy Adler and Ingela Parmryd. Quantifying colocalization by correlation: The pearson correlation coefficient is superior to the mander's overlap coefficient. *Cytometry Part A*, 77A(8):733–742, 2010.

[123] B. Silverman. *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. Chapman and Hall, 1986.

[124] R.A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, January 1922.

[125] Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153, February 1992.

[126] J. Rosenhead. *Robustness analysis: Keeping your options open.*, pages 181–207. In J. Rosenhead and J. Mingers (editors) - Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict. John Wiley & Sons, 2001.

[127] J. Rosenhead. *Robustness to the first degree*, pages 209–223. In J. Rosenhead and J. Mingers (editors) - Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict. John Wiley & Sons, 2001.

[128] H.Y. Wong and J. Rosenhead. A rigorous definition of robustness analysis. *Journal of the Operational Research Society*, 51:176–182, 2000.

[129] NF Samatova, MC Schmidt, W Hendrix, P Breimyer, K Thomas, and BH Park. Coupling graph perturbation theory with scalable parallel algorithms for large-scale enumeration of maximal cliques in biological graphs. *Journal of Physics: Conference Series*, 125(1):012053, 2008.

[130] Sami Hanhijärvi, Gemma C. Garriga, and Kai Puolamäki. Randomization techniques for graphs. In *Proceedings of the 9th SIAM International Conference on Data Mining (SDM '09)*, pages 780–791, 2009.

[131] Evangelos Triantaphyllou and Alfonso Sanchez. A sensitivity analysis approach for some deterministic multi-criteria decision-making methods. *Decision Sciences*, 28(1):151–194, 1997.

[132] Kenneth P. Burnham and David R. Anderson. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, 2nd edition, 2002.

[133] M. Kapushesky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.*, 38(Database issue):D690–698, Jan 2010.

[134] d. a. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37(1):1–13, Jan 2009.

[135] C. J. Willer, Y. Li, and G. R. Abecasis. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, Sep 2010.

[136] Jennifer Neville, Brian Gallagher, Tina Eliassi-Rad, and Tao Wang. Correcting evaluation bias of relational classifiers with network cross validation. *Knowledge and information systems*, 30(1):31–55, 2012.

[137] N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, L. D. Hurst, and M. Tyers. Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS biology*, 4(10):e317, 2006.

[138] Elvan Ceyhan, Carey E. Priebe, and David J. Marchette. A new family of random graphs for testing spatial segregation. *The Canadian Journal of Statistics*, 35(1):27–50, 2007.

[139] R. J. Williams J. A. Dunne and N. D. Martinez. Food-web structure and network theory: The role of connectance and size. *PNAS*, 99(20):12917–12922, 2002.

[140] Steven K. Thompson. Adaptive web sampling. *Biometrics*, 62(4):1224–1234, 2006.

[141] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[142] Stanley Wasserman and Joseph Galaskiewicz. *Advances in social network analysis: research in the social and behavioral sciences.* Sage Publications, 1994.

[143] J. Solassol, W. Jacot, L. Lhermitte, N. Boulle, T. Maudelonde, and A. Mang. Clinical proteomics and mass spectrometry profiling for cancer detection. *Expert Rev. Proteomics*, 3(3):311–320, 2006.

[144] A. Villar-Garea, M. Griese, and A. Imhof. Biomarker discovery from body fluids using mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 849:105–114, 2007.

Acknowledgements

Thanks to:

Dr Italo Zoppis, a nice person, and the mind behind this project;

Prof. Fulvio Magni (funded me), Prof. Giancarlo Mauri (involved me in this project), and Prof. Lilia Alberghina (enabled me to complete my PhD program);

my colleagues, Dr Matteo Chiara and Dr Federico Zambelli, to all members of BEACON lab, and in particular to my supervisor, Prof. Giulio Pavesi;

above all, thank to my family (Mario, Mariangela, Sara), my friends (Samer, Daniele, Andrea, Gianluca, Isabella, Marco, Agnese, Vanessa, Mirella, Katia, Alessandra, Andrea; Laura, Daniele, Davide, Paola, Valentina, Lapo, Jacopo, Ginger, and many more...); to Gian Marco, Elisa, Vittorio, Gianluca, and to my wife Rosangela, for their support and love.

Part II

Papers and Conference Posters

Chapter 5

Published papers

Information Optimization for Mass Spectra Data Alignment

Citation: Italo Zoppis, Erica Gianazza, Massimiliano Borsani, Clizia Chinello, Veronica Mainini, Carmen Galbusera, Carlo Ferrarese, Gloria Galimberti, Sandro Sorbi, Barbara Borroni, Fulvio Magni, Marco Antoniotti, Giancarlo Mauri. Mutual Information Optimization for Mass Spectra Data Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 3, pp. 934-939, May-June 2012. ©2012 IEEE. Reprinted, with permission from the authors.

Document Type: Article (source: Scopus)
ISSN: 15455963
DOI: 10.1109/TCBB.2011.80
PMID: 21519116

Short Papers

Mutual Information Optimization for Mass Spectra Data Alignment

I. Zoppis et al.

Abstract—"Signal" alignments play critical roles in many clinical setting. This is the case of mass spectrometry (MS) data, an important component of many types of proteomic analysis. A central problem occurs when one needs to integrate (MS) data produced by different sources, e.g., different equipment and/or laboratories. In these cases, some form of "data integration" or "data fusion" may be necessary in order to discard some source-specific aspects and improve the ability to perform a classification task such as inferring the "disease classes" of patients. The need for new high-performance data alignments methods is therefore particularly important in these contexts. In this paper, we propose an approach based both on an information theory perspective, generally used in a feature construction problem, and the application of a mathematical programming task (i.e., the weighted bipartite matching problem). We present the results of a competitive analysis of our method against other approaches. The analysis was conducted on data from plasma/ethylenediaminetetraacetic acid of "control" and Alzheimer patients collected from three different hospitals. The results point to a significant performance advantage of our method with respect to the competing ones tested.

Index Terms—Optimization, information theory, medicine, medical informatics, proteomics, data integration, graph algorithms.

1 Introduction

Alzheimer disease (AD) represents one of the most common neurodegenerative disorder in the elderly. It is characterized by progressive memory, language, and other cognitive function impairment, as well as by behavioral and social deterioration [1], [2]. A large number of studies are currently investigating the pathogenetic mechanisms involved in such a complex disease. Abeta 1-42 aggregation, tau hyperphosphorylation, inflammation, oxidative stress, and glutamate-induced excitotoxicity are now considered the main events which probably interact and lead to neuronal death and synaptic loss, ultimately resulting in dementia [3], [4], [5], [6]. Alzheimer is often discovered late, so it is urgent to define biomarkers for an early detection, for a differential diagnosis from other neurodegenerative diseases, and to monitor the course of the disease [7]. Recently, proteomics has become an emerging field in research on clinical diagnostics because of its power to detect and identify differentially expressed proteins/peptides during physiological and pathological processes [8]. Currently, the use of a single biomarker to realize diagnostic models is considered incomplete; consequently, studies are now growing about the discovering of multiple biomarkers which contain a higher level of discriminatory information [9], [10]. Protein profiling with mass spectrometry (MS) represents a promising tool for the biomarkers discovery and for an improved understanding of the disease biological mechanisms. One of the emerging MS-based screening methods allowing high-throughput analysis of peripheral fluids (easily accessible with noninvasive procedures) with a simple and automated process is the ClinProt technique. A successful discovery of a proteomic profile related to an altered state has been obtained in different human diseases with this methodology [11], [12], [13]. In particular, the ClinProt technique can be used to obtain the protein profile of the biological fluids utilizing magnetic beads with an active surface able to extract specific peptides and proteins, which are then analyzed by matrix laser desorption ionization time of flight (MALDI-TOF) MS. A protein/peptide profile is a graph function (Fig. 1) in which each peak (or signal) is bell-shaped with a height which identifies the intensity (related to the "abundance") at a specific mass-to-charge ratio of a biomolecule (protein/peptide) in the original sample. In this paper, we refer to this pair of variables (i.e., intensity and mass value) as *features*.[1]

Observed data from MALDI-TOF are generally organized as values stored in tables like those in Fig. 2. These values may be affected by errors introduced during different experiment phases (or even due to day-to-day instrument variations), causing noise, peak broadening, contaminants, etc. Moreover, since the MALDI-TOF mass spectrometer resolution working in linear mode has a mass accuracy in the range of about $+/-8$ Daltons, the measured m/z of the same entity (protein/peptide) can be slightly different in each spectrum. To allow an easy and effective comparison of different spectra, alignment methods find a common set of peak locations (i.e., m/z values), among sets of spectra, in such a way that all spectra have common m/z values for the same biological entities (see, for instance, [14]). In other words, an alignment finds which features among different spectra share common qualities (identify the same protein/peptide molecule). The search for a suitable solution to this clinical alignment problem can also be motivated through two interrelated lines of thought. The first is noise reduction. Discarding the source-specific aspects will eliminate the noise. The second is, in general, more abstract. Here, different measurement sources can even convey different kinds of information. In our case, the mass-to-charge ratio values taken in different laboratories may refer to the same peptide; and different peptides may even be considered in each lab measurement. What is in common in the sources is what we are really interested in.

The quality of sharing common attributes (commonalities) in data sources has been studied by methods that search for statistical dependencies between them. The earliest was the classical linear canonical correlation analysis [15] which has been extended to nonlinear variants (for example, in [16]) and more general techniques that maximize mutual information (MI) [17]. Moreover, MI [18], "the measure" adopted in our study and described later on, has, of course, already been used in the biomedical domain, e.g., Hilario et al. [19] describe the use of MI for biomarkers prediction. By applying this measure, it is generally possible to enhance the inference on the disease class[2] of patients and even rank the useful

- *I. Zoppis, M. Antoniotti, and G. Mauri are with the Department of Informatics, Systems and Communication, University of Milano-Bicocca, Edificio U14, Viale Sarca 336, Milano 20126, Italy. E-mail: italo.zoppis@gmail.com, marco.antoniotti@unimib.it, mauri@disco.unimib.it.*
- *E. Gianazza, M. Borsani, C. Chinello, V. Mainini, and F. Magni are with the Department of Experimental Medicine, University of Milano-Bicocca, Edificio U8, via Cadore 48, Monza 20900, Italy. E-mail: e.gianazza@campus.unimib.it, massimiliano.borsani@unimi.it, {clizia.chinello, veronica.mainini, fulvio.magni}@unimib.it.*
- *C. Galbusera is with the Department of Neuroscience and Biomedical Technology, University of Milano-Bicocca, Edificio U8, via Cadore 48, Monza 20900, Italy. E-mail: carmen.galbusera@unimib.it.*
- *C. Ferrarese and G. Galimberti are with the Department of Neurology, San Gerardo Hospital, via cadore 48, Monza 20052, Italy. E-mail: {carlo.ferrarese, gloria.galimberti}@unimib.it.*
- *S. Sorbi is with the Department of Neurological and Psychiatric Sciences, University of Florence, Viale Morgagni, 85 Firenze, Florence 50131, Italy. E-mail: sandro.sorbi@unifi.it.*
- *B. Borroni is with the Department of Neurology, Center for Aging Brain and Dementia, University of Brescia, Piazza Spedali Civili 1, Brescia 25125, Italy. E-mail: borroni@med.unibs.it.*

Manuscript received 23 Feb. 2010; revised 18 Aug. 2010; accepted 21 Feb. 2011; published online 19 Apr. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2010-02-0061. Digital Object Identifier no. 10.1109/TCBB.2011.80.

1. With a slide abuse of terminology, we also call the mass-to-charge ratio value with the terms mass value.

2. With the terms "disease class," we refer to a *feature* reflecting the health state of patients; for instance, in this paper, this attribute can take two values: 1) reflecting control patients (patients with no apparent disease) and 2) reflecting AD patients. We also name this attribute "target class."

1545-5963/12/$31.00 © 2012 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

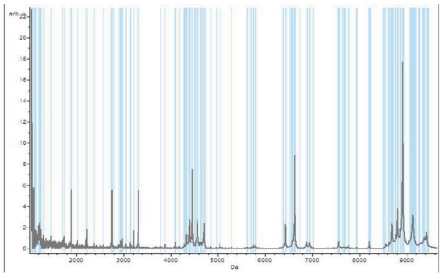


Fig. 1. A typical mass spectrum profile.

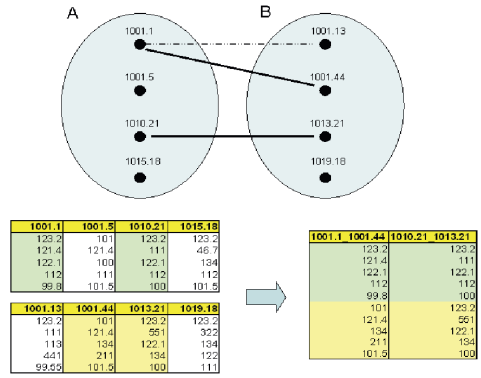
features (e.g, this is the case adopted for several biomarkers discovery problems—see [20]). In our case, following this approach, our option is to also include these aspects of each data source that are "mutually informative" for the integrated target attribute (i.e., disease class values of patients from different labs). In fact, the mass spectra alignment is based on establishing protein/peptide correspondences between data sets by concatenating these intensity values from different profiles, which are most informative for the respective disease classes (cf., Fig. 2). This is the facet we want to analyze in this study, i.e., a suitable way to align signals in order to integrate properly data provided from different labs.

The main source of inspiration for our proposal is the theoretical formalism [21] used for the feature construction and extraction methods (here denoted as FSCM). Broadly speaking FSCM consists of applying both a "relevance function" which, given a set of variables (features), evaluates the relevance of the set, and a construction mechanism to build new characteristic features. In this paper, FSCM gives us the possibility to induce a formal definition for the MS data alignment (denoted as MSDA problem). This definition expresses an MI maximization problem and also gives the advantage of casting MSDA in the general context of the stochastic optimization.

In a nutshell, by sampling the target class y and a feature x, one generally applies the MI $\mathcal{I}(x, y)$ to quantify the information these two objects share together. Therefore, in order to perform an optimization leading to some combination g of a suitable pair of features, e.g., $g(x, z)$, one can try to extract z from a set of features $\{z_1, z_2, \ldots z_m\}$, which improve $\mathcal{I}(g(x, z), y)$. In our case, the right combination $g(x, z)$ may contribute to the common peptide alignments measured in different labs. This is roughly the approach we describe more formally in Section 2. In the same section, we propose a way to approach computationally this task. Its estimation can be quite naturally obtained through the application of the *Maximum Weighted Bipartite Matching* problem (MWBM). Several applications for both MWBM and (in general) the *matching problems* have been described in literature (see, for instance, [22]). In our case by giving a solution for MWBM, we obtain an estimated solution for MSDA. In Section 2, we also recall some fundamentals both concerning MI (useful for the MSDA's formulation) and the framework which helps to design a comparative analysis for evaluating the proposed alignment. In Section 3, we describe the clinical setting and give the results of our tests. These results show a significance performance improvement of our solution with respect to the competing approaches. Some comments and future work are presented finally in Section 4.

2 Materials and Methods

We approach the formulation of MSDA through the use of FSCM which is generally applied in the context of the feature selection/construction problem. This formulation gives the advantage of casting MSDA in the context of MI maximization or, more generally, in the field of stochastic optimization. In this section, we first discuss some useful theoretical bases, then we present the problem formulation and the feasibility to approach its estimation through the application of an algorithmic solution for the MWBM problem.



1001.1	1001.5	1010.21	1015.18
123.2	101	123.2	123.2
121.4	121.4	111	46.7
122.1	100	122.1	134
112	111	112	112
99.8	101.5	100	101.5

1001.13	1001.44	1013.21	1019.18
123.2	101	123.2	123.2
111	121.4	551	322
113	134	122.1	134
441	211	134	122
99.65	101.5	100	111

1001.1	1001.44	1010.21	1013.21
123.2		123.2	123.2
121.4		111	111
122.1		122.1	122.1
112		112	112
99.8		100	100
101		123.2	123.2
121.4		551	551
134		122.1	122.1
211		134	134
101.5		100	100

Fig. 2. Data integration is based, first, on establishing peptide correspondences through a *matching* (bold lines) between sets of mass values (e.g., A and B), then by "concatenating" the respective intensities (columns of the table).

2.1 Mutual Information

Mutual information is a widely used information-theory-based measure for the stochastic dependence of random variables (r.v.s). In this paper, it represents the fundamental tool for the formal definition of the MSDA problem. Formally, given two r.v.s X and Y, MI can be defined as $\mathcal{I}(Y, X) = H(X) + H(Y) - H(X, Y)$, where $H(X)$ is the entropy of X and measures the uncertainty associated with it. When X is a discrete r.v. taking values in $\{x_1, \ldots, x_k\}$ with distribution P_X, then $H(X) = -\sum_{i=1}^{k} P_X(x_i) log_2 P_X(x_i)$. $H(X, Y)$ is the joint entropy of (X, Y) which, for discrete data, assumes the value $H(X, Y) = -\sum_{i,j} P_{X,Y}(x_i, y_j) log_2 P_{X,Y}(x_i, y_j)$, where $P_X(x)$, $P_Y(y)$ and $P_{X,Y}(x, y)$ represent the marginals and the joint distributions for the bivariate (X, Y). Intuitively, MI measures the information shared by two features X and Y. In case, X and Y are independent, then knowing that X does not give any information about Y and vice versa, so their MI is zero. At the other end, if X and Y are identical, then all information conveyed by X is shared with Y: knowing that X determines the value of Y and vice versa.

2.2 Features and Model Construction

Feature construction is important for solving many complex learning tasks. One approach to this problem uses two major components. There needs to be an evaluation mechanism, which, given a set of variables, evaluates the relevance of the set; then a construction mechanism to properly define new variables. We formulate MSDA by using FSCM here below. This should be performed by defining for each lab k the following objects:

1. Let $\mathcal{P}^{(k)}$ be the set of helpful peptides population for k.[3] Each peptide $p \in \mathcal{P}^{(k)}$ has an associated random variable $I_p^{(k)}$, distributed as $f_{I_p^{(k)}}(i_p^{(k)})$, which gives the intensity value of the peptide p.
2. Let $D^{(k)}$ be a Bernoulli r.v. which gives the disease class of the patients.
3. Finally, $M_p^{(k)}$ is a random variable distributed as $f_{M_p^{(k)}}(m_p^{(k)})$ which gives the mass-to-charge ratio for the peptide p.

In order to simplify the notation (with a slight abuse of notation), we write in the following:

$$\mathrm{f}_{p,k}(i) \equiv f_{I_p^{(k)}}\big(i_p^{(k)}\big), \quad \mathrm{f}_{p,k}(I) \equiv f_{I_p^{(k)}}\big(I_p^{(k)}\big).$$

We also write the joint distribution as

$$\mathrm{f}_{p,k}(i,\ d) \equiv f_{I_p^{(k)},\ D^{(k)}}\big(i_p^{(k)},\ d^{(k)}\big),$$

3. We use the superscript to annotate the lab indexes.

and again, with an abuse of notation

$$\mathrm{f}_{p,k}(I,\ D) \equiv f_{I_p^{(k)},\ D^{(k)}}\big(I_p^{(k)},\ D^{(k)}\big).$$

At a first glance, the use of the two mechanisms follows the idea both to express a relationship between features (e.g., disease class and intensity), and then evaluate this expressed relationship. We have to do this taking into account the set of signals which can be aligned. In fact, the construction and relevance mechanisms are specifically formulated as follows:

- **Construction mechanism.** For each pair of labs (for instance, labs 1 and 2)[4] and pairs of peptides $(p,\ q)$ satisfying

$$\big|M_p^{(1)} - M_q^{(2)}\big| \leq 8, \tag{1}$$

we define

$$\begin{aligned} Z_p^{(1)} &= \lg \frac{\mathrm{f}_{p,1}(I,\ D)}{\mathrm{f}_{p,1}(I)\cdot f_{D^{(1)}}(D^{(1)})},\\ Z_q^{(2)} &= \lg \frac{\mathrm{f}_{q,2}(I,\ D)}{\mathrm{f}_{q,2}(I)\cdot f_{D^{(2)}}(D^{(2)})},\\ Z_{p,q}^{(1,2)} &= \lg \frac{\mathrm{f}_{p,1}(I,\ D)}{\mathrm{f}_{p,1}(I)\cdot f_{D^{(1)}}(D^{(1)})}\\ &\quad + \lg \frac{\mathrm{f}_{q,2}(I,\ D)}{\mathrm{f}_{q,2}(I)\cdot f_{D^{(2)}}(D^{(2)})}. \end{aligned} \tag{2}$$

In (2), it is given the dependence between the intensity and the disease class for different labs whenever the mass values are supposed to describe the same peptides entities [i.e., (1)].

- **Relevance mechanism.** It is simply obtained by taking the expectation

$$\begin{aligned} \Big\langle Z_{p,q}^{(1,2)} \Big\rangle &= \left\langle \lg \frac{\mathrm{f}_{p,1}(I,\ D)}{\mathrm{f}_{p,1}(I)\cdot f_{D^{(1)}}(D^{(1)})} \right\rangle\\ &\quad + \left\langle \lg \frac{\mathrm{f}_{q,2}(I,\ D)}{\mathrm{f}_{q,2}(I)\cdot f_{D^{(2)}}(D^{(2)})} \right\rangle, \end{aligned} \tag{3}$$

that is, the sum of MI is shared by $I_p^{(1)}$ with $D^{(1)}$ and $I_q^{(2)}$ with $D^{(2)}$, respectively, i.e., $\mathcal{I}(I_p^{(1)}, D^{(1)}) + \mathcal{I}(I_q^{(2)}, D^{(2)})$.

Once the relevances have been attributed, it is possible, for each set $\mathcal{R}$ of peptide pairs satisfying 1, to take the ones which maximize 3 that is:

$$\underset{(p,q)\in\mathcal{R}}{\operatorname{argmax}}\ \mathcal{I}(I_p^{(1)}, D^{(1)}) + \mathcal{I}(I_q^{(2)}, D^{(2)}) \subseteq \mathcal{P}^{(1)} \times \mathcal{P}^{(2)}. \tag{4}$$

Equation (4) expresses the MSDA problem by giving the pairs of peptides (which can be aligned) whose intensities share most of the information with the disease class of the patients. We emphasize the constraint $\big|M_p^{(1)} - M_q^{(2)}\big| \leq 8$; as described above, due to the instrument resolution, this is the useful range in order for two molecules (protein/peptide) to be referred to the same entity.

2.3 Maximum Weight Bipartite Matching and Data Integration

Since we consider an MI-based optimization for only two labs (in the next section, the numerical evaluation extends the results also to a third lab), the estimation of MI in (4) does not cause computational problems. Estimating MI in this case is straightforward because both the joint and marginal probability table can be obtained by discretizing and tallying, for each peptide, the samples from $\mathrm{f}_{p,1}(i,\ d)$ (or $\mathrm{f}_{q,2}(i,\ d)$), $\mathrm{f}_{p,1}(i)$ ($\mathrm{f}_{q,2}(i)$), and $f_{D^{(1)}}(d^{(1)})$ ($f_{D^{(2)}}(d^{(2)})$), respectively.

A feasible estimated computational solution for MSDA can be quite naturally obtained when one considers the MWBM problem [23]. Therefore, in the following, we reformulate problem (4) in term of bipartite graphs.

A graph $G = (V, E)$ is *bipartite* if there exists partition $V = A \cup B$ with $A \cap B = \emptyset$ and $E \subseteq A \times B$. A matching is a subset $\mathcal{M} \subseteq E$ so that $\forall v \in V$ at the most one edge in $\mathcal{M}$ is incident upon v. The size of a matching is $|\mathcal{M}|$, the number of edges in $\mathcal{M}$. When it comes to consider the weighted bipartite graphs (i.e., a function $w : E \to \Re$ exists), one can define the *weight of a matching* $\mathcal{M}$ as the sum of the weights of edges in $\mathcal{M}$: $s(\mathcal{M}) = \sum_{e\in\mathcal{M}} w(e)$. It is therefore possible to consider the following problem.

2.3.1 MWBM

Given a weighted bipartite graph G, find a matching $\mathcal{M}$ of maximum weight.

In our case, the constrain (1) induces a family of relations $\mathcal{R}_1, \mathcal{R}_2, \ldots, \mathcal{R}_n$ on $\mathcal{P}^{(1)} \times \mathcal{P}^{(2)}$ where for each $1 \leq i \leq n$:

$$\mathcal{R}_i = \big\{(p,q) : \big|M_p^{(1)} - M_q^{(2)}\big| \leq 8\big\}, \tag{5}$$

or equivalently

$$\mathcal{R}_i = \big\{\big(M_p^{(1)}, M_q^{(2)}\big) : \big|M_p^{(1)} - M_q^{(2)}\big| \leq 8\big\}. \tag{6}$$

In other words, due to (1), we handle different families of pairs of random variables where each pair expresses the mass values of potentially equivalent entities (i.e., peptides could be the same molecule) provided from different labs. Therefore, we can view our (mass-to-charge) data as observations for estimating (6). Hence, instead of (6), we use the following:

$$\tilde{\mathcal{R}}_i = \big\{\big(m_p^{(1)}, m_q^{(2)}\big) : \big|m_p^{(1)} - m_q^{(2)}\big| \leq 8\big\}, \tag{7}$$

that is, by considering

$$\begin{aligned} V_1 &= \big\{m_p^{(1)} | \exists q, j : \big(m_p^{(1)}, m_q^{(2)}\big) \in \tilde{\mathcal{R}}_j\big\},\\ V_2 &= \big\{m_q^{(2)} | \exists p, j : \big(m_p^{(1)}, m_q^{(2)}\big) \in \tilde{\mathcal{R}}_j\big\}, \end{aligned} \tag{8}$$

and $E = \bigcup \tilde{\mathcal{R}}_i$, we have a bipartite graph $G = (V_1 \cup V_2, E)$. Finally, we get an instance for MWBM by labeling each $(m_p^{(1)}, m_q^{(2)}) \in E$ as follow:

$$\begin{aligned} &w\big(\big(m_p^{(1)},\ m_q^{(2)}\big)\big)\\ &\quad = \sum_{t,d} \tilde{\mathrm{f}}_{p,1}(t,\ d) \log \frac{\tilde{\mathrm{f}}_{p,1}(t,\ d)}{\tilde{\mathrm{f}}_{p,1}(t)\cdot \tilde{f}_{D^{(1)}}(d^{(1)})}\\ &\qquad + \sum_{t,d} \tilde{\mathrm{f}}_{q,2}(t,\ d) \log \frac{\tilde{\mathrm{f}}_{q,2}(t,\ d)}{\tilde{\mathrm{f}}_{q,2}(t)\cdot \tilde{f}_{D^{(2)}}(d^{(2)})}, \end{aligned}$$

with $\tilde{\mathrm{f}}$ the associated empirical distributions.

This way, (4) can be estimated with one of the many general applied techniques for MWBM [23]. Therefore, the data integration process is performed first by establishing peptide correspondences through their mass values (i.e., a matching), and then by "concatenating" the respective intensity values (Fig. 2).

2.4 Tools of Evaluation

In this section, we briefly report on the environments and methods we considered and used for the numerical evaluation process. We save for the last section some comments on the results obtained.

The performances for evaluating the comparison among different alignment methods were obtained through the design of a Rapid Miner (v4.4) process [24]. Rapid Miner is a machine learning environment where a knowledge discovery process (KD

4. Almost all of our numerical experiments are performed on pairs of labs. Even the general formulation to n-tuple is straightforward, we maintain here this constraint to simplify the annotation.

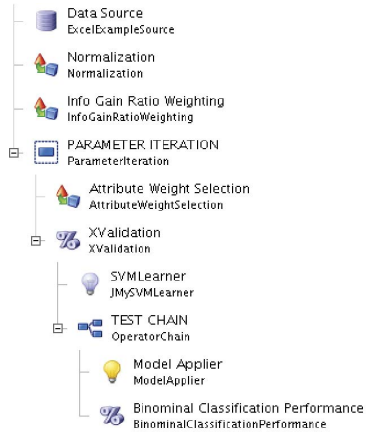


Fig. 3. Evaluation process. Each block in the figure explains a step in our alignment evaluation.

process) is modeled by a complex nested chain (tree) of objects called operators. These operators implement several KD processes, such as data preprocessing, performance evaluation, visualization, learning algorithms, etc. Fig. 3 represents the KD process used, in our case, to analyze and extrapolate the performance values for each data combination method. In other words, it describes the experiments we designed in order to compare different data alignment techniques. In Fig. 3, blocks correspond to simple process steps in the whole design: each operator receives an input and delivers an output to the forward operator. The information flow is similar to an *in depth first* search [25] of normal trees. Here, we give a short description of what each operator implements in our evaluation process.

- **Data Source Operator** reads data from files. In our case, it is an combined data set obtained with the application of a specific data alignment method.
- **Normalization Operator** normalizes data in $[-1, 1]$.
- **Information Gain Operator** computes an MI-based score for weighting the relevance of each feature (i.e., signal). We use this step in order to base the inference process on the top $2 \leq K \leq 12$ highest value features. The extrapolation step is then realized with the Attribute Weight Selection Operator.
- **Parameter Iteration Operator** uses some defined parameters and performs the inner operators for all possible combinations of them (e.g., in our case for different number of signals). We iterated the inner operators, changing the number of features to consider the inference process.
- **XValidation Operator.** XValidation Operator encapsulates a cross-validation process: the input data set S is split up into subsets $\{S_1, S_2, \ldots, S_n\}$. The inner operators are applied n times using at each iteration i the set S_i as the test set and $S \backslash S_i$ as the training set.
- **SVM Operator** implements a Support Vector Machine algorithm (see, for example, [26]) to deliver an inference model. We used SVM as a black box inference process to measure the performance for each combined input data set.
- **Model Applier Operator** applies the model delivered by the SVM operator.
- **Binomial Classification Performance Operator** collects the performance evaluation for the classification task and outputs the performance measures. We measure here the performance by using the Area Under ROC Curve (AURC) and the Precision index.

3 Experimental Results

3.1 Clinical Setting

Samples were collected after receiving informed consent from all the subjects participating in the study from three different hospitals using a standardized protocol. A cohort of six control subjects and nine AD patients was recruited from the Università degli Studi di Firenze—School of Medicine network (Florence, Italy), 23 controls and 18 AD patients from San Gerardo Hospital (Monza, Italy), and a total of 6 controls and 15 AD patients from the Center for Aging Brain and Dementia (Brescia, Italy). Plasma was obtained from blood collected in EDTA.

3.1.1 Plasma Purification

Sample purification was performed in duplicate at room temperature with ClinProt MB-HIC8 (Magnetic-Beads-based Hydrophobic Interaction Chromatography) kit. All processes were automatically executed by using a ClinProt Robot as previously described [13].

3.1.2 MALDI-TOF MS and Data Processing

The plasma protein profiles were obtained by an MALDI-TOF Reflex IVTM mass spectrometer (Bruker Daltonics). The instrument was externally calibrated using a mixture of standard peptides/proteins. Mass spectra were acquired in positive linear mode in the m/z range of 1,000-10,000 Daltons; accumulation of signals from different spot positions resulted in a total averaging spectrum. The spot was preirradiated with higher laser power to improve the spectra quality before each acquisition cycle. Multiple spectra comparison was performed using ClinProToolsTM 2.1 software (Bruker Daltonics). First, each raw spectrum was normalized and all spectra were then recalibrated (realignment) using prominent internal m/z values. Subsequently, baseline subtraction and peak detection were achieved before peak area calculation. The software calculates the mean spectrum for each subject's data set, and then, selects the spectrum that is most similar to the average one to be used for further evaluations. ClinProTools automatically provided a list of peaks sorted according to the statistical relevance to differentiate between classes with their corresponding p-value.

3.2 Numerical Evaluations

The data sets obtained from the proteomic analysis of these biological samples will be identified with the name of the city where the labs come from, i.e., Florence, Monza, and Brescia. Our intent is to evaluate the application of the MI-based data fusion (labeled as "MI-based" in this section) by comparing the inference results with other two different methods. This evaluation has been initially obtained by integrating all the following combinations of data sets: Monza + Florence (as MF data), Monza + Brescia (MB data), Florence +Brescia (FB data), and Monza + Florence + Brescia data (MFB data). Next, we used the two competitive approaches, called, respectively, Equal Mass Fusion (labeled as "EM-based") and T-Test based (or "TT-based") to complete the evaluation procedure.

- *EM-based fusion.* The features from different labs have been unified whenever the associated mass values were equal.
- *TT-based fusion.* For all pair of features whose mass difference ranges in an interval of $+/-8$ units, we compared the means from two different samples by a statistical t-Test. Then, we unified these pairs of features with the maximum value of significance.

In order to evaluate the performances of the above approaches, we used integrated data sets as input to the same inference procedure (SVM operator). As reported in Section 2.4, the inference is performed to predict the disease class of the patients. The

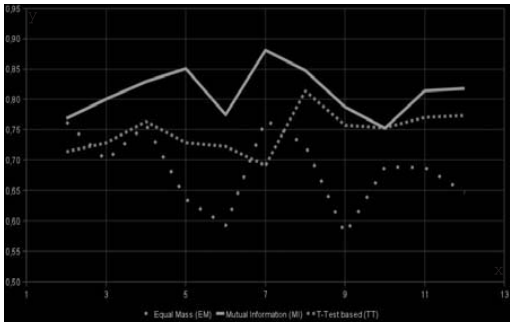


Fig. 4. Average AURC (y-axis) for different number of features (x-axis) reported for each method.

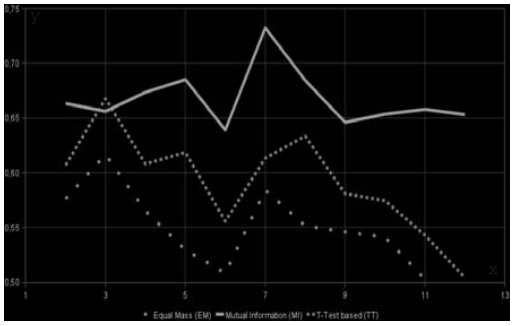


Fig. 5. Average precision (y-axis) for different number of features (x-axis) reported for each method.

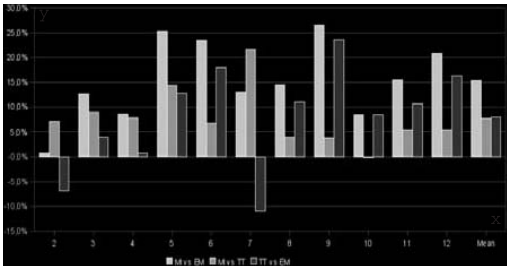


Fig. 6. Average AURC; percentage (y-axis) for different number of features (x-axis). The last point on the x-axis refers to the average value computed over all former values.

different methods are fed by their integrated data sets, and are ranked by comparing the results thus obtained. The results presented in this section extend those in [27] through the evaluations of the methods for different numbers of ranked features (see, for instance, [21]). Therefore, performances have been compared by computing the precision index (i.e., that fraction of examples classified as positive that are truly positive) with respect to case (AD) patients, and AURC, respectively. We recall that the ROC curve can be represented by plotting the fraction of true positives' examples versus the fraction of false positives' examples (see, for instance, [28]).

Average values for AURC are shown in Fig. 4: the MI-based approach appears to be generally better than the competitive methods. Similar results are confirmed when evaluating the precision index (Fig. 5). In Figs. 6 and 7, the percentages regarding how a first method m_1 outperforms the second m_2 in a pairwise comparison of m_1 versus m_2 are shown. For example, in Fig. 6—specifically comparing MI versus TT—it is shown that MI behaves, on average, 7 percent better than TT for two

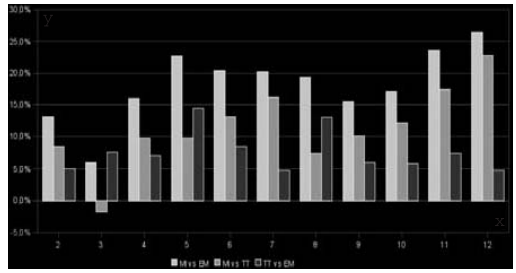


Fig. 7. Average precision; percentage (y-axis) for different number of features (x-axis).

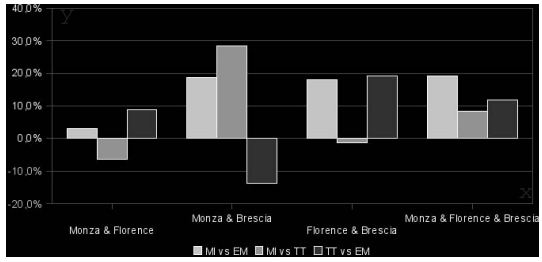


Fig. 8. AURC performance (y-axis) for each integrated data set (x-axis).

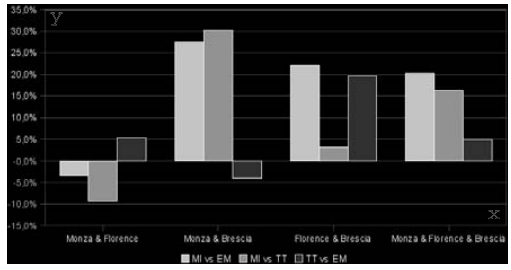


Fig. 9. Precision performance (y-axis) for each integrated data set (x-axis).

features. On the contrary when considering seven features, TT has, on average, a 11 percent lower performance than EM. These percentage are computed by averaging the respective performance indexes (AURC in Fig. 6 and Precision in Fig. 7) over all integrated data set. In Figs. 8 and 9, comparison for specific integrated data is finally considered. The aggregation of all data sets (reported as "Monza & FIrenze & Brescia") seems to give better results than data from a subset of them increasing approximately of about 20 percent for AURC (Fig. 8) and Precision, respectively (Fig. 9).

4 Conclusions

In this paper, we have presented the results of our development for the mutual application optimization in the study of mass spectra alignment of data provided from different labs. Our contributions are the following:

- We studied the problem of signal alignments in a setting where a sensible fusion criterion permits to align peptide profile measurements of patients from different labs.
- We formalized the general problem using a feature construction methodology, i.e., combining these features with maximum information content with the resulting combined target patient classes.

- We further applied the Maximum Weight Bipartite problem formalization in order to give an estimation for the general setting described above.
- We finally showed the validity of the proposed method on three different real data sets by comparing performances with two other approaches. These comparisons have the target to show that with the given formulation (MSDA), one has suitable solutions (e.g., MWBM) which do not fail like a "rough" method (i.e., EM) that performs poorly. On the contrary, these applications encourage us by performing better than a specific statistical t-test-based approach. We detailed the evaluation process for the alignment comparison through a Rapid Miner process. This process explains, through the way in which its base blocks are organized, how we faced—we believe, a quite complex phase for evaluating results (embedding normalization, cross validation, support vector inference, and feature ranking).

This study was realized on a small sample; thus, it is necessary to validate our results with a wider number of Alzheimer's patients and other techniques. Extending these numerical experiments should strengthen the evidence for the proposed approach validity for this kind of inference task. In addition, it is important to verify the diagnostic efficacy of these predictive models in a blind manner on samples from subjects with different neurodegenerative pathologies.

REFERENCES

[1] R.N. Kalaria, G.E. Maestre, R. Arizaga, R.P. Friedland, D. Galasko, K. Hall, J.A. Luchsinger, A. Ogunniyi, E.K. Perry, F. Potocnik, M. Prince, R. Stewart, A. Wimo, Z.X. Zhang, and P. Antuono, "World Federation of Neurology Dementia Research Group. Alzheimer's Disease and Vascular Dementia in Developing Countries: Prevalence, Management, and Risk Factors," *Lancet Neurology*, vol. 7, no. 9, pp. 812-826, 2008.

[2] R.J. Caselli, T.G. Beach, R. Yaari, and E.M. Reiman, "Alzheimer's Disease a Century Later," *J. Clinical Psychiatry*, vol. 67, pp. 1784-1800, 2007.

[3] S. Small and K. Duff, "Linking Abeta and Tau in Late-Onset Alzheimer's Disease: A Dual Pathway Hypothesis," *Neuron*, vol. 60, no. 4, pp. 534-542, 2008.

[4] T. Wyss-Coray, "Inflammation in Alzheimer Disease: Driving Force, Bystander or Beneficial Response?," *Nature Medicine*, vol. 12, no. 9, pp. 1005-1015, 2006.

[5] D. Praticò, "Evidence of Oxidative Stress in Alzheimer's Disease Brain and Antioxidant Therapy: Lights and Shadows," *Annals of New York Academy of Sciences*, vol. 1147, pp. 70-78, 2008.

[6] E. Koutsilieri and P. Riederer, "Excitotoxicity and New Antiglutamatergic Strategies in Parkinson's Disease and Alzheimer's Disease," *Parkinsonism and Related Disorders*, vol. 13, pp. S329-S331, 2007.

[7] S. Ray, M. Britschgi, C. Herbert, Y. Takeda-Uchimura, A. Boxer, K. Blennow, L. Friedman, D. Galasko, M. Jutel, A. Karydas, J.A. Kaye, J. Leszek, B.L. Miller, L. Minthon, J.F. Quinn, G.D. Rabinovici, W.H. Robinson, M.N. Sabbagh, Y.T. So, D.L. Sparks, M. Tabaton, J. Tinklenberg, J.A. Yesavage, R. Tibshirani, and T. Wyss-Coray, "Classification and Prediction of Clinical Alzheimer's Diagnosis Based on Plasma Signaling Proteins," *Nature Medicine*, vol. 13, no. 11, pp. 1359-1362, 2007.

[8] M. Latterich, M. Abramovitz, and B. Leyland-Jones, "Proteomics: New Technologies and Clinical Applications," *European J. Cancer*, vol. 44, pp. 2737-2741, 2008.

[9] K. Landers, M. Burger, M. Tebay, D. Purdie, B. Scells, H. Samaratunga, M. Lavin, and R. Gardiner, "Use of Multiple Biomarkers for a Molecular Diagnosis of Prostate Cancer," *Int'l J. Cancer*, vol. 114, pp. 950-956, 2005.

[10] D.M. Good, V. Thongboonkerd, J. Novak, J.L. Bascands, J.P. Schanstra, J.J. Coon, A. Dominiczak, and H. Mischak, "Body Fluid Proteomics for Biomarker Discovery: Lessons from the Past Hold the Key to Success in the Future," *J. Proteome Research*, vol. 6, no. 12, pp. 4549-4555, 2007.

[11] M.E.D. Noo, B.J. Mertens, A. Ozalp, M.R. Bladergroen, M.P. van der Werff, C.J. van de Velde, A.M. Deelder, and R.A. Tollenaar, "Detection of Colorectal Cancer Using Maldi-Tof Serum Protein Profiling," *European J. Cancer*, vol. 42, no. 8, pp. 1068-1076, 2006.

[12] G.L. Freed, L. Cazers, C. Fichandler, T. Fuller, C. Sawyer, B.C.J. Stack, S. Schraff, O.J. Semmes, J.T. Wadsworth, and R. Drake, "Differential Capture of Serum Proteins for Expression Profiling and Biomarker Discovery in Pre- and Posttreatment Head and Neck Cancer Samples," *Laryngoscope*, vol. 118, no. 1, pp. 61-68, 2008.

[13] N. Bosso, C. Chinello, S. Picozzi, E. Gianazza, V. Mainini, C. Galbusera, F. Raimondo, R. Perego, S. Casellato, F. Rocco, S. Ferrero, S. Bosari, P. Mocarelli, M.G. Kienle, and F. Magni, "Human Urine Biomarkers of Renal Cell Carcinoma Evaluated by Clinprot," *Proteomics—Clinical Application*, vol. 2, nos. 7/8, pp. 1036-1046, 2008.

[14] N. Barbarini, P. Magni, and R. Bellazzi, "A New Approach for the Analysis of Mass Spectrometry Data for Biomarker Discovery," *Proc. AMIA Ann. Symp.*, pp. 26-30, 2006.

[15] H. Hotelling, "Relation between Two Sets of Variates," *Biometrika*, vol. 28, pp. 321-377, 1936.

[16] W.W. Hsieh, "Nonlinear Canonical Correlation Analysis by Neural Networks," *Neural Networks*, vol. 13, pp. 1095-1105, 2000.

[17] P.A. Viola, W.M. Wells III, "Alignment by Maximization of Mutual Information," *Int'l J. Computer Vision*, vol. 24, no. 2, pp. 137-154, 1997.

[18] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 2000.

[19] M. Hilario, A. Kalousis, M. Mller, and C. Pellegrini, "Machine Learning Approaches to Lung Cancer Prediction from Mass Spectra," *Proteomics*, vol. 3, no. 9, pp. 1716-1719, 2003.

[20] K. Torkkola, "Feature Extraction by Non-Parametric Mutual Information Maximization," *J. Machine Learning Research*, vol. 3, pp. 1415-1438, 2003.

[21] *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh, eds., Springer, 2006.

[22] L. Lovász and M. Plummer, *Matching Theory*. Akadémiai Kiadó, 1986.

[23] J.A. McHugh, *Algorithmic Graph Theory*. Prentice Hall, 1990.

[24] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks," *KDD '06: Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, L. Ungar, M. Craven, D. Gunopulos, and T. Eliassi-Rad, eds., pp. 935-940, 2006.

[25] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*. MIT Press, 1990.

[26] N. Cristianini and J. Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge Univ. Press, 2000.

[27] I. Zoppis, E. Gianazza, C. Chinello, V. Mainini, C. Galbusera, C. Ferrarese, G. Galimberti, A. Sorbi, B. Borroni, F. Magni, and G. Mauri, "A Mutual Information Approach to Data Integration for Alzheimer's Disease Patients," *Lecture Notes in Computer Science*, pp. 431-435, Springer, 2009.

[28] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and Roc Curves," *Proc. 23rd Int'l Conf. Machine Learning (ICML '06)*, pp. 233-240, 2006.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Analysis of Correlation Structures in Renal Cell Carcinoma patient data

Citation: Italo Zoppis, Massimiliano Borsani, Erica Gianazza, Clizia Chinello, Francesco Rocco, Giancarlo Albo, André M. Deelder, Yuri E. M. van der Burgt, Fulvio Magni, Marco Antoniotti and Giancarlo Mauri. Analysis of Correlation Structures in Renal Cell Carcinoma Patient Data. *BIOINFORMATICS 2012*, 251-256.

Document Type: Conference Paper (source: Scopus)
ISBN: 978-989842590-4
DBLP: conf/biostec/2012bi

ANALYSIS OF CORRELATION STRUCTURES IN RENAL CELL CARCINOMA PATIENT DATA

Italo Zoppis[1], Massimiliano Borsani[1], Erica Gianazza[2], Clizia Chinello[2], Marco Antoniotti[1], André M. Deelder[3], Yuri E. M. van der Burgt[3], Fulvio Magni[2] and Giancarlo Mauri[1]

[1]*Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milano, Italy*

[2]*Department of Experimental Medicine, University of Milano-Bicocca, Monza, Italy*

[3]*Department of Parasitology, Biomolecular Mass Spectrometry Unit, Leiden University Medical Center, Leiden, The Netherlands*

Keywords: Proteomics, Mass Spectrometry, Hypotheses Testing, Clinical Analysis, Correlation, Bipartite Graphs

Abstract: Mass Spectrometry (MS)-based technologies represent a promising area of research in clinical analysis. They are primarily concerned with measuring the relative intensity (abundance) of many protein/peptide molecules associated with their mass-to-charge ratios over a particular range of molecular masses. These measurements (generally referred as *proteomic signals* or *spectra*) constitute a huge amount of information which requires adequate tools to be investigated and interpreted. Following the methodology for testing hypotheses, we investigate the *proteomic signals* of the most common type of Renal Cell Carcinoma, the *Clear Cell* variant (ccRCC). Specifically, the aim of our investigation is to detect changes of the signal correlations from control to case group (ccRCC or non–ccRCC). To this end, we sample and represent each population group through a graph providing, as it will be defined below, the observed *signal correlation structure*. This way, graphs establish abstract frames of reference in our analysis giving the opportunity to test hypotheses over their properties. In other terms, changes are detected by testing graph property modifications from group to group. We show the results by reporting the *mass-to-charge* values which identify bounded regions where changes have been detected. The main interest in handling these regions is to perceive which signal ranges are associated with some specific factors of interest (e.g., studying differentially expressed peaks between case and control groups) and thus, to suggest potential biomarkers for future analysis or for clinical monitoring. Data were collected, from patients and healthy volunteers at the Ospedale Maggiore Policlinico Foundation (Milano, Italy).

1 INTRODUCTION

Renal Cell Carcinoma (RCC) is the most common tumor in the adult kidney and accounts for about 3-4% of all adult malignancies (Brannon and Rathmell, 2010). The most frequent histological subtype (60-80%) is the Clear Cell variant (ccRCC). There are currently no biomarkers available for its early detection, for an efficient prognosis, and for optimal predictive therapeutic approaches (Drucker, 2005). At present, proteomics represents a good tool for defining biomarkers in biological fluids which can characterize and predict multifactorial diseases. In this context, *Mass Spectrometry* (MS) techniques have recently been playing an important role in studying biological samples. They are primarily concerned with measuring the relative intensity (abundance) of many protein/peptide molecules associated with their mass-to-charge ratios over a particular Dalton range. The resulting measurements are often displayed as a graph

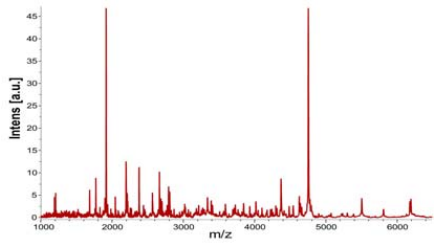


Figure 1: A typical *protein/peptide profile*.

– a *protein/peptide profile* like the one in Fig. 1, in which each *peak* (or *signal*) identifies the pair of values given by the intensity (related to the abundance) of a molecule (y–axis) with its specific molecular mass-to-charge ratio (x–axis). The final interest in handling the huge amount of data produced from these analyses is to perceive which peaks are associated with some specific factors of interest (e.g.,

studying differentially expressed peaks between case and control groups) and thus, to suggest potential biomarkers for future analysis (Latterich et al., 2008). However, to our knowledge, most of these studies omit to consider the following key-points.

Constrained Classification. Case / Control discrimination requirements for real-world problems are often constrained by a given true positive or false positive rate to ensure that the classification error for the most important class is within a desired limit.

Relational Information. Many domains are best described by relational models in which instances of multiple types are related to each other in complex ways – see for example (Getoor and Taskar, 2007). In this case, some features of one entity are often correlated with features of related entities. It is intuitive that, just as some features are not helpful for mining data sets, some relations might provide informations for clustering or classification algorithms. When it comes to analyze differentially expressed peaks in a case/control classification problem, comparisons are generally performed between protein/peptide profiles of different groups – or between statistics summarizing the peaks' property of a group, (Solassol et al., 2006). Actually, different neighborhoods in the m/z spectra can be (anti)correlated each other and, this property, in turn, may change from group to group. In such a situation, the incorporation of relational information may increase the performance of the system for "difficult" data sets.

In order to manage the above issues, we formulate our framework as follow.

1. The *constrain requirement* is met following a standard *test of hypothesis* approach. This way, one must *decide* between a null hypothesis and an alternative hypothesis. A level of significance α (called the size of the test) is imposed on the false alarm probability (*type I error*), and one seeks a test that satisfies this constraint. The experimental design which derive from this formulation provide us with a tool for detecting regions of the proteomic spectra characterized by properties differentially expressed from group to group. Specifically, in these region *correlations* between signals are a "powerful" discrimination factor between groups. This detection is our primary interest in this paper.
2. *Relational informations* are introduced by giving new graph representations for the observed samples. This way, as is used to represent relationships of many interacting entities, we express correlations between signals in the m/z spectra of a patient group. Throughout, we call these representations *correlation structures* (shortly, *templates*). Arguments of our hypotheses state conjectures over specific graph (i.e., template) properties. Therefore, by testing hypotheses over properties, we can decide whether these graphs have been changed from control to case groups (i.e, either ccRCC or non-ccRCC groups).

Given the above concerns, this paper is laid out as follows. In sections 2 we introduce the preliminaries and notations. In section 3 we formulate the problem. In section 4 we report the clinical setting and some numerical results. Finally, in section 5 we conclude the paper by discussing some issues of this work.

2 BASIC DEFINITIONS AND NOTATION

Graphs are important structures to model a wide range of natural phenomena, particularly when one has to represent complex systems of interactions among entities. Throughout this paper $G = (V_1 \cup V_2, E)$ denotes a *oriented bipartite graph*; that is, V_1 and V_2 are two sets of *vertices* such that the set of all *arcs* $E \subseteq V_1 \times V_2$ connect vertices in one set with vertices in the other: i.e., E is a set of ordered pairs (v_i, v_j) with $v_i \in V_1$ and $v_j \in V_2$ constrained to not contain any of the arcs (v_i, v_j) and (v_j, v_i). Given an *oriented bipartite graph* $G = (V_1 \cup V_2, E)$, the *subgraph* of G given by $\tilde{G} = (\tilde{A}, \tilde{E})$, with $\tilde{A} \subseteq V_1 \cup V_2$ and $\tilde{E} \subseteq E$ is a *biclique* if, for all $v_1 \in (\tilde{A} \cap V_1)$ and $v_2 \in (\tilde{A} \cap V_2)$ then $(v_1, v_2) \in \tilde{E}$. Biclique are, therefore, "extreme" forms of highly inter-connected bipartite graphs and they will of interest in defining indexes for our analisys. The number of vertexes $N_v = |V_1 \cup V_2|$ and the number of arcs $N_e = |E|$ are generally called the *order* and the *size* of the graph. Moreover, graphs can be, generally, "summarized" in a compact way by various *graph properties*. Among all the properties in literature (Brandes and Erlebach, 2005), here we focus on *cohesion*. A well known index to characterize this notion is that of *density*. We treat the subject in order to give a "local" scale of characterization for it. While, in general, with a "global" density, we can characterize the cohesion on the whole graph, with a *local density* index as we will define below, we wish to analyze the cohesion (i.e., by testing hypotheses), on differently located parts of the graph. Before introducing formally this notion we give the following definition.

Definition 1 (Neighborhood). *Let* $G = (V_1 \cup V_2, E)$

*be an oriented bipartite graph with V_1, V_2 two well-ordered sets of vertexes. We call $M_{i,j,k}(G) = (\tilde{A}, \tilde{E})$ a $(i,j,k)-$*neighborhood *(or simply, a neighborhood $M_{i,j,k}$ centered in (v_i, v_j)) the subgraphs of G induced by $\tilde{A} = \tilde{V}_{i,k} \cup \tilde{V}_{j,k}$ where $\tilde{V}_{i,k} = \{v_{i-k}, \ldots, v_i, \ldots, v_{i+k}\}$ and $\tilde{V}_{j,k} = \{v_{j-k}, \ldots, v_j, \ldots, v_{j+k}\}$*[1].

We are now able to give the following definition.

Definition 2 (Local density)**.** *Let $G = (V_1 \cup V_2, E)$ be an oriented bipartite graph and $M_{i,j,k} = (\tilde{A}, \tilde{E})$ a neighborhood of size S centered in (v_i, v_j), we define the local density of G in $M_{i,j,k}$ as*

$$den(M_{i,j,k}) = \frac{S}{|\tilde{V}_{i,k} \times \tilde{V}_{j,k}|}. \quad (1)$$

The local density is based on the ratio of the number of arcs among a subset of vertices to the total number of possible arcs. This way they provide a measure of "how close" $M_{i,j,k}$ is to being an *oriented biclique*. Since our primary interest is to detect which regions of the spectra express different properties from control to case group (in our case, correlation structure properties) we stress this point with the following definition.

Definition 3 (Bipartite Graph Region)**.** *Let $G = (V_1 \cup V_2, E)$ be an* oriented bipartite graph *with V_1, V_2 two well-ordered sets of vertexes. We say that S is a region of G if it is the subgraph $S = (\tilde{V}_1 \cup \tilde{V}_2, \tilde{E})$ induced through the two sequences of vertexes $\tilde{V}_1$ and $\tilde{V}_2$.*

For a formal point of view, definition 3 says nothing more than *S* is a *subgraph* induced by a set of vertexes. We give this definition purely as a matter of convenience to point out that any *region* of the proteomic spectra (i.e., a sequence of mass-to-charge ratio values) is represented here through the *region* of a bipartite graph. We use widely this term in section 3 to formulate our testing procedures.

3 PROBLEM FORMULATION

In this section we formally define the problem inside the standard *test of hypotheses* framework. The subjects of our formulation are tests concerning graphs properties which can be easily obtained from the following new samples representations. We start by considering a population of interest divided into two groups; respectively *case* and *control* subjects. This population expresses the signal intensity values observable in different regions over the spectra. We sample and represent each population group through graphs which provide the observed *signal correlation structure* as will be defined below in section 3.1. This way, graphs establish abstract frames of reference in our analysis giving the opportunity to test hypotheses over their properties (section 3.2). In other terms, changes are detected by testing graph property modifications from group to group. The whole procedure provide the mass-to-charge Dalton ranges bounding the regions where significant changes have been detected.

3.1 Correlation Structure Representation

As is used to represent structures of many interacting entities, we can express correlations inside patients' groups through a graph whose vertexes are specific mass-to-charge ratios and arcs "express" correlations between signal intensities with these specific mass-to-charge values. We call the resulting representation, the (observed) *correlation structure* (briefly, *template*). More formally, we denote the groups of control and case subjects with I^{ctrl} and I^{case} respectively. We assume that each group (for instance I^{ctrl}) can be expressed through a product $I^{\text{ctrl}}_{m_1} \times I^{\text{ctrl}}_{m_2} \times \ldots \times I^{\text{ctrl}}_{m_n}$ of *spaces* $I^{\text{ctrl}}_{m_i}$, $i \in [n]$ [2], given by all potential intensity values whose mass-to-charge ratio is m_i. We also assume that each $I^{\text{ctrl}}_{m_i}$ is endowed with a distribution function $f^{\text{ctrl}}_{I_{m_i}}$. More in general, let us give the following definition for any group of patients *g* on which is defined a distribution $f^{\text{g}}_{I_{m_i}}$.

Definition 4 (Template)**.** *By sampling from each pair $(f^g_{I_{m_i}}, f^g_{I_{m_j}})$, with $i \in [n]$, $j \in [n]$, two sets of i.i.d. random variables $\{I^g_{m_{i,1}}, I^g_{m_{i,2}}, \ldots, I^g_{m_{i,n}}\}$ and $\{I^g_{m_{j,1}}, I^g_{m_{j,2}}, \ldots, I^g_{m_{j,n}}\}$, we call* template *(of g) the bipartite graph $R^g = (V_1 \cup V_2, E)$ with vertexes $V_1 = \{m_1, m_2, \ldots, m_n\}$ and $V_2 = \{m'_1, m'_2, \ldots, m'_n\}$. Moreover, $(m_i, m'_j) \in E$ only if the absolute value of the Pearson's correlation coefficient exceeds a threshold δ. That is,*

$$\rho^g_{i,j} = \frac{\sum_{k=1}^{n} (I^g_{m_{i,k}} - \overline{I^g_{m_i}})(I^g_{m_{j,k}} - \overline{I^g_{m_j}})}{\sqrt{\sum_{k=1}^{n} (I^g_{m_{i,k}} - \overline{I^g_{m_i}})^2} \sqrt{\sum_{k=1}^{n} (I^g_{m_{j,k}} - \overline{I^g_{m_j}})^2}} \geq \delta, \quad (2)$$

where $\overline{I^g_{m_i}}$ and $\overline{I^g_{m_j}}$ are the sample means.

Notice that, given the template $R^{\text{g}} = (V_1, V_2, E)$ and any *region S* of R^{g}, we can easily provide a set of densities $\{d_1, d_2, \ldots, d_n\}$ by observing a set of neighborhoods in *S*. For example, in Fig. 2 is reported

[1] We also refer to the pair (v_i, v_j) and the constant *k* as, respectively, the center and the ray of the *neighborhood*

[2] We use the bracket notation $[n]$ to denote the set $\{1, \ldots, n\}$ of the first *n* positive integers.

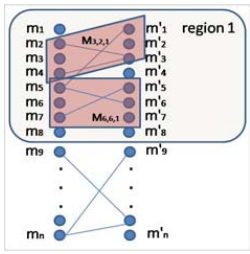


Figure 2: The bipartite graph for RCC data (template) with one region and two neighborhoods.

a subgraph of R^{g} with one region and two neighborhoods M_1^g and M_2^g.[3] Yet it is clear that, these neighborhoods provide the set of local density values $D_S^{\text{g}} = \{\text{den}(M_1^g), \text{den}(M_2^g)\}$. We assume that D_S^{g} are observations from a distribution (of densities) referred to the region S. Throughout, we will consider for any pair of templates R^{ctrl} and R^{case} the set of densities D_S^{ctrl} and D_S^{case} as samples of observations realized in a common region S to test local hypotheses over a (density) population.

3.2 Hypothesis testing

We recall that, statistical hypotheses (noted as H_0 and H_A) are competing statements concerning the population parameters. The rationale for establishing our hypotheses is deciding whether a pathology (for instance, ccRCC) has modified the cohesion of a control group's correlation structure. Since we use density to analyze cohesions, we should also say that for two groups of densities, to be consistent with the above rationale, it suffices that $\mu^{\text{ctrl}} \neq \mu^{\text{case}}$, where μ^{ctrl} and μ^{case} are the means in the control and case groups of densities. Therefore, given (i) the (paired) samples of densities $D^{\text{ctrl}} = \{X_1, X_2, ..., X_n\}$ from controls, and $D^{\text{case}} = \{Y_1, Y_2, ..., Y_n\}$ from cases, (ii) their differences $D = \{D_i : D_i = X_i - Y_i, X_i \in D^{\text{ctrl}}, Y_i \in D^{\text{case}}\}$, (iii) the sample mean $\tilde{D}$ and (iv) the sample standard deviation of difference scores S_d, we can reject the *null* $H_0 : \mu^{\text{ctrl}} = \mu^{\text{case}}$ (no change) in favor of the *alternative* $H_A : \mu^{\text{ctrl}} \neq \mu^{\text{case}}$ using

$$T = \frac{\tilde{D}}{S_d/\sqrt{n}} \tag{3}$$

as *test statistic* which, in turn, follows a Student's *t*-distribution with $n-1$ degree of freedom if H_0 is true. Thus, we apply a classical two-sample, paired t-test, rejecting the null when the realization t of the statistic in expression 3 is such that $|t| > t_{1-\alpha/2}(n-1)$, where $t_{1-\alpha/2}(n-1)$ is the quantile of Student's t-distribution with $n-1$ degrees of freedom. As argued above, the use of local densities gives us the opportunity to analyze the cohesion in different parts of the graph. This way, we can consider different regions over the spectra – through different "local statistics", and perform different tests. Specifically, as noted in section 3.1, given a common region S for both (the templates) R^{ctrl} and R^{case}, we obtain two sets of densities D_S^{ctrl} and D_S^{case}. As previously stated, using these data as observations provided by sampling both the control and the case groups in S, we are able to apply the test $H_0 : \mu_S^{\text{ctrl}} = \mu_S^{\text{case}}$ against $H_A : \mu_S^{\text{ctrl}} \neq \mu_S^{\text{case}}$ for any region S; that is, by observing different regions, we test the cohesion modifications from group to group in different parts of the spectra. Given the above arguments, we can define different classes of case/control tests thought the following procedures:

- Control vs. ccRCC Tests (briefly noted CVR Tests)
  1. We represent R^{ctrl} by sampling from each pair $(f_{l_{m_i}}^{ctrl}, f_{l_{m_j}}^{ctrl})$ – in the control group, the sets of i.i.d rvs $\{I_{m_i,1}^{\text{ctrl}}, I_{m_i,2}^{\text{ctrl}}, \ldots, I_{m_i,n}^{\text{ctrl}}\}$ and $\{I_{m_j,1}^{\text{ctrl}}, I_{m_j,2}^{\text{ctrl}}, \ldots, I_{m_j,n}^{\text{ctrl}}\}$.
  2. We represent R^{rcc} by sampling from each pair $(f_{l_{m_i}}^{\text{rcc}}, f_{l_{m_j}}^{\text{rcc}})$ – in the ccRCC group, the sets of i.i.d rvs $\{I_{m_i,1}^{\text{rcc}}, I_{m_i,2}^{\text{rcc}}, \ldots, I_{m_i,n}^{\text{rcc}}\}$ and $\{I_{m_j,1}^{\text{rcc}}, I_{m_j,2}^{\text{rcc}}, \ldots, I_{m_j,n}^{\text{rcc}}\}$.
  3. Given any region S, common both to R^{ctrl} and R^{rcc}, we obtain the local densities $D_S^{\text{ctrl}} = \{\text{den}(M_1^{\text{ctrl}}), \text{den}(M_2^{\text{ctrl}}), \ldots, \text{den}(M_n^{\text{ctrl}})\}$ and $D_S^{\text{rcc}} = \{\text{den}(M_1^{\text{rcc}}), \text{den}(M_2^{\text{rcc}}), \ldots, \text{den}(M_n^{\text{rcc}})\}$. Then for each S, we employ these sets (as observations from a density population) together with Eq. 3 (as test statistic) in the following tests: $H_0 : \mu_S^{\text{ctrl}} = \mu_S^{\text{rcc}}$ Vs. $H_A : \mu_S^{\text{ctrl}} \neq \mu_S^{\text{rcc}}$, where μ_S^{ctrl} and μ_S^{rcc} are, respectively, the (population) means of the densities in the control and ccRCC groups.
- Control vs. non-ccRCC Tests (CVNR Tests)
  1. We represent R^{ctrl} by sampling from each pair $(f_{l_{m_i}}^{ctrl}, f_{l_{m_j}}^{ctrl})$ – in the control group, the sets of i.i.d rvs $\{I_{m_i,1}^{\text{ctrl}}, I_{m_i,2}^{\text{ctrl}}, \ldots, I_{m_i,n}^{\text{ctrl}}\}$ and $\{I_{m_j,1}^{\text{ctrl}}, I_{m_j,2}^{\text{ctrl}}, \ldots, I_{m_j,n}^{\text{ctrl}}\}$.
  2. We represent R^{nrc} by sampling from each pair $(f_{l_{m_i}}^{\text{nrc}}, f_{l_{m_j}}^{\text{nrc}})$ – in the non-ccRCC group, the sets of i.i.d rvs $\{I_{m_i,1}^{\text{nrc}}, I_{m_i,2}^{\text{nrc}}, \ldots, I_{m_i,n}^{\text{nrc}}\}$ and $\{I_{m_j,1}^{\text{nrc}}, I_{m_j,2}^{\text{nrc}}, \ldots, I_{m_j,n}^{\text{nrc}}\}$.

[3] For sake of clarity to specify the group g from which the neighborhood M is drawn, we also use the notation M^g.

3. Given any region S, common both to R^{ctrl} and R^{nrc}, we obtain the local densities $D_S^{\text{ctrl}} = \{\text{den}(M_1^{\text{ctrl}}), \text{den}(M_2^{\text{ctrl}}), \ldots, \text{den}(M_n^{\text{ctrl}})\}$ and $D_S^{\text{nrc}} = \{\text{den}(M_1^{\text{nrc}}), \text{den}(M_2^{\text{nrc}}), \ldots, \text{den}(M_n^{\text{nrc}})\}$. Then for each S, we employ these sets (as observations from a density population) together with Eq. 3 (as test statistic) in the following tests: $H_0 : \mu_S^{\text{ctrl}} = \mu_S^{\text{nrc}}$ Vs. $H_A : \mu_S^{\text{ctrl}} \neq \mu_S^{\text{nrc}}$, where μ_S^{ctrl} and μ_S^{nrc} are, respectively, the means of the densities in the control and non-ccRCC population groups.

- ccRCC vs. non-ccRCC Tests (RVNR Tests)

1. We represent R^{rcc} by sampling from each pair $(f_{I_{m_i}}^{rcc}, f_{I_{m_j}}^{rcc})$ – in the ccRCC group, the sets of i.i.d rvs $\{I_{m_i,1}^{\text{rcc}}, I_{m_i,2}^{\text{rcc}}, \ldots, I_{m_i,n}^{\text{rcc}}\}$ and $\{I_{m_j,1}^{\text{rcc}}, I_{m_j,2}^{\text{rcc}}, \ldots, I_{m_j,n}^{\text{rcc}}\}$.
2. We represent R^{nrc} by sampling from each pair $(f_{I_{m_i}}^{\text{nrc}}, f_{I_{m_j}}^{\text{nrc}})$ – in the non-ccRCC group, the sets of i.i.d rvs $\{I_{m_{i,1}}^{\text{nrc}}, I_{m_{i,2}}^{\text{nrc}}, \ldots, I_{m_{i,n}}^{\text{nrc}}\}$ and $\{I_{m_{j,1}}^{\text{nrc}}, I_{m_{j,2}}^{\text{nrc}}, \ldots, I_{m_{j,n}}^{\text{nrc}}\}$.
3. Given any region S, common both to R^{rcc} and R^{nrc}, we obtain the local densities $D_S^{\text{rcc}} = \{\text{den}(M_1^{\text{rcc}}), \text{den}(M_2^{\text{rcc}}), \ldots, \text{den}(M_n^{\text{rcc}})\}$ and $D_S^{\text{nrc}} = \{\text{den}(M_1^{\text{nrc}}), \text{den}(M_2^{\text{nrc}}), \ldots, \text{den}(M_n^{\text{nrc}})\}$. Then for each S, we employ these sets (as observations from a density population) together with Eq. 3 (as test statistic) in the following tests: $H_0 : \mu_S^{\text{rcc}} = \mu_S^{\text{nrc}}$ Vs. $H_A : \mu_S^{\text{rcc}} \neq \mu_S^{\text{nrc}}$, where μ_S^{rcc} and μ_S^{nrc} are, respectively, the means of the densities in the ccRCC and non-ccRCC population groups.

We point out that, each of the above class is characterized to have the same alternative conjecture but test statistics related to different parts of the graph. We shall also say that, while evaluating higher performance tests we may also observe in which regions of the spectra there are the best chances of seeing discriminative effects between alternatives.

4 CLINICAL SETTING AND NUMERICAL RESULTS

The above analysis has been applied to samples collected, after informed consent from all subjects participating in the study, at the Ospedale Maggiore Policlinico Foundation (Milano, Italy) using a standardized protocol. As a first step the morning urine midstream (100 mL) was collected in tubes. After centrifugation at 3000 rpm for 10 minutes samples were divided into aliquots. For peptide and protein profiling the eluates from Weak Cation Exchange magnetic beats extraction were automatically spotted onto a Matrix–Assisted Laser Desorption Ionization (MALDI) target plate. All samples were analyzed using an UltraFlex II MALDI-TOF/TOF MS instrument (Bruker Daltonics) and mass spectra were acquired in positive linear mode in the m/z range of 1000-12000. ClinProTools 2.2 software (Bruker Daltonics) was used for all MS data interpretation procedures (Bosso et al., 2008).

4.1 Clinical data

The samples cohort consists of 85 control subjects (58 men, 27 women) and 102 Renal Cell Carcinoma patients (64 men, 38 women). Mean age for controls was 45 with a range of 30–68 years, while for patients 64 with a range of 33–88 years. It was possible to classify pathological group in patients affected by clear cell (ccRCC) and other different histological subtypes (rispectively 79 ccRCC and 23 non-ccRCC). ccRCC samples were classified according to the 2002 TNM (tumor-node-metastasis) system classification.

4.2 Numerical Results

Before discussing the numerical results, it might be useful to remember that the decisions of a statistical test depends on a number of factors; e.g., the sample size, the test statistic, the significance level and the critical value. Moreover, we introduced new parameters which may influence the result as well; i.e., the threshold δ (employed for the template representation) and the neighborhood ray K. We also stress that, in each class CVR, CVNR and RVNR (as defined in section 3.2), tests follow common conjectures (e.g., $\mu^{\text{ctrl}} = \mu^{\text{rcc}}$ and $\mu^{\text{ctrl}} \neq \mu^{\text{rcc}}$) but they use statistics referred to different regions over the spectra. With the above concerns in mind, we summarize the targets of our experiments as follows.

1. For each class of tests, we evaluate (empirically) which threshold δ, and ray K are employed to detect the lowest number of correlation structure changes from control to case groups. In other terms, for different pairs of δ and K we count the number of significant tests rejecting the null hypothesis. For this, we constrain δ to range within a set of higher Pearson's correlation coefficients.
2. By using the values of δ and K obtained above, we detect the mass-to-charge ratio bounds which identify modified regions over the spectra. That is, regions where we have detected a correlation structure modification at a specific level of significance.

Indeed, we first established a fixed number of regions (i.e., 7), a set of arbitrary thresholds $T = \{0.75, 0.76, 0.77, 0.78, 0.79, 0.80\}$ and a set of arbitrary rays $R = [6]$. Then, for each combination of $\delta \in T$ and $K \in R$, we evaluated (for each class of tests) the number of significant tests rejecting the null hypothesis over the spectra. In tab. 1, we report, for each class, both the pair (δ, K) employed to detect the lowest number (i.e., $n = 1$) of tests rejecting the null, and the mass-to-charge ranges which identify the rejection regions at a 5% significance level.

5 CONCLUSIONS

This study showed the possibility to use the extracted peptides to separate healthy subjects from tumor patients and mostly to distinguish non-ccRCC from RCC. Specifically, testing hypotheses on a specific graph property (i.e., density), we derived decision procedures able to provide the clinical modeler with lists of Dalton ranges where it has been detected distinguishing regions. We point out that, from a clinical perspective, in order to apply this approach (for example, to decide the membership group of new subjects), it will be necessary to compute a correlation matrix (whose components are given by Eq. 2) over a set of technical replicates. This will be the most obvious extension for our next work, when new (biological and technical) samples will be available. Moreover, we can summarize, as follow, some further extensions which we are immediately interested to: (I) We need to determine conclusively the identity of the lists of signals in any differentially expressed region. The theoretical framework of section 3 was employed to detect spectral signals for their biological importance (for instance, to suggest potential biomarkers for future analyses) even their identity is not yet ensured. Identification of the peptides/proteins, generating these signals, is a very laborious process implying the analysis of the urine extract with different MS approaches. Therefore, in order to recognize candidate multiple biomarkers, for a specific disease, it's important first to determine their diagnostic "power" and then to investigate better their biological role in the disease mechanisms. (II) The dominant approach to classifier design in clinical studies has been to *minimize* the probability of error – see for example, (Dudoit et al., 2002). Yet it is clear that failing to detect a malignant tumor has drastically different consequences than erroneously flagging a benign tumor. In other words, classification requirements are often *constrained* by a given true positive (*type I error*) and false positive rate (*type II error*) to ensure that the classification error for the most important class is within a desired limit. In order, for our procedures to take into account all of these two requirements, it is necessary to constrain the *type II error*. We point out that, here by constraining only the *type I error* at a standard level of significance, we applied a methodology approach mainly to provide the list of modified regions.

Table 1: Mass-to-Charge regions for Control vs. Case

CVR $\delta = 0.75, K = 2$		CVNR $\delta = 0.75, K = 2$		RVNR $\delta = 0.75, K = 2$	
From	*To*	*From*	*To*	*From*	*To*
1719	2084	1719	2084	4625	5374

ACKNOWLEDGEMENTS

The present work has been supported by grants FIRB n. RBRN07BMCT_011 (Rete Nazionale per lo Studio del Proteoma Umano) of the Italian Ministry of Research.

REFERENCES

Bosso, N., Chinello, C., Picozzi, S., Gianazza, E., Mainini, V., Galbusera, C., Raimondo, F., Perego, R., Casellato, S., Rocco, F., Ferrero, S., Bosari, S., Mocarelli, P., Kienle, M. G., and Magni, F. (2008). Human urine biomarkers of renal cell carcinoma evaluated by clinprot. *Proteomics - Clinical Application*, 2:1036–1046.

Brandes, U. and Erlebach, T., editors (2005). *Network Analysis: Methodological Foundations*, volume 3418 of *Lecture Notes in Computer Science*. Springer.

Brannon, A. and Rathmell, W. (2010). Renal cell carcinoma: where will the state-of-the-art lead us? *Curr. Oncol. Rep.*, (12):193–201.

Drucker, B. (2005). Renal cell carcinoma: current status and future prospects. *Cancer Treat. Rev.*, 31:536–545.

Dudoit, S., Fridlyand, J., and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. of the American Stat. Association*, 97(457):77–87.

Getoor, L. and Taskar, B. (2007). *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Latterich, M., Abramovitz, M., and Leyland-Jones, B. (2008). Proteomics: New technologies and clinical applications. *Eur. Jour. Cancer.*, (44):2737–2741.

Solassol, J., Jacot, W., Lhermitte, L., Boulle, N., Maudelonde, T., and Mang, A. (2006). Clinical proteomics and mass spectrometry profiling for cancer detection. *Expert Rev. Proteomics*, 3(3):311–320.

Characterization of Distinguishing Regions for Renal Cell Carcinoma Discrimination

Citation: Italo Zoppis, Massimiliano Borsani, Erica Gianazza, Clizia Chinello, Giancarlo Albo, Francesco Rocco, Andre M. Deelder, Yuri E. M. van der Burgt, Marco Antoniotti, Fulvio Magni, Giancarlo Mauri. Poster: Characterization of distinguishing regions for Renal Cell Carcinoma discrimination. *IEEE 2nd International Conference on Computational Advances in Bio and medical Sciences*, 2012.
©2012 IEEE. Reprinted, with permission from the authors.

Document Type: Conference Paper (source: Scopus)
ISBN: 978-146731321-6
DOI: 10.1109/ICCABS.2012.6182664

Poster: Characterization of Distinguishing Regions for Renal Cell Carcinoma Discrimination

Italo Zoppis*, Massimiliano Borsani*, Erica Gianazza†, Clizia Chinello†, Giancarlo Albo‡, Francesco Rocco‡
André M. Deelder§, Yuri E. M. van der Burgt§, Marco Antoniotti*, Fulvio Magni† and Giancarlo Mauri*
*Department of Informatics, System and Communication, University of Milano Bicocca, Milano, Italy
†Department of Experimental Medicine, University of Milano Bicocca, Milano, Italy
‡Department of Specialistic Surgical Sciences, "Ospedale Maggiore Policlinico" Foundation, Milano, Italy
§Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands

***Keywords*-proteomics; mass spectrometry; test of hypothesis; clinical data; mutual information, bipartite graphs;**

Mass Spectrometry (MS)-based technologies represent a promising area of research in clinical analysis. They are primarily concerned with measuring the relative intensity (i.e., *signals*) of many protein/peptide molecules associated with their mass-to-charge ratios. These measurements provide a huge amount of information which requires adequate tools to be interpreted. Following the methodology for *testing hypotheses*, we investigate the *proteomic signals* of the most common type of Renal Cell Carcinoma, the *Clear Cell* variant (ccRCC) [1]. By using *mutual information*, we detect changes in dependence values between signals from control to case groups (*ccRCC* or *non–ccRCC*). To this end, we sample and represent each population group through graphs, thus providing the observed *dependence structures* (many real domains are best described by relational models [2]). This way, graphs establish abstract frames of reference in our analysis giving the opportunity to test hypotheses over their properties. In other words, changes are detected by testing graph property modifications from group to group. We report the *mass-to-charge* values which identify *bounded regions* where changes have been detected. The main interest in handling such regions is to perceive which signal ranges are associated with some specific factors of interest (e.g., studying differentially expressed peaks between *cases* and *controls*) and thus, to suggest potential biomarkers for future analysis [3]. This study has been applied to samples collected at the "Ospedale Maggiore Policlinico" Foundation (Milano, Italy) using a standardized protocol. All samples were analyzed using an *UltraFlex II MALDI-TOF/TOF MS* instrument and mass spectra were acquired in the m/z range of 1000-12000. The samples cohort consists of 85 control subjects and 102 Renal Cell Carcinoma patients. It was possible to classify pathological group in patients affected by clear cell (ccRCC) and other different histological subtypes (respectively 79 ccRCC and 23 non-ccRCC). Table I reports the selected rejection regions (i.e., tests reject the null) at the 5% significance level. Testing hypotheses suggested by the data may induce statistical bias. For this reason, we evaluate the results to independent samples. We investigate whether test decisions are statistically independent from the region's property (i.e., distinguishing (DR) or non–distinguishing (ND) regions) when new samples are given. In other words, we want to know whether the property of a region can be statistically associated to test decisions when new samples are available. After that a new sample is provided, we verify test decisions over both the detected distinguishing regions and these regions out of the m/z bounding values previously detected. Table II summarizes the (Fisher's exact test) results confirming a significant association ($\alpha = 0.05$ level) between decisions and region's property for both the class of tests. This work was supported by grants from the Italian Ministry of University and Research (PRIN n. 69373, FIRB n. RBRN07BMCT_011, FAR 2006–2011), EuroKUP COST Action (BM0702) and the NEDD project ("Regione Lombardia").

Table I
REJECTION REGIONS FOR CONTROL VS. CASE TESTS

Control Vs. ccRCC		Control Vs. non–ccRCC		ccRCC Vs. non–ccRCC	
From m/z	*To* m/z	*From* m/z	*To* m/z	*From* m/z	*To* m/z
2644.49	3214.26	1719.45	2084.34	1719.45	2084.34
3270.53	4018.88	4050.39	4540.1	1832.33	2217.2

Table II
NUMBER OF TESTS ACCEPTING H_0 (H_A) VS REGION'S PROPERTY

	CVR		CVNR		RVNR	
Region	H_0	H_A	H_0	H_A	H_0	H_A
DR	2	17	2	17	0	19
ND	11	8	13	6	11	8
p-value	0.005		0.0006		0.0001	

REFERENCES

[1] B. Drucker, "Renal cell carcinoma: current status and future prospects," *Cancer Treat. Rev.*, vol. 31, pp. 536–545, 2005.

[2] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning*. The MIT Press, 2007.

[3] M. Latterich, M. Abramovitz, and B. Leyland-Jones, "Proteomics: New technologies and clinical applications." *Eur. Jour. Cancer.*, no. 44, pp. 2737–2741, 2008.

978-1-4673-1321-6/12/$31.00 ©2012 IEEE

Conference Posters list

2010

- Erica Gianazza, Yuri E.M. van der Burgt, Italo Zoppis, Massimiliano Borsani, Giancarlo Mauri, Clizia Chinello, Valeria Squeo, Gianpaolo Zanetti, Stefano Signorini, André M. Deelder, Marzia Galli Kienle, and Fulvio Magni. Urinary MALDI-TOF profiles of renal cell carcinoma patients obtained by an automated weak-cation exchange magnetic bead purification workflow. *Eurokup Cyprus 2010*. Pissouri (Cyprus)

- M. Antoniotti, M. Borsani, E. Gianazza, F. Magni, G. Mauri, I. Zoppis. Applying Random Graphs for proteomic Data Analysis of RCC Patients. *IBS 2010*. Rimini (Italy). DOI:10.1016/j.jbiotec.2010.09.616

2011

- Erica Gianazza; Yuri E.M. Van Der Burgt; Marco R. Bladergroen; Hans Dalebout; Italo Zoppis; Clizia Chinello; Valeria Squeo; Gianpaolo Zanetti; Giancarlo Mauri; Massimiliano Borsani; Stefano Signorini; Marzia Galli Kienle; Fulvio Magni;

André M. Deelder. Quality control on MALDI-TOF RPC18 profiles of urinary peptides from renal cell carcinoma patients. *ASMS2011 - 59th Conference on Mass Spectrometry and Allied Topics*. Denver, CO (USA)

Part III

Miscellanea

Chapter 6

Supplementary material

6.1 Mass Spectrometry data alignment

Precision and Recall - All labs (MF, MB, FB and MFB)						
k	Precision			Recall		
	MI vs EM	MI vs TT	TT vs EM	MI vs EM	MI vs TT	TT vs EM
2	13.10%	8.50%	5.03%	-5.83%	11.62%	-19.74%
3	6.03%	-1.77%	7.66%	32.53%	29.15%	4.77%
4	16.08%	9.72%	7.04%	12.28%	5.94%	6.74%
5	22.71%	9.70%	14.41%	40.32%	20.74%	24.70%
6	20.42%	13.04%	8.49%	22.35%	17.35%	6.05%
7	20.21%	16.24%	4.74%	19.84%	19.44%	0.50%
8	19.41%	7.43%	12.94%	14.14%	3.67%	10.87%
9	15.48%	10.07%	6.01%	21.71%	6.86%	15.94%
10	17.20%	12.08%	5.82%	6.11%	5.86%	0.27%
11	23.58%	17.48%	7.39%	11.55%	6.55%	5.35%
12	26.48%	22.79%	4.78%	11.71%	14.30%	-3.02%

Table 6.1: Performance comparison (percentage) between Precision and Recall mean values, measured for each method by varying k. Positive values indicates a better performance of the first method versus the second (for example, given A vs. B: A better than B), negative values the opposite (B better than A). Data are drawn in picture 6.1 and 6.1.

Precision and Recall - Labs comparison				
Methods	Monza & Florence	Monza & Brescia	Florence & Brescia	Monza & Florence & Brescia
	Precision			
MI vs EM	-3.45%	27.60%	22.18%	20.37%
MI vs TT	-9.32%	30.34%	3.15%	16.25%
TT vs EM	5.36%	-3.94%	19.65%	4.92%
	Recall			
MI vs EM	17.49%	36.85%	12.97%	3.75%
MI vs TT	9.63%	47.56%	2.81%	-3.60%
TT vs EM	8.70%	-20.43%	10.45%	7.09%

Table 6.2: Performance comparison (percentage) between Precision and Recall mean values, measured for each method by varying the subsets of labs aligned.
Positive values indicates a better performance of the first method versus the second (for example, given A vs. B: A better than B), negative values the opposite (B better than A). Data are drawn in picture 6.1 and 6.1.

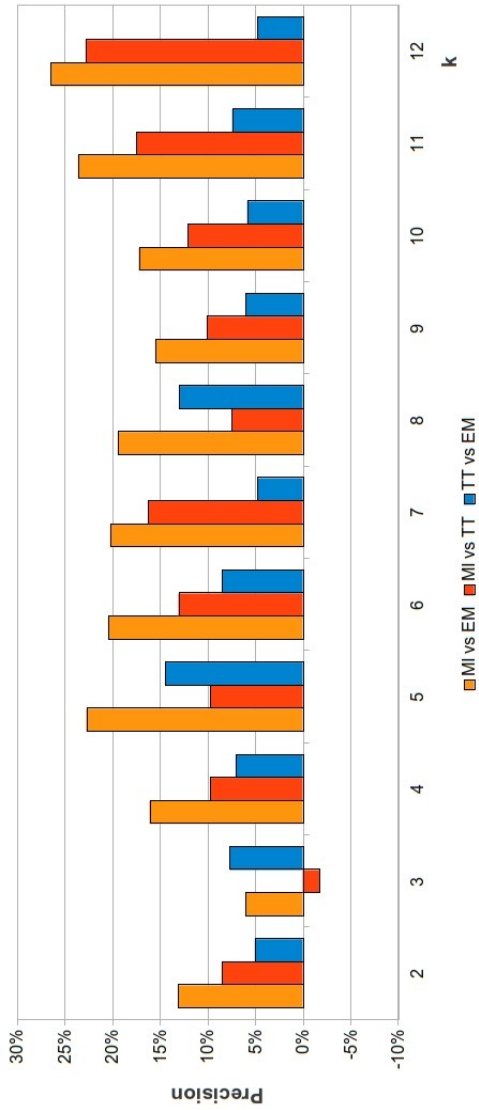


Figure 6.1: Performance comparison (percentage) between Precision mean values, measured for each method by varying k.
MI (light and dark orange) always performed better than the competing methods, with peaks between 20-27%. See also table 6.1.

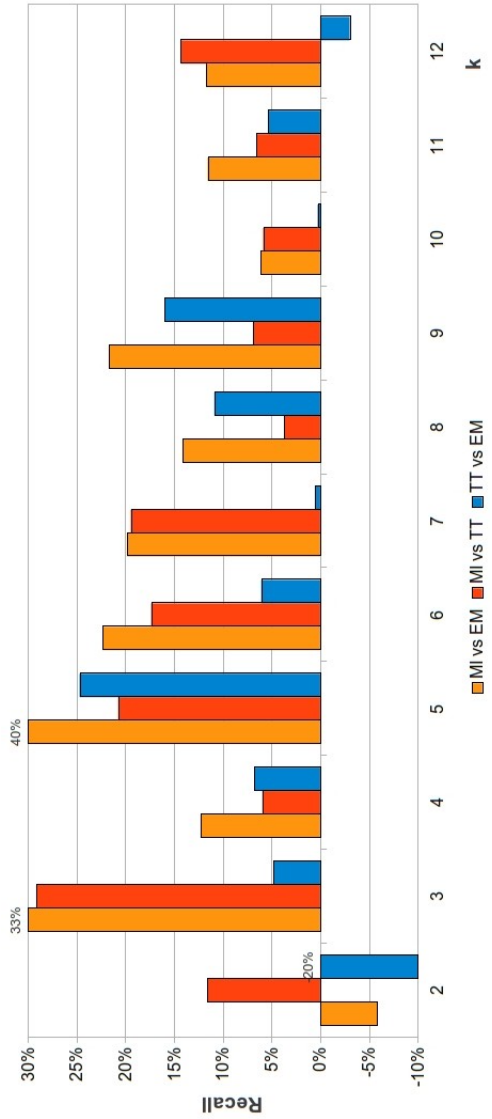


Figure 6.2: Performance comparison (percentage) between Recall mean values, measured for each method by varying k.
MI (light and dark orange) always performed better than the competing methods, with peaks between 20-40%. See also table 6.1.

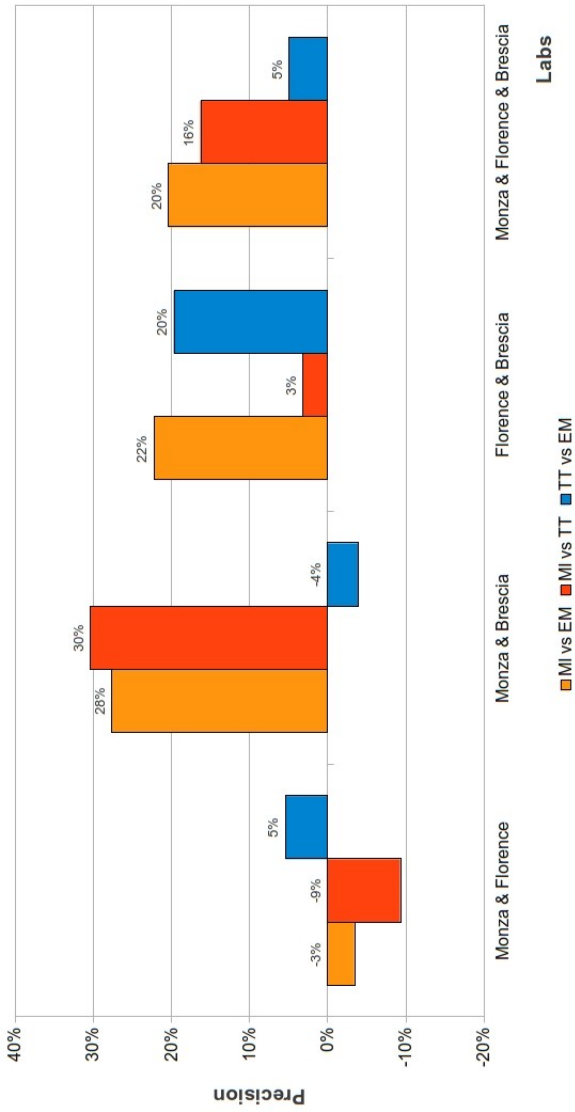


Figure 6.3: Performance comparison (percentage) between Precision mean values, measured for each method by varying the subsets of labs aligned.
MI (light and dark orange) performed better than the competing methods two (MB, FB) on three match (pairs of labs), and, overall, in the three labs alignment (MFB; 15-20 % better). See also table 6.2.

Figure 6.4: Performance comparison (percentage) between Recall mean values, measured for each method by varying the subsets of labs aligned.
MI (light and dark orange) always performed better than the competing methods (pairs of labs), and with slight differences in the three labs alignment (MFB; +4 or -4 %). See also table 6.2.

Chapter 7

Different topics

MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa

Citation: Paolo D'Onorio de Meo, Mattia D'Antonio, Francesca Griggio, Renato Lupi, Massimiliano Borsani, Giulio Pavesi, Tiziana Castrignanò, Graziano Pesole, and Carmela Gissi MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa. Nucl. Acids Res. (2012) 40(D1): D1168-D1172

Document Type: Article (source: Scopus)
DOI: 10.1093/nar/gkr1144
PMID: 22123747

Nucleic Acids Research Advance Access published November 28, 2011

Nucleic Acids Research, 2011, 1–5
doi:10.1093/nar/gkr1144

MitoZoa 2.0: a database resource and search tools for comparative and evolutionary analyses of mitochondrial genomes in Metazoa

Paolo D'Onorio de Meo[1], Mattia D'Antonio[2], Francesca Griggio[3], Renato Lupi[3], Massimiliano Borsani[3], Giulio Pavesi[3], Tiziana Castrignanò[1], Graziano Pesole[2,4] and Carmela Gissi[3,*]

[1]CASPUR, Consorzio interuniversitario per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, [2]Dipartimento di Biochimica e Biologia Molecolare 'E. Quagliariello', Università degli Studi di Bari, Bari, [3]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università degli Studi di Milano, Milan and [4]Istituto di Biomembrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy

Received August 10, 2011; Revised September 26, 2011; Accepted November 9, 2011

ABSTRACT

The MITOchondrial genome database of metaZOAns (MitoZoa) is a public resource for comparative analyses of metazoan mitochondrial genomes (mtDNA) at both the sequence and genomic organizational levels. The main characteristics of the MitoZoa database are the careful revision of mtDNA entry annotations and the possibility of retrieving gene order and non-coding region (NCR) data in appropriate formats. The MitoZoa retrieval system enables basic and complex queries at various taxonomic levels using different search menus. MitoZoa 2.0 has been enhanced in several aspects, including: a re-annotation pipeline to check the correctness of protein-coding gene predictions; a standardized annotation of introns and of precursor ORFs whose functionality is post-transcriptionally recovered by RNA editing or programmed translational frameshifting; updates of taxon-related fields and a BLAST sequence similarity search tool. Database novelties and the definition of standard mtDNA annotation rules, together with the user-friendly retrieval system and the BLAST service, make MitoZoa a valuable resource for comparative and evolutionary analyses as well as a reference database to assist in the annotation of novel mtDNA sequences. MitoZoa is freely accessible at http://www.caspur.it/mitozoa.

INTRODUCTION

The mitochondrial genome (mtDNA) of Metazoa is a major target of studies focused on phylogenetic reconstructions, population genetics and molecular evolution (1). Whole-genome sequencing projects of this relatively small and mostly circular molecule have been undertaken since the development of the Sanger sequencing method (2,3) and have seen an explosive increase with the establishment of next-generation sequencing technologies (4–8). To date, over 4000 entries described as complete mitochondrial genomes are collected in the EMBL nucleotide database (release 108), with about 10 000 additional entries corresponding to human mt genome variants.

The MITOchondrial genome database of metaZOAns (MitoZoa; MZ; http://www.caspur.it/mitozoa) is a unique resource that provides manually curated data on gene annotation, gene order, gene content and non-coding regions (NCR) of complete and nearly-complete ($\geq$7 kb) mtDNA entries of all available metazoan species. One representative entry is present for those metazoan species/subspecies for which the mtDNA has been sequenced in several individuals (9).

Most mtDNA databases focus only on metazoan subgroups. For example, AMiGA collects only arthropod mtDNA sequences (10); MamMiBase focuses on mammals (11); HmtDB and Human mtDB on human (12,13); MitoFish on fishes (http://mitofish.aori.u-tokyo.ac.jp/). Only the no longer updated OGRe (14) and the currently non-functional Mitome (15) databases collected complete mtDNAs of all metazoans. In addition, the NCBI Organelle Genome Resource (16,17) and GOBASE (18) databases contain all mitochondrial and

*To whom correspondence should be addressed. Tel: +39 02 50314918; Fax: +39 02 50314912; Email: carmela.gissi@unimi.it

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at Bib Univ del Dip di Produzione vegetale on November 29, 2011

chloroplastic genomes from all taxonomic groups. However, GOBASE and the Organelle Resource do not attempt to address, or fail in the correction of the large number of misannotations present in mtDNA entries (1,9,14,19). On the contrary, MitoZoa collects sequences from all metazoan species, and systematically identifies and resolves gene misannotations. It also offers several additional types of information and search options absent in other available mtDNA databases (9). Indeed, an associative retrieval system provides a set of tools to carry out basic and complex queries. Thus, MitoZoa users can easily retrieve gene order, NCR sequences, NCR location data, gene/genome sequences, reannotation information and other mito-genomic characteristics, for a given metazoan taxon or for congeneric species.

MitoZoa has already proved to be a useful tool for the scientific community, particularly for studies using mtDNA as a phylogenetic marker (20–23), but also for molecular evolutionary (24,25) and evolutionary ecology analyses (26) including studies on the parallel evolution of minimal mt rRNA secondary structures in metazoans, and on the development of software for environmental metagenomics analyses.

MitoZoa presents several innovative features compared to other mtDNA databases, including a user-friendly retrieval system with one general and three specialized search menus (9). Innovative features of MitoZoa, already described in (9), include:

(1) Extensive controls and correction of gene annotations using a mtDNA-specific re-annotation pipeline.
(2) Standard messages and new entry fields, unambiguously reporting all modifications and data enrichments of the original entries, and making these changes easily searchable by MitoZoa users. The 'MitoZoa Reannotation Summary' (MRS) is one of the main novelties of the EMBL-like MitoZoa entry format.
(3) NCRs of any size are annotated under the new 'NCR' FTkey, thus they can be retrieved with the specialized 'NCR Menu' using several selection criteria.
(4) Gene names are standardized using hidden aliases, thus all sequences of a given gene can be simply retrieved using the 'Gene Content Menu'.
(5) The mtDNA gene order is stored as a string of standardized gene names using a FASTA-like format. Thus, entries sharing a given gene order can be retrieved with the 'Gene Order Menu'.
(6) mtDNAs of congeneric species can be easily selected by the 'General Search Menu', thanks to the creation of the new 'ConGeneric' field.

Several new features have been introduced in MitoZoa 2.0, including: (i) the implementation of a sequence similarity search service by BLAST; (ii) the improvement of the gene re-annotation strategy and of the related pipeline; (iii) the inspection of protein-coding genes; (iv) the systematic and standardized annotation of introns and 'precursor ORFs' post-transcriptionally restored by RNA editing or programmed translational frameshifting (PTF) (27,28); and (v) updating of entries.

NEW FEATURES IN MITOZOA 2.0

BLAST service

The MitoZoa web resource now includes a dedicated BLAST page. The BLAST service allows sequence similarity searches not only against the MitoZoa database (i.e. the full 'mtDNA' sequence of each MitoZoa entry) but also against five additional MitoZoa-derived data sets (Table 1). Each of these additional data sets contains functionally homogeneous mitogenomic 'sub-sequences', such as NCRs or gene categories. Moreover, each sequence of these five additional data sets is described in the header by the entry Accession number, the species name and also the MitoZoa-defined standardized gene name or NCR code (Table 1). These gene names/NCR codes will greatly help the use of BLAST results for annotation of newly produced mt sequences, and for re-annotation of existing mtDNA sequences.

It should be emphasized that all BLAST data sets derived from MitoZoa are automatically updated in concert with MitoZoa. As an example, Table 1 reports the size of the BLAST data sets built from MitoZoa release 9.1. The BLAST service uses the most recent version (2.2.25) of the BLAST+ package (29,30).

Quality checks of protein-coding gene annotation

Unlike the previous MitoZoa reannotation pipeline (9), MitoZoa 2.0 now includes specific checks that verify the correctness of protein-coding gene (CDS) annotations. As a result, possible CDS name errors are fixed and CDS boundaries are also significantly improved.

The quality check pipeline involves both automatic and manual steps, described in detail in Supplementary Data. In particular, examination of CDS multi-alignments allows the detection of two types of CDS inconsistencies resolved in MitoZoa in the following ways:

- Modification of the CDS boundaries: by shifting the annotated start/stop codon, we can recover highly conserved N/C-terminal protein regions identified in the CDS multi-alignment of a given large taxon. Similarly, we can also eliminate extra N/C-terminal protein regions not present in all other multi-aligned CDS. Thus, the encoded protein is accordingly lengthened or shortened.
- Warning message on 'loss of highly conserved aminoacidic regions(s) that can be recovered by frameshift(s)': highly conserved protein region(s) identified in certain multi-alignments are lost in some CDS but can be easily recovered by CDS frameshift(s). Most of such CDS frameshifts are likely due to inaccurate sequencing, as they are located close to sequencing error hot spots (i.e. long homopolymers >8 nt). However, other frameshift cases cannot be easily explained and could represent real losses of functional regions. Thus, we have not modified the boundaries of these CDS but have highlighted them in the MRS

Downloaded from http://nar.oxfordjournals.org/ at Bib Univ del Dip di Produzione vegetale on November 29, 2011

Table 1. Mitochondrial data sets searchable with BLAST, together with the data set size in MitoZoa Release 9.1

Data set name	FTkey used as data set source	Additional data to the sequence header	No. of sequences
mtDNA	Full entry	mtDNA	2894
CDS_nt	CDS	Standard gene name	37 022
tRNA	tRNA	Standard gene name	61 228
rRNA	rRNA	Standard gene name	5699
NCR ≥ 25 nt	NCR ≥ 25 nt	NCR code[a]	8761
Protein	CDS translation, excluding pseudogenes	Standard gene name	37 016

[a]The NCR code defined by MitoZoa relates to species, flanking genes and NCR length (in bp). See also the online MitoZoa Help.

Table 2. Inconsistencies of protein-coding genes (CDS) corrected or pointed out with a warning message in MitoZoa Release 9.1

CDS inconsistency	No. of CDS	No. of entries
Modification of name	2[a]	1[a]
Modification of strand and boundaries	2[b]	1[b]
Modification of boundaries	203	184
Internal stop codons resolved by adding a 'join'[c]	9[d]	8
Unusual start codon resolved by deleting a 'join'[c]	2[e]	2[e]
Warning on 'loss of highly conserved regions'	107	84
MitoZoa Release 9.1	27 022	2894

[a]Exchanged annotation between *atp8* and *atp6* in the snake *Anilius scytale* (FJ755180, v2 EMBL entry).
[b]*atp8* and *nad3* of the gastropod *Platevindex mortoni* (GU475132).
[c]Special cases of the category 'modification of boundaries'. The 'join' operator, defined by GenEMBL, is used to exclude internal positions from CDS or other FTkeys.
[d]In *nad2* of the gastropod *Ilyanassa obsoleta* (NC_007781), the addition of the 'join' operator is also accompanied by modification of the start codon position. In all remaining cases, the CDS boundary modification consists of only the addition of the 'join' operator.
[e]In both cases (DQ340844 and NC_000844), the presence of the join operator was due to the hypothesis of the existence of a four-base start codon in *cox1*, recently rejected by experimental data (32).

('MitoZoa Reannotation Summary') field using a specific warning message (see figure 1 of the online MitoZoa Help). Consequently, MitoZoa users can easily select these CDS, and are warned to pay special attention to the analyses of these CDS and their possible flanking NCRs.

Our CDS quality check strategy identified a total of 207 CDSs that need 'modifications of name/boundaries', and 107 CDS that invoke a warning on the 'loss of highly conserved aminoacidic regions' (Table 2). We emphasize that most CDS modifications and warning notes cause the disappearance of flanking NCRs or gene overlaps. In addition, 4 CDS errors have effects on the determination of gene order ('gene name' and 'gene strand' modifications in Table 2). Finally, 9 CDSs were likely incorrect because they showed multiple internal stop codons (Table 2). Therefore, the CDS re-annotation process has significant consequences on the CDSs themselves (and their use in phylogenetic reconstruction), the determination of flanking NCRs, and even on the overall gene order.

As a final point, we would emphasize that CDS re-annotation has required the definition of specific criteria for mt CDS determination based on the peculiarities of the mt transcriptional and maturation processes (31–33). These criteria can be also regarded as tentative rules for the standardization of mt CDS annotation and are detailed in the Supplementary Data.

Standardized annotation of introns and frameshifts

Group I and II self-splicing introns as well as frameshift sites post-transcriptionally resolved by RNA editing or programmed translational frameshifting (PTF) (27,28,34) occur in some protein-coding genes of few metazoan taxa. However, original entries often contain non-standard annotations of these phenomena, rendering automated parsing difficult. In MitoZoa 2.0, we have implemented a specific pipeline, detailed in the Supplementary Data, to identify and standardize such annotations.

These CDS peculiarities are now clearly recorded in the MRS field with appropriate standardized messages (see figure 1 of the Online MitoZoa Help), thus they can be easily retrieved by MitoZoa users. Moreover, we have created a new FTkey 'prec_ORF' in order to annotate all 'precursor ORFs' with frameshift site(s) corrected by RNA editing or PTF. This new FTkey allows the automatic retrieval and analysis of these 'precursor ORF' sequences. As discussed in the Supplementary Data, we have used the 'prec_ORF' annotation to study the reliability of the currently hypothesised RNA editing/PTF cases. Thus, we are confident that this MitoZoa novelty will help the correct annotation of future cases of RNA editing/PTF.

In the current MitoZoa release, we have identified and annotated 40 CDS with introns and 198 CDS with frameshift sites (see Supplementary Tables S1–S3).

MitoZoa format novelties

For each MitoZoa entry, the gene order is reported in a FASTA-like format as a string of standardized gene names (9). In MitoZoa 2.0, the gene order format has been improved adding to the header a token that indicates the linear topology (L) or the partial status (P) of the entry. This novelty helps to identify linear and partial mtDNAs from the inspection of gene order header. It can be advantageous to users interested in extensive analyses of the gene order in large taxonomic groups.

MitoZoa entry updates

Pre-existing MZ entries are now updated at each new MZ release. This update is essential to allow reliable entry

Downloaded from http://nar.oxfordjournals.org/ at Bib Univ del Dip di Produzione vegetale on November 29, 2011

selections with the Taxonomy, the Organism Species (OS) and the ConGeneric (CG) fields of the 'General Search Menu'.

In particular, the update of the Taxonomy field is indispensable because it comes from the Taxonomy database (http://www.ncbi.nlm.nih.gov/taxonomy), where even high taxonomic levels are frequently reorganized by NCBI curators. Furthermore, the OS field of existing entries are sometimes modified by the authors of entries owing to revised taxonomic assignment of the biological sample used for sequence production. Specific standardized messages are added to the MRS field to track these changes and allow easily retrieval (see figure 1 of the online MitoZoa Help).

As an example of the extent of MZ entry update, the migration of the 2633 pre-existing entries from MitoZoa Rel. 7 to Rel. 8 involved changes of 300 entries (11.4%) in the OC field, and 65 entries (2.5%) in the OS field (plus OC, if necessary).

Miscellanea

The MZ re-annotation pipeline includes some completely manual steps involving literature check, evaluation of unusual mtDNA characteristics, and *de novo* annotation of interesting entries. All these steps depend on curator expertise and are time-consuming. Thus, we have set up specific file formats and scripts to assist curators. Some examples of manually revised entries are reported in Supplementary Table S4.

The previous MitoZoa list of the mt genetic codes has been updated adding a new genetic code absent in the translation table list compiled by the NCBI (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). This code, named '5bis', has been recently identified in the nematode *Radopholus similis* by Jacob *et al.* (35).

SUMMARY AND FUTURE DIRECTIONS

MitoZoa provides carefully revised annotations of all mt gene categories, thus it ensures high accuracy of gene sequences, NCRs and gene order data extracted from MitoZoa. Moreover, all corrections and improvements of the entries are indicated by standardized messages (mainly located in the MRS field), further assisting MitoZoa users in the analysis of the revised elements.

The Mitozoa retrieval system permits the easy selection both of highly studied mt protein-coding genes and some often overlooked mt features such as NCR sequences and gene order, even for large taxonomic data sets. Among these features, NCR sequences and gene order data are difficult or impossible to retrieve from other mt databases. Indeed, MitoZoa permits flexible queries not feasible by any other system. For example, the selection of the teleost L-strand replication origin sequences can be achieved through the 'NCR Menu' searching for all NCRs longer than 20 bp, located between *trnN* and *trnC*, and belonging to the taxon Teleostei. Likewise, all metazoan mtDNAs having the mammalian-distinctive 'WANCY' region can be simply extracted through the 'Gene Order Menu' searching for entries having the '*trnW -trnA -trnN -trnC -trnY*' gene string.

We believe that both the correction of annotation inconsistencies and the user-friendly retrieval system makes Mitozoa a valuable resource for researchers interested in phylogenetic reconstructions and also in peculiar aspects of mtDNA evolution. MitoZoa could also direct the mitochondrial community to new investigations, thanks to the emphasis on taxa/genes characterized by problematic annotations or unusual features. Finally, the implementation of the BLAST sequence similarity search could make MitoZoa a reference database for the annotation of novel mt genomes, and the definition of widely shared mt annotation rules whose requirement has been often invoked in the past (19). Indeed, as stressed in the section on CDS quality check, the correction of gene boundaries requires the definition of general annotation rules based on the knowledge of the mt transcription and translation processes.

In the future, we plan to develop new tools for the examination of gene order and to implement services for the analyses of retrieved sequences (programs for sequence multi-alignment, prediction of secondary structures, etc). Suggestions from MitoZoa users on new options for data visualization and extraction will be also taken into account.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables S1–S4.

FUNDING

Ministero dell'Istruzione, dell'Università e della Ricerca, Italy: Programma di Ricerca Scientifica di Rilevante Interesse Nazionale 2009; Progetto DM19410. Università degli Studi di Milano, Italy: Programma dell'Università per la Ricerca FIRST 2007. Consiglio Nazionale delle Ricerche, Italy: LifeWatch. Funding for open access charge: Università degli Studi di Milano, Italy: Programma dell'Università per la Ricerca (to C.G.).

Conflict of interest statement. None declared.

REFERENCES

1. Gissi,C., Iannelli,F. and Pesole,G. (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity*, **101**, 301–320.
2. Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.
3. Bibb,M.J., Van Etten,R.A., Wright,C.T., Walberg,M.W. and Clayton,D.A. (1981) Sequence and gene organization of mouse mitochondrial DNA. *Cell.*, **26**, 167–180.
4. Jex,A.R., Hall,R.S., Littlewood,D.T. and Gasser,R.B. (2010) An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Res.*, **38**, 522–533.
5. Jex,A.R., Littlewood,D.T. and Gasser,R.B. (2010) Toward next-generation sequencing of mitochondrial genomes—focus

Downloaded from http://nar.oxfordjournals.org/ at Bib Univ del Dip di Produzione vegetale on November 29, 2011

on parasitic worms of animals and biotechnological implications. *Biotechnol Adv*., **28**, 151–159.
6. McComish,B.J., Hills,S.F., Biggs,P.J. and Penny,D. (2010) Index-free de novo assembly and deconvolution of mixed mitochondrial genomes. *Genome Biol. Evol*., **2**, 410–424.
7. Morin,P.A., Archer,F.I., Foote,A.D., Vilstrup,J., Allen,E.E., Wade,P., Durban,J., Parsons,K., Pitman,R., Li,L. *et al.* (2010) Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Res*., **20**, 908–916.
8. Timmermans,M.J., Dodsworth,S., Culverwell,C.L., Bocak,L., Ahrens,D., Littlewood,D.T., Pons,J. and Vogler,A.P. (2010) Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. *Nucleic Acids Res*., **38**, e197.
9. Lupi,R., D'Onorio de Meo,P., Picardi,E., D'Antonio,M., Paoletti,D., Castrignanò,T., Pesole,G. and Gissi,C. (2010) MitoZoa: a curated mitochondrial genome database of metazoans for comparative genomics studies. *Mitochondrion*, **10**, 192–199.
10. Feijao,P.C., Neiva,L.S., de Azeredo-Espin,A.M. and Lessinger,A.C. (2006) AMiGA: the arthropodan mitochondrial genomes accessible database. *Bioinformatics*., **22**, 902–903.
11. Vasconcelos,A.T., Guimaraes,A.C., Castelletti,C.H., Caruso,C.S., Ribeiro,C., Yokaichiya,F., Armoa,G.R., Pereira Gda,S., da Silva,I.T., Schrago,C.G. *et al*. (2005) MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies. *Bioinformatics*., **21**, 2566–2567.
12. Attimonelli,M., Accetturo,M., Santamaria,M., Lascaro,D., Scioscia,G., Pappada,G., Russo,L., Zanchetta,L. and Tommaseo-Ponzetta,M. (2005) HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. *BMC Bioinformatics*., **6**, S4.
13. Ingman,M. and Gyllensten,U. (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. *Nucleic Acids Res*., **34**, D749–751.
14. Jameson,D., Gibson,A.P., Hudelot,C. and Higgs,P.G. (2003) OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucleic Acids Res*., **31**, 202–206.
15. Lee,Y.S., Oh,J., Kim,Y.U., Kim,N., Yang,S. and Hwang,U.W. (2008) Mitome: dynamic and interactive database for comparative mitochondrial genomics in metazoan animals. *Nucleic Acids Res*., **36**, D938–D942.
16. Wolfsberg,T.G., Schafer,S., Tatusov,R.L. and Tatusov,T.A. (2001) Organelle genome resource at NCBI. *Trends Biochem Sci*., **26**, 199–203.
17. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*., **37**, D32–D36.
18. O'Brien,E.A., Zhang,Y., Wang,E., Marie,V., Badejoko,W., Lang,B.F. and Burger,G. (2009) GOBASE: an organelle genome database. *Nucleic Acids Res*., **37**, D946–D950.
19. Boore,J. (2006) Requirements and standards for organelle genome databases. *OMICS*, **10**, 119–126.
20. Irisarri,I., San Mauro,D., Green,D.M. and Zardoya,R. (2010) The complete mitochondrial genome of the relict frog *Leiopelma archeyi*: insights into the root of the frog Tree of Life. *Mitochondrial DNA*, **21**, 173–182.
21. Irisarri,I., Vences,M., San Mauro,D., Glaw,F. and Zardoya,R. (2011) Reversal to air-driven sound production revealed by a molecular phylogeny of tongueless frogs, family Pipidae. *BMC Evolutionary Biol*., **11**, 114.
22. Jia,W.Z., Yan,H.B., Guo,A.J., Zhu,X.Q., Wang,Y.C., Shi,W.G., Chen,H.T., Zhan,F., Zhang,S.H., Fu,B.Q. *et al.* (2011) Complete mitochondrial genomes of Taenia multiceps, T. hydatigena and T. pisiformis: additional molecular markers for a tapeworm genus of human and animal health significance. *BMC Genomics*, **11**, 447.
23. Rawlings,T.A., MacInnis,M.J., Bieler,R., Boore,J.L. and Collins,T.M. (2010) Sessile snails, dynamic genomes: gene rearrangements within the mitochondrial genome of a family of caenogastropod molluscs. *BMC Genomics*, **11**, 440.
24. Castellana,S., Vicario,S. and Saccone,C. (2011) Evolutionary patterns of the mitochondrial genome in Metazoa: exploring the role of mutation and selection in mitochondrial protein coding genes. *Genome Biol. Evol*., **3**, 1067–1079.
25. Klimov,P.B. and Knowles,L.L. (2011) Repeated parallel evolution of minimal rRNAs revealed from detailed comparative analysis. *J. Heredity*, **102**, 283–293.
26. Bengtsson,J., Eriksson,K.M., Hartmann,M., Wang,Z., Shenoy,B.D., Grelet,G.A., Abarenkov,K., Petri,A., Alm Rosenblad,M. and Nilsson,R.H. (2011) Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie Van Leeuwenhoek*, **100**, 471–475.
27. Russell,R.D. and Beckenbach,A.T. (2008) Recoding of translation in turtle mitochondrial genomes: programmed frameshift mutations and evidence of a modified genetic code. *J. Mol. Evol*., **67**, 682–695.
28. Rosengarten,R.D., Sperling,E.A., Moreno,M.A., Leys,S.P. and Dellaporta,S.L. (2008) The mitochondrial genome of the hexactinellid sponge *Aphrocallistes vastus*: evidence for programmed translational frameshifting. *BMC Genomics*., **9**, 33.
29. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
30. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*., **10**, 421.
31. Montoya,J., Lopez-Perez,M.J. and Ruiz-Pesini,E. (2006) Mitochondrial DNA transcription and diseases: past, present and future. *Biochim Biophys Acta*., **1757**, 1179–1189.
32. Stewart,J.B. and Beckenbach,A.T. (2009) Characterization of mature mitochondrial transcripts in *Drosophila*, and the implications for the tRNA punctuation model in arthropods. *Gene*., **445**, 49–57.
33. Temperley,R.J., Wydro,M., Lightowlers,R.N. and Chrzanowska-Lightowlers,Z.M. (2010) Human mitochondrial mRNAs–like members of all families, similar but different. *Biochim Biophys Acta*., **1797**, 1081–1085.
34. Mindell,D.P., Sorenson,M.D. and Dimcheff,D.E. (1998) An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. *Mol. Biol. Evol*., **15**, 1568–1571.
35. Jacob,J.E., Vanholme,B., Van Leeuwen,T. and Gheysen,G. (2009) A unique genetic code change in the mitochondrial genome of the parasitic nematode *Radopholus similis*. *BMC Res Notes*., **2**, 192.

Downloaded from http://nar.oxfordjournals.org/ at Bib Univ del Dip di Produzione vegetale on November 29, 2011

Comparative Profiling of Pseudomonas aeruginosa Strains Reveals Differential Expression of Novel Unique and Conserved Small RNAs (acknowledgements)

S. Ferrara, M. Brugnoli, A. De Bonis, F. Righetti, F. Delvillani, G. Dehò, D. Horner, F. Briani, G. Bertoni (2012). Comparative profiling of Pseudomonas aeruginosa strains reveals differential expression of novel unique and conserved small RNAs. PLoS ONE, 7, 5:e36553.

PMID: 22590564

Acknowledgments

The authors acknowledge all members of the lab for helpful discussion and technical support, F. Dal Pero and A. Albiero for 454-pyrosequencing and data analysis, respectively, and G. Pavesi and M. Borsani for L1 oligonucleotide design.

PhD school notes on PhD thesis format

NOTES ON THE PhD THESIS FORMAT

We strongly advise the following format for the PhD thesis:

Booklet size: B5 (17.6 x 25 cm).
Cover: see attached example for outside cover.
Contents: (for more details see attached outline) divided into:

- Inside cover
- Contents (Index)
- Summary: **maximum** one page (it must be a summary!)
- Part I (Introductory, see below): about 50 pages (or more, if necessary), font size 12, 1.5 line spacing.
- Part II (publications, submitted manuscripts): as necessary.
- Part III: Supplementary materials (optional)

Printed double-sided,
Binding: paperback (not hard cover).
Language: English (but see the cover templates: some terms are in Italian).
It is usual to include credits, in addition to the acknowledgments present in the published papers. This is obviously allowed, but remember that the thesis is an official document and a sobering tone is recommended.
The attached example for outside cover (splayed out) has to be scaled up for printing. The format of the attached thesis outline is A4 and will be reduced upon printing.

Remember that **the core of your thesis is Part II**. Published papers should be included as **pdf copy of the printed published version**, and may include published supplementary material, if appropriate.
Part III may include manuscripts in preparation, data not yet organized in the form of a manuscript, **side researches** not included in the main frame of your thesis, bulky data set not reported in Part II.
Part I should be a general introduction and presentation of your thesis work. **Part I should provide the background and rationale of your work, and link together the material presented in parts II and III**.

For any additional information and to inspect recommended examples of theses, please contact Margherita Russo (margherita.russo@unimi.it).

updated on September 19, 2012

Page numbering starts here

Part I (about 50 pages or as necessary)

Abstract (**max 1 page**)
State of the Art
Aim of the Project
Main Results
Conclusions and Future Prospects
References
(Acknowledgement)

Part II

Content (list of papers included, with indication of "Published in: …", or "Submitted to…").
Published paper_1
Published paper_2
…
Submitted Manuscript_1
Submitted Manuscript_2
…

Part III (if needed)

Supplementary data, manuscripts in preparation, additional tables and figures, etc

"Io stimo più il trovar un vero,
benché di cosa leggiera,
ché il disputar lungamente
delle massime questioni
senza conseguir verità nissuna"

Galileo Galilei

["I appreciate more finding an effective truth,
though simplier it might be,
than long debating
on lofty topics
without reaching any conclusion."]

www.unimi.it http://users.unimi.it/dottscbiolemol/