ANALYSIS OF CORRELATION STRUCTURES IN RENAL CELL CARCINOMA PATIENT DATA

Italo Zoppis[1], Massimiliano Borsani[1], Erica Gianazza[2], Clizia Chinello[2], Francesco Rocco[4], Giancarlo Albo[4], André M. Deelder[3], Yuri E. M. van der Burgt[3], Marco Antoniotti[1], 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, Leiden University Medical Center, Leiden, The Netherlands*

[4]*Department of Specialistic Surgical Sciences, "Ospedale Maggiore Policlinico" Foundation, Milano, Italy*

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 – a *protein/peptide profile* (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 pe-

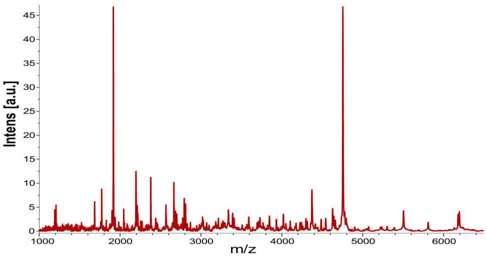

Figure 1: A typical *protein/peptide profile*.

Zoppis I., Borsani M., Gianazza E., Chinello C., Rocco F., Albo G., M. Deelder A., E. M. van der Burgt Y., Antoniotti M., Magni F. and Mauri G..
ANALYSIS OF CORRELATION STRUCTURES IN RENAL CELL CARCINOMA PATIENT DATA.
DOI: 10.5220/0003856702510256
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 251-256
ISBN: 978-989-8425-90-4

rceive 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 *constrained classification* 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 $\tilde{V}_{i,k}=\{v_{i-k},\dots,v_i,\dots,v_{i+k}\}$ and $\tilde{V}_{j,k}=\{v_{j-k},\dots,v_j,\dots,v_{j+k}\}$*[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*

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

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 $I^{\text{ctrl}}_{m_1}\times I^{\text{ctrl}}_{m_2}\times\dots\times I^{\text{ctrl}}_{m_n}$ 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 $\{I^g_{m_{i,1}},I^g_{m_{i,2}},\dots,I^g_{m_{i,n}}\}$ and $\{I^g_{m_{j,1}},I^g_{m_{j,2}},\dots,I^g_{m_{j,n}}\}$, we call* template *(of g) the bipartite graph $R^g=(V_1\cup V_2,E)$ with vertexes $V_1=\{m_1,m_2,\dots,m_n\}$ and $V_2=\{m^{'}_1,m^{'}_2,\dots,m^{'}_n\}$. Moreover, $(m_i,m^{'}_j)\in E$ only if the absolute value of the Pearson's correlation coefficient exceeds a threshold δ. That is,*

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

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

[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 $\{1,\dots,n\}$ of the first *n* positive integers.

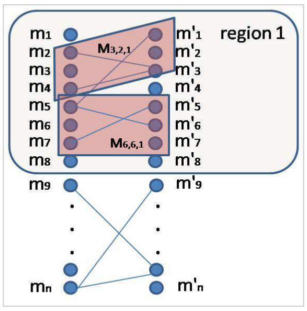

Figure 2: The bipartite graph for RCC data (template) with one region and two neighborhoods.

densities $\{d_1, d_2, \ldots, d_n\}$ by observing a set of neighborhoods in S. For example, in Fig. 2 is reported 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 $D_S^{\mathrm{g}} = \{\mathrm{den}(M_1^g), \mathrm{den}(M_2^g)\}$. 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 $\mu^{\mathrm{ctrl}} \neq \mu^{\mathrm{case}}$, 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 $D^{\mathrm{ctrl}} = \{X_1, X_2, ..., X_n\}$ from controls, and $D^{\mathrm{case}} = \{Y_1, Y_2, ..., Y_n\}$ from cases, (ii) their differences $D = \{D_i : D_i = X_i - Y_i, X_i \in D^{\mathrm{ctrl}}, Y_i \in D^{\mathrm{case}}\}$, (iii) the sample mean $\tilde{D}$ and (iv) the sample standard deviation of difference scores S_d, we can reject the *null* $H_0 : \mu^{\mathrm{ctrl}} = \mu^{\mathrm{case}}$ (no change) in favor of the *alternative* $H_A : \mu^{\mathrm{ctrl}} \neq \mu^{\mathrm{case}}$ 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 $H_0 : \mu_S^{\mathrm{ctrl}} = \mu_S^{\mathrm{case}}$ against $H_A : \mu_S^{\mathrm{ctrl}} \neq \mu_S^{\mathrm{case}}$ 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 through the following procedures:

- **Control vs. ccRCC Tests (Noted as CVR Tests).**
  1. We represent R^{ctrl} by sampling from each pair $(f_{I_{m_i}}^{ctrl}, f_{I_{m_j}}^{ctrl})$ – in the control group, the sets of i.i.d rvs $\{I_{m_i,1}^{\mathrm{ctrl}}, I_{m_i,2}^{\mathrm{ctrl}}, \ldots, I_{m_i,n}^{\mathrm{ctrl}}\}$ and $\{I_{m_j,1}^{\mathrm{ctrl}}, I_{m_j,2}^{\mathrm{ctrl}}, \ldots, I_{m_j,n}^{\mathrm{ctrl}}\}$.
  2. We represent R^{rcc} by sampling from each pair $(f_{I_{m_i}}^{\mathrm{rcc}}, f_{I_{m_j}}^{\mathrm{rcc}})$ – in the ccRCC group, the sets of i.i.d rvs $\{I_{m_i,1}^{\mathrm{rcc}}, I_{m_i,2}^{\mathrm{rcc}}, \ldots, I_{m_i,n}^{\mathrm{rcc}}\}$ and $\{I_{m_j,1}^{\mathrm{rcc}}, I_{m_j,2}^{\mathrm{rcc}}, \ldots, I_{m_j,n}^{\mathrm{rcc}}\}$.
  3. Given any region S, common both to R^{ctrl} and R^{rcc}, we obtain the local densities $D_S^{\mathrm{ctrl}} = \{\mathrm{den}(M_1^{\mathrm{ctrl}}), \mathrm{den}(M_2^{\mathrm{ctrl}}), \ldots, \mathrm{den}(M_n^{\mathrm{ctrl}})\}$ and $D_S^{\mathrm{rcc}} = \{\mathrm{den}(M_1^{\mathrm{rcc}}), \mathrm{den}(M_2^{\mathrm{rcc}}), \ldots, \mathrm{den}(M_n^{\mathrm{rcc}})\}$. 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: $H_0 : \mu_S^{\mathrm{ctrl}} = \mu_S^{\mathrm{rcc}}$ Vs. $H_A : \mu_S^{\mathrm{ctrl}} \neq \mu_S^{\mathrm{rcc}}$, 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 $(f_{I_{m_i}}^{ctrl}, f_{I_{m_j}}^{ctrl})$ – in the control group, the sets of i.i.d rvs $\{I_{m_i,1}^{\mathrm{ctrl}}, I_{m_i,2}^{\mathrm{ctrl}}, \ldots, I_{m_i,n}^{\mathrm{ctrl}}\}$ and $\{I_{m_j,1}^{\mathrm{ctrl}}, I_{m_j,2}^{\mathrm{ctrl}}, \ldots, I_{m_j,n}^{\mathrm{ctrl}}\}$.
  2. We represent R^{nrc} by sampling from each pair $(f_{I_{m_i}}^{\mathrm{nrc}}, f_{I_{m_j}}^{\mathrm{nrc}})$ – in the non-ccRCC group, the sets of i.i.d rvs $\{I_{m_{i,1}}^{\mathrm{nrc}}, I_{m_{i,2}}^{\mathrm{nrc}}, \ldots, I_{m_{i,n}}^{\mathrm{nrc}}\}$ and $\{I_{m_{j,1}}^{\mathrm{nrc}}, I_{m_{j,2}}^{\mathrm{nrc}}, \ldots, I_{m_{j,n}}^{\mathrm{nrc}}\}$.

[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 $D_S^{\text{ctrl}} = \{\text{den}(M_1^{\text{ctrl}}), \text{den}(M_2^{\text{ctrl}}), \dots, \text{den}(M_n^{\text{ctrl}})\}$ and $D_S^{\text{nrc}} = \{\text{den}(M_1^{\text{nrc}}), \text{den}(M_2^{\text{nrc}}), \dots, \text{den}(M_n^{\text{nrc}})\}$. 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 $(f^{rcc}_{I_{m_i}}, f^{rcc}_{I_{m_j}})$ – in the ccRCC group, the sets of i.i.d rvs $\{I^{\text{rcc}}_{m_i,1}, I^{\text{rcc}}_{m_i,2}, \dots, I^{\text{rcc}}_{m_i,n}\}$ and $\{I^{\text{rcc}}_{m_j,1}, I^{\text{rcc}}_{m_j,2}, \dots, I^{\text{rcc}}_{m_j,n}\}$.
2. We represent R^{nrc} by sampling from each pair $(f^{\text{nrc}}_{I_{m_i}}, f^{\text{nrc}}_{I_{m_j}})$ – in the non-ccRCC group, the sets of i.i.d rvs $\{I^{\text{nrc}}_{m_{i,1}}, I^{\text{nrc}}_{m_{i,2}}, \dots, I^{\text{nrc}}_{m_{i,n}}\}$ and $\{I^{\text{nrc}}_{m_{j,1}}, I^{\text{nrc}}_{m_{j,2}}, \dots, I^{\text{nrc}}_{m_{j,n}}\}$.
3. Given any region S, common both to R^{rcc} and R^{nrc}, we obtain the local densities $D_S^{\text{rcc}} = \{\text{den}(M_1^{\text{rcc}}), \text{den}(M_2^{\text{rcc}}), \dots, \text{den}(M_n^{\text{rcc}})\}$ and $D_S^{\text{nrc}} = \{\text{den}(M_1^{\text{nrc}}), \text{den}(M_2^{\text{nrc}}), \dots, \text{den}(M_n^{\text{nrc}})\}$. 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 (respectively 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. By 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*, we applied a methodology approach mainly to provide the list of modified regions.

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

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

ACKNOWLEDGEMENTS

This work was supported by grants of the Italian Ministry of Research: PRIN 2006, FIRB 2007 (RBRN07BMCT_11), FAR 2006–2011, and in part by both the EuroKUP COST Action (BM0702) and the NEDD project of the "Regione Lombardia".

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 - Clin. App.*, 2:1036–1046.

Brandes, U. and Erlebach, T., editors (2005). *Network Analysis: Methodological Foundations*, volume 3418 of *Lect. 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. Assoc.*, 97(457):77–87.

Getoor, L. and Taskar, B. (2007). *Introduction to Statistical Relational 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.