Procedia Computer Science

Volume 51, 2015, Pages 683–692

ICCS 2015 International Conference On Computational Science

Robust Conclusions in Mass Spectrometry Analysis

Italo Zoppis[1], Riccardo Dondi[3], Massimiliano Borsani[1], Erica Gianazza[2], Clizia Chinello[2], Fulvio Magni[2], and Giancarlo Mauri[1]

[1] Department of Informatics, University of Milano Bicocca, Milano, Italy.
[2] Department of Experimental Medicine, University of Milano Bicocca, Milano, Italy.
[3] Human and Social Sciences Department, University of Bergamo, Bergamo - Italy.

Abstract

A central issue in biological data analysis is that uncertainty, resulting from different factors of variability, may change the effect of the events being investigated. Therefore, *robustness* is a fundamental step to be considered. Robustness refers to the ability of a process to cope well with uncertainties, but the different ways to model both the processes and the uncertainties lead to many alternative conclusions in the robustness analysis.

In this paper we apply a framework allowing to deal with such questions for mass spectrometry data. Specifically, we provide robust decisions when testing hypothesis over a case/control population of subject measurements (i.e. proteomic profiles). To this concern, we formulate (i) a reference model for the observed data (i.e., graphs), (ii) a reference method to provide decisions (i.e., test of hypotheses over graph properties) and (iii) a reference model of variability to employ sources of uncertainties (i.e., random graphs). We apply these models to a real-case study, analyzing the mass spectrometry *profiles* of the most common type of Renal Cell Carcinoma; the *Clear Cell* variant.

Keywords: Data analysis, inference, robust decisions, graph, mass spectrometry.

1 Introduction

Uncertainty characterizes many experimental processes and may change the effects of the events being investigated. For this reason, *robustness analysis* needs to be considered in an appropriate manner[1]. Even though, in its general form, the word *robustness* refers to the ability of a process to cope well with uncertainties, the different ways in which both the process and the uncertainty are modeled may lead to many alternative definitions of the word itself. Hampel [9] defines the word *robustness* within a general *statistical context*. That definition can be summarized by considering the *robustness* as *the stability theory of statistical procedures*. It systematically investigates the effects of deviations from modeling assumptions on known procedures and, if necessary, develops new, better procedures. Recently, the relationship between robustness and

[1]For an extended introduction to robustness analysis, see for example [19, 16].

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2015

doi:10.1016/j.procs.2015.05.185

multiple criteria decision analysis has been observed by a number of researchers [18, 15]. For instance Kouvelis and Yu [13] studied the robustness in the context of discrete optimization. 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.

In this paper we deal with such questions for mass spectrometry data. In this context data produced by mass spectrometers are affected by errors and noise due to sample preparation and instrument approximation, therefore data pre-processing and robust data analysis are fundamentals. Here, we provide robust decisions when testing hypothesis over a case/control population of subject measurements i.e. proteomic profiles. For our purposes protein/peptide profiles can be thought of as collections of *peaks*, where each peak identifies the pair of values given by the intensity (related to the abundance) of a molecule with its specific molecular mass-to-charge ratio. To this concern, we properly formulate (i) a reference model of the observed data (i.e., graphs), (ii) a reference method to provide decisions (i.e., test of hypotheses over graph properties) and (iii) a reference model of variability to employ sources of uncertainties (i.e., random graphs). Throughout, we apply a real-case study by analyzing the mass spectrometry profiles of the most common type of Renal Cell Carcinoma: the *Clear Cell* variant (ccRCC) [2]. In Sec. 2 we introduce basic definitions and preliminary notations. In Sec. 3 we formulate the reference model for the observed data, the reference method to provide decisions and the reference model of variability. Sec. 4 is specifically dedicated to our case study (i.e. RCC analysis). In Sec. 5 we report results and numerical details of the statistical procedures applied to the RCC data. Finally we conclude the work in Sec. 6 by providing discussions for future analysis.

2 Basic Definitions

Throughout this paper $G = (V_1 \cup V_2, E)$ denotes a *bipartite graph*. V_1 and V_2 are two *totally ordered* sets of *vertices*[3] such that the set of *edges* $E \subseteq V_1 \times V_2$ connect vertices in one set with vertices in the other: i.e., E is a set of pairs (v_i, v_j) with $v_i \in V_1$ and $v_j \in V_2$. Given a *bipartite graph* $G = (V_1 \cup V_2, E)$, the *sub-graph* 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 $(v_1, v_2) \in \tilde{E}$ [4]. The number of vertices $N_v = |V_1 \cup V_2|$ and the number of edges $N_e = |E|$ are generally called the *order* and the *size* of the graph. Graphs can be generally "summarized" in a compact way by various *graph properties*. Among all the properties in the literature [4], here we focus on *cohesion*. A well known index to characterize this notion is that of *density*. Before introducing formally this concept we give the following definition.

Definition 1 (Neighborhood)**.** *Let $G = (V_1 \cup V_2, E)$ be a bipartite graph. 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 sub-graphs of G induced by $\tilde{A} = \tilde{V}_{i,k} \cup \tilde{V}_{j,k}$ where $\tilde{V}_{i,k} = \{v_{i-k}, \ldots, v_i, \ldots, v_{i+k}\} \subseteq V_1$, $\tilde{V}_{j,k} = \{v_{j-k}, \ldots, v_j, \ldots, v_{j+k}\} \subseteq V_2$* [5].

[2] Renal Cell Carcinoma (RCC) is typically asymptomatic and surgery usually increases patients life only for early stage tumors. However, some cystic and solid renal lesions cannot be confidently differentiated from clear cell RCC. Therefore early detection of robust markers to distinguish malignant kidney tumors is needed. See for instance [5].

[3] This requirement is motivated by the necessity to employ new definitions.

[4] Bicliques are therefore "extreme" forms of highly inter-connected bipartite graphs.

[5] We also refer to the pair (v_i, v_j) and the constant k as, respectively, the *center* and the *ray* of the *neighborhood* $M_{i,j,k}$.

Definition 2 (Density). *Let $G = (V_1 \cup V_2, E)$ be a bipartite graph and $M_{i,j,k} = (\tilde{A}, \tilde{E})$ a neighborhood of size $N_e = |\tilde{E}|$, centered in (v_i, v_j), with $\tilde{A} = \tilde{V}_{i,k} \cup \tilde{V}_{j,k}$. We define the density of $M_{i,j,k}$ as*

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

Definition 3 (Bipartite Graph Region). *Let $G = (V_1 \cup V_2, E)$ be a* bipartite graph *and V_1, V_2 two totally ordered sets. We say that $S_{a,b} = (\tilde{V}_1 \cup \tilde{V}_2, \tilde{E})$ is a region of G if it is the subgraph induced through the sequences of vertices $\tilde{V}_1 = \{v_a, v_{a+1}, \ldots, v_b\} \subseteq V_1$, $v_a \leq v_b$ and $\tilde{V}_2 = \{v_a^{'}, v_{a+1}^{'}, \ldots, v_b^{'}\} \subseteq V_2$, $v_a^{'} \leq v_b^{'}$.*

We also say that, given two bipartite graphs $G_1 = (V_1 \cup V_2, E_1)$ and $G_2 = (V_1 \cup V_2, E_2)$, then S_{x_1,y_1} and S_{x_2,y_2} are *common* regions of both G_1 and G_2 only if $x_1 = x_2$ and $y_1 = y_2$.

3 Problem Formulation

3.1 Reference Model for the Observed Data

As is used to represent sets of many interacting entities, we can express relationships between MS signals of a given case or control group through a graph whose vertices are observed mass-to-charge ratios, and edges represent dependencies between signal intensities with different mass-to-charge values. Broadly speaking, we assume there is a system under study that may be represented by a graph $G = (V, E)$ (say population graph), however instead of having all of G available to us, we take measurements that effectively yield a sample of vertices $\tilde{V} \subseteq V$ and edges $\tilde{E} \subseteq E$, which we compile into our data model $R = (\tilde{V}, \tilde{E})$. In fact, we extend this assumption by considering both a case G^{case} and a control G^{ctrl} population graph. That is, the measurements yield two models R^{case} and R^{ctrl} which we respectively refer to as the *case* and the *control template*. Here the target is to sample neighborhoods from both the case template R^{case} and the control template R^{ctrl}, in such a way that we can define test of hypotheses by conjecturing statements over the "population graph" parameters, for instance being given by the neighborhood densities associated to the case and control "population graphs".

More formally, we observe a set of N mass-to-charge ratios $\{m_1, m_2, \ldots m_N\}$ and for each mass-to-charge ratio m_s, $1 \leq s \leq N$, we measure a vector of intensity levels $\{i_{m_s,1}, i_{m_s,2}, \ldots, i_{m_s,n}\}$. Moreover, let $\mathcal{I}^{case}$ and $\mathcal{I}^{ctrl}$ be the case and control groups, respectively. We assume that each group (for example, $\mathcal{I}^{ctrl}$) can be expressed through a product $\mathcal{I}^{ctrl}_{m_1} \times \mathcal{I}^{ctrl}_{m_2} \times \ldots \times \mathcal{I}^{ctrl}_{m_N}$ of *spaces* $\mathcal{I}^{ctrl}_{m_s}$, $s \in [N]$[6] endowed with a distribution function $f^{ctrl}_{I_{m_s}}$, where I_{m_s} is the random variable expressing the signal intensity for a protein/peptide molecule with mass to charge ratio m_s [7].

A simple but commonly used measure to quantify dependencies between variables is the *mutual information* [7]. Thus, we employ mutual information to quantify dependencies between pairs of signals with different mass-to-charge value. Let us consider the following definition for any group of patients g on which is defined a distribution $f^g_{I_{m_s}}$.

Definition 4 (Template). *Let g be a group of subjects. For each $s, t \in [N]$ let $\{i_{m_s,1}, \ldots, i_{m_s,n}\}$ and $\{i_{m_t,1}, \ldots, i_{m_t,n}\}$ be two sets of observations obtained from $f^g_{I_{m_s}}$ and $f^g_{I_{m_t}}$. We call* template *(of g) the bipartite graph $R^g = (V_1 \cup V_2, E)$ with $V_1 = \{m_1, m_2, \ldots, m_N\}$, $V_2 =$*

[6] We use the bracket notation $[N]$ to denote the set $\{1, \ldots, N\}$.

[7] We follow the standard convention that random variables are denoted by upper case letters (e.g., random intensities I_{m_s}), and observations (i.e., intensity values i_{m_s}) are denoted by lower case letters.

$\{m_1^{'}, m_2^{'}, \ldots, m_N^{'}\}$ *and* $(m_s, m_t^{'}) \in E$, $s,t \in [N]$, *only if* $\mathrm{MI}(f^g_{I_{m_s}} || f^g_{I_{m_t}}) \geq \delta$. *Where* $\mathrm{MI}(f^g_{I_{m_s}} || f^g_{I_{m_t}})$ *gives, for any pair* m_s *and* m_t, *the estimation of the mutual information (MI) between* $f^g_{I_{m_s}}$ *and* $f^g_{I_{m_t}}$ [8].

3.2 Reference Method to Provide Decisions

We assume that *templates* can be thought of as graphs from a larger underlying "population graphs". Moreover, we recall that statistical hypotheses (noted as H_0 and H_A) are competing statements concerning the population parameters. This way templates establish abstract frames of reference, giving the opportunity to employ statistical test of hypotheses over their properties. The rationale for establishing our hypotheses is deciding whether the *case template* (e.g., ccRCC template) shows modified neighborhood cohesion as compared to the *control template*. Since we use edge densities to analyze cohesion, we should also say that for two sets of neighborhoods to be consistent with the above rationale, it is enough that $\mu^{ctrl} \neq \mu^{case}$, where μ^{ctrl} and μ^{case} are the neighborhood mean densities in the population graphs of the case/control groups. Moreover, given a template $R = (V_1 \cup V_2, E)$, we can easily provide a set of density measurements $D = \{\mathrm{den}(M_1^R), \mathrm{den}(M_2^R), \ldots, \mathrm{den}(M_n^R)\}$ by sampling a set of neighborhoods $\{M_1^R, M_2^R, \ldots, M_n^R\}$ from R[9]. Hence in order to check a significant difference of the neighborhood mean density between groups of neighborhoods provided from two different population graphs we can simply *match* pairs of neighborhoods in the (associated sample) template models R^{case} and R^{ctrl}. Therefore, given the following quantities:

- the (paired) samples of density measurements (from controls) $D^{ctrl} = \{\mathrm{den}(M_1^{ctrl}), \mathrm{den}(M_2^{ctrl}), \ldots, \mathrm{den}(M_n^{ctrl})\}$,
- the (paired) samples of density measurements (from cases) $D^{case} = \{\mathrm{den}(M_1^{case}), \mathrm{den}(M_2^{case}), \ldots, \mathrm{den}(M_n^{case})\}$,
- their differences $D = \{D_i : D_i = X_i - Y_i, X_i \in D^{ctrl}, Y_i \in D^{case}\}$,
- the sample mean $\tilde{D}$ and
- the sample standard deviation of difference scores Σ_d,

then we can reject the *null* $H_0 : \mu^{ctrl} = \mu^{case}$ (no change) in favor of the *alternative* $H_A : \mu^{ctrl} \neq \mu^{case}$ using $T = \frac{\tilde{D}}{\Sigma_d/\sqrt{n}}$ 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 T 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.

The use of regions gives us the opportunity to analyze the neighborhood 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, given two regions S_1 and S_2 *common* both to R^{ctrl} and R^{case}, we can obtain two sets of densities $D^{ctrl}_{S_1}$ and $D^{case}_{S_2}$ simply by sampling neighborhoods from S_1 and S_2 respectively. As previously stated, using these data as observations provided by sampling neighborhoods both from control and case groups in S_1 and S_2, we are able to apply the test $H_0 : \mu^{ctrl}_{S_1} = \mu^{case}_{S_2}$ against $H_A : \mu^{ctrl}_{S_1} \neq \mu^{case}_{S_2}$ for any pair

[8] In this paper we apply the *histogram estimation*. More sophisticated techniques are discussed in [17].

[9] Here we use the notation M_i^R to emphasize that the ith neighborhood is sampled from R. Below we will use similar notation when the sampling is done from a region S of e.g., a control template R^{ctrl}. In that case the sample is denoted by $\{M_1^{ctrl,S}, M_2^{ctrl,S}, \ldots, M_n^{ctrl,S}\}$.

of regions S_1 and S_2; that is, by observing regions differently located over the graphs, we test the neighborhood cohesion modifications from group to group in different parts of the spectra.

We summarize the *Case VS Control* tests through the following procedure.

1. We represent R^{ctrl} (as in section 3.1) by observing (from the control group) for each pair (m_i, m_j) the intensities $\{i_{m_i,1}, i_{m_i,2}, \ldots, i_{m_i,n}\}$ and $\{i_{m_j,1}, i_{m_j,2}, \ldots, i_{m_j,n}\}$.

2. We represent R^{case} (as in section 3.1) by observing (from the case group) for each pair (m_i, m_j) the intensities $\{i_{m_i,1}, i_{m_i,2}, \ldots, i_{m_i,n}\}$ and $\{i_{m_j,1}, i_{m_j,2}, \ldots, i_{m_j,n}\}$.

3. Given any pair of regions S_1 and S_2, common both to R^{ctrl} and R^{case}, we obtain the local densities: *a)* $D^{case}_{S_1} = \{\text{den}(M_1^{case,S_1}), \text{den}(M_2^{case,S_1}), \ldots, \text{den}(M_n^{case,S_1})\}$ and
b) $D^{ctrl}_{S_2} = \{\text{den}(M_1^{ctrl,S_2}), \text{den}(M_2^{ctrl,S_2}), \ldots, \text{den}(M_n^{ctrl,S_2})\}$
Then for each considered pairs S_1 and S_2, we employ these sets together with the Student's t statistic (as *test statistic*) in the following tests:

$$H_0 : \mu^{case}_{S_1} = \mu^{ctrl}_{S_2} \quad \text{Vs.} \quad H_A : \mu^{case}_{S_1} \neq \mu^{ctrl}_{S_2}, \tag{2}$$

where $\mu^{ctrl}_{S_1}$ and $\mu^{case}_{S_2}$ are the neighborhood mean densities for respectively the control and case (population) graphs.

3.3 Reference Model of Variability

Starting from an observed template $R^g = (V_1 \cup V_2, E)$ we wish to define a random graph able to preserve (within a defined range) a property of it. A random graph Γ can be defined as a probability space of the form $(\mathcal{G}, \Pr)$ where $\mathcal{G}$ is an "ensemble" of possible graphs and Pr is a *probability measure* [2]. The difference among various models depends largely on how we specify $\mathcal{G}$ and Pr [14, 12]. Some methods for example, let Pr be *uniform* (all graphs have equal probability), restricting $\mathcal{G}$ to contain only those graphs with specific properties. Other approaches induce Pr through the application of some generative mechanisms. Here we follow the latter approach.
As we just stated above in section 3, our interest is to analyze the neighborhood cohesion of an observed graph (i.e., template) R^g. Hence, we attempt to preserve the densities of R^g by preserving (on average) its size. Among the many methods for defining a random graph from any observed graph (see, for example, the problem of graph randomization [10]), while preserving some properties, we simply "modify" randomly the edges of R^g. We realize this perturbation in such a way that the expected number of "modifications" takes values inside specific ranges. The following definition formalizes this (random) mechanism.

Definition 5 ((s, t, R^g)–Preserving Random Graph)**.** *Let $R^g = (V_1 \cup V_2, E)$ 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^g)–preserving r.g. $\tau(s, t)$ if the expected number of edge additions and deletions in R^g take values in $[s, t]$.*

Please note that, since any region S of R^g is simply a sub-graph of R^g, we can also apply definition 5 to regions. Let S be any region of R^g, we say that $\tau(s, t)$ is an (s, t, S)–preserving RG if the condition of definition 5 is verified for S. Moreover, it is easy to prove the following result.

Property 1. *Let $R^g = (V_1 \cup V_2, E)$ be a template. We should obtain an (s, t, R^g)–preserving r.g. $\tau(s, t)$ by constraining the "perturbation" probability p in definition 5 in such a way that $\frac{s}{n^2} \leq p \leq \frac{t}{n^2}$, where $n^2 = |V_1 \times V_2|$.*

Proof. Let X_e be the r.v. which expresses the "modification" of any $e \in V_1 \times V_2$ in $R^g = (V_1 \cup V_2, E)$, i.e., $X_e = 1$ if e is added to (or deleted from) E; $X_e = 0$ otherwise. By assumption $X_e \sim \text{Ber}(p)$ (i.e., Bernoulli r.v. of parameter p). Let $\frac{s}{n^2} \leq p \leq \frac{t}{n^2}$ for any pair of integers $s \leq t$. Then we have $X = \sum_{e \in V_1 \times V_2} X_e$; that is X counts the number of edges which have been "modified" (added/deleted) in R^g. Since $|V_1 \times V_2| = n^2$, we have that $X \sim \text{Bin}(n^2, p)$ (i.e., a Binomial r.v. of parameters n^2 and p), then we have $s \leq E[X] = n^2 p \leq t$. □

As previously reported in section 3.2, we can also formulate the test of hypotheses when the perturbation mechanism in definition 5 is applied. To tackle this problem, we turn e.g., to a well-known form of the *Monte Carlo* method. Here, we summarize the *Control vs Case* tests through the following procedure.

1. We define R^{ctrl} as in step 1 of the *Controls VS. Cases* procedure (section 3.2).
2. We define R^{case} as in step 2 of the *Controls VS. Cases* procedure (section 3.2).
3. For any pair of regions S_1 and S_2 – common both to R^{ctrl} and R^{case} (i.e., two templates), we generate two (Monte Carlo) samples:

   - (i) n realizations $\{\tilde{\tau}_1^{(1)}, \tilde{\tau}_1^{(2)}, \ldots, \tilde{\tau}_1^{(n)}\}$ of $\tau_1(s, t)$ and
   - (ii) n realizations $\{\tilde{\tau}_2^{(1)}, \tilde{\tau}_2^{(2)}, \ldots, \tilde{\tau}_2^{(n)}\}$ of $\tau_2(s, t)$,

   where $\tau_1(s, t)$ and $\tau_2(s, t)$ are respectively (s, t, S_1)–preserving and (s, t, S_2)–preserving random graphs. Then, we derive the estimates: $\tilde{\theta}_1(i) = \text{den}(\tilde{\tau}_1^{(i)})$, $i \in [n]$ (respectively, $\tilde{\theta}_2(i) = \text{den}(\tilde{\tau}_2^{(i)})$) and apply them as observations from the distribution $\text{den}(\tau_1(s, t))$ (respectively, $\text{den}(\tau_2(s, t))$). Finally we test the following conjectures (for all the considered pairs S_1 and S_2):

   $$H_0 : \mu_{S_1}^{ctrl} = \mu_{S_2}^{case} \quad \text{Vs.} \quad H_A : \mu_{S_1}^{ctrl} \neq \mu_{S_2}^{case} \tag{3}$$

   where $\mu_{S_1}^{ctrl}$ and $\mu_{S_2}^{case}$ are the neighborhood mean densities for respectively the control and case (population) graphs.

4 RCC Analysis

This study has been applied to data provided and elaborated from a cohort of subjects screened at the "Ospedale Maggiore Policlinico", San Gerardo Hospital (Monza, Italy) and Desio Hospital (Desio, Italy). The samples consists of 85 control subjects and 102 RCC patients. Patients have been divided into groups according to their pathologies: clear cell RCC ($n = 79$) and other different histological subtypes i.e., non-ccRCC ($n = 23$). Specific mass spectrometry techniques and elaborations (see e.g. [6]) were applied to obtain the following data sets. i) Data collection 1: Control and ccRCC MS data (85 control vs 79 ccRCC patients); ii) Data collection 2: Control and non-ccRCC MS data (85 control vs 23 non-ccRCC patients); iii) Data collection 3: ccRCC and non-RCC MS data (79 ccRCC patients vs 23 non-ccRCC patients).

We summarize our RCC analysis as follows [10].

[10] We report the procedure for data collection number 1. Generalizations to others dataset are straightforward.

1. We represent R^{ctrl} (as in section 3.1) by observing (from control) for each pair (m_i, m_j) the intensities $\{i_{m_i,1}, i_{m_i,2}, \ldots, i_{m_i,n}\}$ and $\{i_{m_j,1}, i_{m_j,2}, \ldots, i_{m_j,n}\}$.

2. We represent R^{rcc} (as in section 3.1) by observing (from ccRCC) for each pair (m_i, m_j) the intensities $\{i_{m_i,1}, i_{m_i,2}, \ldots, i_{m_i,n}\}$ and $\{i_{m_j,1}, i_{m_j,2}, \ldots, i_{m_j,n}\}$.

3. Given any pair of regions S_1 and S_2, common both to R^{ctrl} and R^{rcc}, we obtain the local densities:

   a) $D^{rcc}_{S_1} = \{\mathrm{den}(M_1^{rcc,S_1}), \mathrm{den}(M_2^{rcc,S_1}), \ldots, \mathrm{den}(M_n^{rcc,S_1})\}$, and
   b) $D^{ctrl}_{S_2} = \{\mathrm{den}(M_1^{ctrl,S_2}), \mathrm{den}(M_2^{ctrl,S_2}), \ldots, \mathrm{den}(M_n^{ctrl,S_2})\}$.

   Then for each pair S_1 and S_2, we employ these sets together with the Student's t statistic in the following tests:

   $$H_0 : \mu^{rcc}_{S_1} = \mu^{ctrl}_{S_2} \quad \text{Vs.} \quad H_A : \mu^{rcc}_{S_1} \neq \mu^{ctrl}_{S_2}, \tag{4}$$

   where $\mu^{ctrl}_{S_1}$ and $\mu^{rcc}_{S_2}$ are the neighborhood mean densities for respectively the control and ccRCC (population) graphs.

4. For any pair of regions S_1 and S_2 – common both to R^{ctrl} and R^{rcc} (i.e., two templates), we generate two (Monte Carlo) samples:

   - (i) n realizations $\{\tilde{\tau}_1^{(1)}, \tilde{\tau}_1^{(2)}, \ldots, \tilde{\tau}_1^{(n)}\}$ of $\tau_1(s,t)$ and
   - (ii) n realizations $\{\tilde{\tau}_2^{(1)}, \tilde{\tau}_2^{(2)}, \ldots, \tilde{\tau}_2^{(n)}\}$ of $\tau_2(s,t)$,

   where $\tau_1(s,t)$ and $\tau_2(s,t)$ are respectively (s,t,S_1)–preserving and (s,t,S_2)–preserving random graphs. Then, we derive the estimates: $\tilde{\theta}_1(i) = \mathrm{den}(\tilde{\tau}_1^{(i)})$, $i \in [n]$ (respectively, $\tilde{\theta}_2(i) = \mathrm{den}(\tilde{\tau}_2^{(i)})$) and apply them as observations from the distribution $\mathrm{den}(\tau_1(s,t))$ (respectively, $\mathrm{den}(\tau_2(s,t))$). Finally we test the following conjectures (for all the considered pairs S_1 and S_2):

   $$H_0 : \mu^{ctrl}_{S_1} = \mu^{rcc}_{S_2} \quad \text{Vs.} \quad H_A : \mu^{ctrl}_{S_1} \neq \mu^{rcc}_{S_2} \tag{5}$$

   where $\mu^{ctrl}_{S_1}$ and $\mu^{rcc}_{S_2}$ are the neighborhood mean densities for respectively the control and ccRCC (population) graphs

5 Numerical Experiments

In Sec. 3, 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 against template perturbation. With this concern in mind, here discuss some numerical results obtained after perturbations of the template's parameters.

We divided each template into a fixed number of regions (i.e., 8). This way, for a given pair R^{ctrl} and R^{case}, tests were defined over 8 different pairs of regions, common both to R^{ctrl} and R^{case}. We also defined a set of arbitrary thresholds $T = \{0.01, 0.006, 0.001, 0.0001, 0.00001\}$ and a set of arbitrary (neighborhood) rays $R = [4]$. Thus, for each combination of $\delta \in T$ and $k \in R$,

we considered (for each class of tests) 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 mass-to-charge ratio bounds identifying modified regions over the spectra at a specific level of significance. Each class of test was applied to its associated data collection. Table 1a reports for each class the spectra bounds which constrain pairs of regions, common both to the case (i.e., ccRCC or non-ccRCC) and control groups, where tests have rejected the null hypothesis at the 5% significance level, thus stating that a difference in the neighborhood mean density of these regions exists between case and control groups. We referred to these regions as *distinguishing regions*, shortly reported as DR. We called the regions which are not constrained by the bounds in table 1a, *non-distinguishing regions*, shortly reported as NDR. These are regions where tests have accepted the null hypothesis thus stating that no difference in the considered cohesion exists between case and control groups. To check the robustness of test decisions (equivalently, the robustness of the distinguishing/non-distinguishing capabilities for all the considered common regions) we verified whether after the application of the variability model in definition 5, the capability of any pair of common regions to distinguish (i.e., DRs) or to not distinguish (i.e., NDRs) – between control and case, was preserved. Clearly, after a perturbation, we can still obtain new distinguishing / non-distinguishing regions following the test decisions, as previously performed before the application of the considered mechanism. Therefore, the question of interest was to asses the association between the two categorical variables that can be considered in the above argumentation: (i) from one side, the *distinguishing capability before the perturbation* – which divide the regions into DRs and NDRs respectively and, (ii) from the other, the *test of hypotheses decision after the perturbation* – which in turn may accept or reject that the considered DRs and NDRs still preserve their distinguishing capabilities, as previously observed. These are the type of questions that Fisher's exact test (for small samples) is designed to answer. Hence, we verified whether the three pairs of DRs (common both to control and case groups), identified by the (m/z) bounds in Table 1a, still preserved their DR capabilities. Similarly, we verified whether the remaining NDRs still preserved their NDR capabilities after the realized perturbation. Tables 1b summarize the results for each class CVR, CVNR and, RVNR respectively. These tables report for different values of the perturbation probabilities the number of tests accepting H_0 (H_A) against the region's property after the applied perturbations. For example, given a perturbation probability of 0.05 (Table 1b), none of the 12 tests, performed over the three common DRs, modified their conclusions (still rejecting the null). On the other hand, only 1 of the 20 tests, performed over NDRs, gave a different conclusion, i.e., accepting the alternative hypothesis. To apply Fisher's test we evaluated the probability of obtaining our results or more extreme results. This gives the one-side p–values reported in Table 1c. Fisher's exact test confirms a significant association, at the 0.001 level, between decisions and region's property with perturbation equal to 0.005 and 0.1 for CVR tests and up to perturbations equal to 0.3 for both the CVNR and RVNR tests. For these probabilities values, tests still preserve their decisions - as previously obtained, providing robust conclusions in the experimental design. We can also say that, by considering as acceptable those template modifications for which at most, about 22 (i.e., by defining a $\tau(0, 22)$ r.g. in property 1) of the possible edges are removed one is able to obtain, thanks to property 1, the perturbation probability $p \leq 0.1$ for observing robust conclusions for the CVR class. Moreover, with the same frequency of modifications we still preserve process decisions for the other classes of tests. In other words, in order for the modeler to evaluate whether a set of acceptable parameters e.g., $\{s, t\}$ provides reliable conclusions he can simply apply the perturbation mechanism over the observed model, and check whether for these parameters a

perturbation probability satisfying property 1 give rise to decisions, which still preserve the conclusions previously obtained.

6 Conclusions

The robustness of a biological system is mainly defined as a property of a biological function [11]. For this reason robustness here relates to the determination of the effect of certain perturbations on the spectra signals of peptide/protein relationships. In fact, many domains are best described by "relational" models in which instances of multiple types are related to each other in complex ways. In such situations "relational data" may increase e.g. the performance of classification or clustering models for "difficult" data sets (e.g. [8, 21, 20, 1]). To this concern it is important to note that classification or clustering problems not only are two main fundamental issues for the machine learning community, but they even have been studied from computational point of view [3]. Graphs provide a canonical representation for "relational data" and their employment to reinterpret traditional data seems to be promising in order to better understand and summarize relationships amongst very large number of observations. In this paper graphs 3.2 was employed to define a reference method for our analysis. Using this approach we obtained "differentially expressed" spectra regions between case/control groups. Many questions still need to be addressed in future analysis. For example, we employed the reference method to obtain distinguishing regions for their biological evidence, but a better molecular understanding of the added/deleted edges (i.e., molecular dependencies) should be clearly more appealing from a biological point of view. Further analysis should concern the use of some parameters which we have arbitrary defined as constant values. Probably the most critical parameter is the threshold (delta) for representing in section 3.1 dependencies in the template graphs. The selection of these values can have different effects on the model accuracy. Finally from a clinical prospective, since proteins that are differentially expressed as a consequence of a disease have great potential as new bio-markers, we need to conclusively determine both the classification predictive power of the RCC distinguishing regions and their biological identity aimed to explore the structure and function of these potential bio-markers.

References

[1] M. Antoniotti, M. Carreras, A. Farinaccio, G. Mauri, D. Merico, and I. Zoppis. An application of kernel methods to gene cluster temporal meta-analysis. *Comp. & Oper. Res.*, 37(8):1361–1368, 2010.

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

[3] P. Bonizzoni, G. Della Vedova, R. Dondi, and T. Jiang. On the approximation of correlation clustering and consensus clustering. *J. of Comp. and System Sciences*, 74(5):671 – 696, 2008.

[4] U. Brandes and T. Erlebach, editors. *Network Analysis: Methodological Foundations*, volume 3418 of *Lect. Notes in Computer Science*. Springer, 2005.

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

[6] C. Chinello, C. Galbusera, E. Gianazza, V. Mainini, I. Zoppis, S. Picozzi, F. Rocco, G. Galasso, S. Bosari, S. Ferrero, R. Perego, F. Raimondo, C. Bianchi, M. Pitto, S. Signorini, P. Brambilla, P. Mocarelli, K. Galli, and F. Magni. Serum biomarkes of renal cell carcinoma assessed using a protein profiling approach based on clinprot technique. *Urology*, 75(04):842 – 847, 2009.

[7] T. M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, 2006.

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

CVR		CVNR		RVNR	
From (m/z)	*To* (m/z)	*From* (m/z)	*To* (m/z)	*From* (m/z)	*To* (m/z)
1719.45	2084.34	1719.45	2084.34	1719.45	2084.34
2644.49	3214.26	2092.18	2563.79	2644.49	3214.26
3270.53	4018.88	4050.39	4540.1	3270.53	4018.88

(a) Regions with different mean intensity values.

Perturbation	0.05		0.1		0.2		0.3	
Hp. after perturbation	H_0	H_A	H_0	H_A	H_0	H_A	H_0	H_A
CVR test								
DRs	0	12	0	12	1	11	1	11
NDRs	19	1	13	7	12	8	12	8
CVNR test								
DRs	0	12	0	12	0	12	0	12
NDRs	20	0	17	3	17	3	12	8
RVNR test								
DRs	0	12	0	12	0	12	0	12
NDRs	19	1	15	5	12	8	15	5

(b) Contingency tables for the Fisher's exact test.

CVR	*CVNR*	*RVNR*	*Perturbation*
6.00E-08	0.00000001	6.00E-08	0.005
0.00022316	0.00000199	0.00002741	0.1
0.00457479	0.00000199	0.0005579	0.2
0.00457479	0.0005579	0.00002741	0.3

(c) Fisher's exact test at 0.001 level: p-values

Table 1: RCC Analysis.

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

[10] S. Hanhijärvi, G.C. Garriga, and K. Puolamäki. Randomization techniques for graphs. In *Proc. of the 9th SIAM Int. Conf. on Data Mining (SDM '09)*, pages 780–791, 2009.

[11] H. Kitano. Biological robustness. *Nat Rev Genet*, 5(11):826–37, 2004.

[12] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models.* Springer, 2009.

[13] P. Kouvelis and G. Yu. *Robust Discrete Optimization and Its Applications.* Springer, 1996.

[14] D.J. Marchette. *Random Graphs for Statistical Pattern Recognition.* Wiley–Interscience, 2004.

[15] 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.

[16] J.V. Rosenhead. Robustness analysis: Keeping your options open. In *Rational Analysis for a Problematic World Revisited: Problem Structuring Methods for Complexity, Uncertainty and Conflict*, pages 181–207. John Wiley, 2001.

[17] B.W. Silverman. *Density estimation for statistics and data analysis.* Chapman and Hall, 1986.

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

[19] H.-Y. Wong and J. Rosenhead. A rigorous definition of robustness analysis. *J. of the Oper. Res. Soc.*, 51:176–182, 2000.

[20] I. Zoppis and G. Mauri. Clustering dependencies with support vectors. *LNEE*, 6:155–165, 2008.

[21] I. Zoppis, D. Merico, M. Antoniotti, B. Mishra, and G. Mauri. Discovering relations among go-annotated clusters by graph kernel methods. *LNCS*, 4463:158–169, 2007.