Polymer 49 (2008) 5575–5587

Contents lists available at ScienceDirect

Polymer

journal homepage: www.elsevier.com/locate/polymer

Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case

Maykel Cruz-Monteagudo [a,b,c], Cristian Robert Munteanu [a,c,d], Fernanda Borges [c], M. Natália D.S. Cordeiro [d], Eugenio Uriarte [a], Kuo-Chen Chou [e], Humberto González-Díaz [a,e,*]

[a] Unit of Bioinformatics & Connectivity Analysis (UBICA), Institute of Industrial Pharmacy, Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[b] CEQA, Faculty of Chemistry and Pharmacy, UCLV, Santa Clara 54830, Cuba
[c] Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, 4150-047 Porto, Portugal
[d] REQUIMTE/Science Faculty, Chemistry Department, University of Porto, 4169-007 Porto, Portugal
[e] Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ARTICLE INFO

Article history:
Received 1 July 2008
Received in revised form
26 September 2008
Accepted 30 September 2008
Available online 18 October 2008

Keywords:
QSPR
Biopolymers mixtures
Complex networks

ABSTRACT

The Quantitative Structure–Property Relationships (QSPRs) based on Graph or Network Theory are important for predicting the properties of polymeric systems. In the three previous papers of this series (Polymer 45 (2004) 3845–3853; Polymer 46 (2005) 2791–2798; and Polymer 46 (2005) 6461–6473) we focused on the uses of molecular graph parameters called topological indices (TIs) to link the structure of polymers with their biological properties. However, there has been little effort to extend these TIs to the study of complex mixtures of artificial polymers or biopolymers such as nucleic acids and proteins. In this sense, Blood Proteome (BP) is one of the most important and complex mixtures containing protein polymers. For instance, outcomes obtained by Mass Spectrometry (MS) analysis of BP are very useful for the early detection of diseases and drug-induced toxicities. Here, we use two Spiral and Star Network representations of the MS outcomes and defined a new type of TIs. The new TIs introduced here are the spectral moments (π_k) of the stochastic matrix associated to the Spiral graph and describe non-linear relationships between the different regions of the MS characteristic of BP. We used the MARCH-INSIDE approach to calculate the π_k(SN) of different BP samples and S2SNet to determine several Star graph TIs. In the second step, we develop the corresponding Quantitative Proteome–Property Relationship (QPPR) models using the Linear Discriminant Analysis (LDA). QPPRs are the analogues of QSPRs in the case of complex biopolymer mixtures. Specifically, the new QPPRs derived here may be used to detect drug-induced cardiac toxicities from BP samples. Different Machine Learning classification algorithms were used to fit the QPPRs based on π_k(SN), showing J48 decision tree classifier to have the best performance. These results suggest that the present approach captures important features of the complex biopolymers mixtures and opens new opportunities to the application of the idea supporting classic QSPRs in polymer sciences.

1. Introduction

A very important field of computational polymer sciences is devoted to the development of Quantitative Structure–Property Relationship (QSPR) models, linking the structure of polymers with their properties [1–4] or predicting the properties of catalysts of polymerization reactions [5a]. It is quite effective to use Graph or Complex Networks theory to deal with complicated chemical and biological polymeric systems because it can provide an intuitive image and help people gain useful insights into the mechanism concerned [5b,c]. There are numerous works that use Topological Indices (TIs) or different connectivity measures or Connectivity Indices (CIs) of graph or networks to derive QSPR models for small molecules [6] or polymers [7,8]. The reader may find two recent reviews focused on the network or graph based QSPRs, ranging from small-sized drugs to biopolymers and Complex Networks [9,10]. In

[*] Correspondence to: Humberto González-Díaz, Faculty of Pharmacy, University of Santiago de Compostela, Santiago de Compostela, 15782, Spain. Tel.: +34 981 563100; fax: +34 981 594912.
E-mail address: gonzalezdiazh@yahoo.es (H. González-Díaz).

0032-3861/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2008.09.070

this sense, it is also of major importance the comparative study reported by Bonchev and Buck [11]. Specifically, in the three previous papers of this series: its part 1 published in Ref. [12]; part 2 in Ref. [13]; and part 3 in Ref. [14] we focused on the uses of molecular TIs derived from the stochastic matrix associated to a graph or network representation of one polymer structure. However, in general, the application of TIs to the study of complex mixtures of artificial polymers or biopolymers, such as nucleic acids and proteins is still an emerging field dealing with a rather complicated problem. In this sense, the continuation of the previous series with a work aimed to extend QSPR models based on stochastic TIs to complex mixtures or polymers or biopolymers falls by its own weight.

In particular, Blood Proteome (BP) is one of the most important and complex mixtures containing protein polymers. Circulating carrier proteins have been recently found to act as a reservoir for the accumulation and amplification of bound low mass biomarkers, integrating, amplifying and storing diagnostic information like a capacitor stores electricity [15]. The blood proteome is changing constantly as a consequence of the perfusion of the organ undergoing drug-induced damage and this process subsequently adds, subtracts, or modifies the circulating proteome. Thus, even if these small peptide fragments are many degrees of separation removed from the actual insult, they can retain the specificity for the disease because this process can arise from a specific type of biomarker amplification based on the uniqueness of the tissue microenvironment where the organ toxicity occurs [16]. Consequently, BP represents a potential target for the early detection of diseases and drug-induced toxicities. Because body fluids such as serum, saliva or urine are a protein-rich information reservoir that contains the traces of what the blood has encountered on its constant perfusion and percolation throughout the body [16] and the optimal performance in the low mass range exhibited by Mass Spectroscopy (MS) [17], the use of this method applied to proteomics may offer the best chance of discovering these early stage changes. For instance, outcomes obtained by MS analysis of BP could be useful for the early detection of diseases and drug-induced toxicities.

The application of graph theory to MS was first proposed by Bartels for peptide sequencing [18]. The basic idea consists in transforming a mass spectrum into a graph called "*spectrum graph*". Basically, each peak in the experimental spectrum is represented as a vertex (or several vertices) in the spectrum graph and a directed edge is established between two vertices if the mass difference of the two vertices equals the mass of one or several amino acids. Several algorithms that make use of spectrum graphs have been designed for *de novo* peptide sequencing. Among the most popular are "SeqMS" [19], "Lutefisk" [20], "Sherenga" [21] and more recently "PepNovo" [22]. The construction of the spectrum graph of all these algorithms shares the basic idea proposed by Bartels with their respective particularities. However, most of the above-mentioned methods deal mainly with the MS of one protein and not with complex mixtures of biopolymers relevant to several clinical problems. By contrary, drug toxicity and specifically cardiotoxicity are serious adverse effects of chemotherapy involving the complex mixtures of biopolymers present in BP. It encompasses a spectrum of disorders, ranging from relatively benign arrhythmias to potentially lethal conditions, such as myocardial ischemia/infarction and cardiomyopathy [23]. The toxicity of chemotherapeutic drugs can cause loss of myocytes sarcolemmal integrity, release of bioactive markers into the extracellular environment (tissue and circulation) and ultimately leading to the necrosis of myocytes [24]. The extent and severity of the necrosis can be monitored by the levels of bioactive markers [25]. However, the number of new biomarkers reaching routine clinical use remains unacceptably low [26]. Due to the thousands of intact and cleaved proteins in the BP, finding the single disease-related protein could be like searching for a needle in a haystack, requiring the separation and identification of each protein biomarker. In addition, most commonly used toxicity biomarkers appear only when significant organ damage occurred. For these reasons, to identify patterns by using the serum proteome spectrum instead of directly identify a single marker candidate, represents a more attractive and realistic choice for this purpose. In this sense, Petricoin et al. successfully identified patterns of low molecular weight biomarkers as ion peak features within the spectra, and used these patterns as the diagnostic endpoint itself for the early detection of drug-induced cardiac toxicities [27], ovarian [28] and prostate cancer [29].

In the present work we continue the previous series of work with a redirection of our attention to the application of new stochastic TIs and QSPR method to BP. We aim to use it in generating a prediction model based on a graph theoretical approach instead of directly identify patterns within the MS outputs. In our previous work [30] we introduced an alternative graph theoretical representation of BP in analogy to the four-color maps introduced by Randic et al. for DNA sequences representation [31]. Here, we derive a new family of stochastic TIs using Markov chain theory and this new graphical representation of BP. The new TIs are the spectral moments ($^{sr}\pi_k$) of the stochastic associated to the Spiral graph of BP. These numerical indices are then used as patterns in the derivation of a Quantitative Proteome–Property Relationship (QPPR) to illustrate the usefulness of this approach in complex mixtures of biopolymers. In addition, we compare these results with the topological indices of MS embedded Star Networks/Graphs (eSG) [32] computed in this work and with the LN and SN Shannon entropy calculated in a previous work [30]. The same eSG were used to create models for proteins [33] and enzymes [34].

The best QPPR model can be used for the early detection of drug-induced cardiac toxicities given the MS outcomes of a BP sample. This type of QPPR models may be considered as the biopolymer mixture analogue of classic QSPR models. More specifically, we can describe it as a Quantitative Proteome–Toxicity Relationships (QPTRs) in analogy to classic Quantitative Structure–Toxicity Relationship (QSTR) models. The graphic representation of the approach proposed in this work for the early detection of drug-induced cardiac toxicities is shown in Fig. 1.

2. Methods

2.1. Blood proteome MS dataset

For the generation of the MS Spiral networks of BP (BPMSSNs), the calculation of the π_k values and the development of the QPTR models, we used tab-delimited data files containing mass/charge (m/z) and peak intensity (I) values exported from serum rat proteome high-resolution spectra reported by Petricoin et al. [27]. According to Petricoin et al., the data files are generated by first exporting the raw data file generated from the QSTAR MS into tab-delimited files that generated approximately 350,000 data points per spectrum. The binning process condenses the number of data points to 7105 points per sample. The high-resolution spectra are binned using a function of 400 parts per million (ppm), so that all data files possess identical m/z values (e.g., the m/z bin sizes linearly increase from 0.28 at m/z 700 to 4.75 at m/z 12 000) [27]. Using the Spontaneously Hypertensive Rat (SHR) model, in which animals were challenged with doxorubicin or with mitoxantrone ± dexrazoxane (a routinely used cardioprotectant), over 200 samples collected and stored frozen over a 4-year period ($N = 203$) were analyzed. This study system has both well-known pathological and serum biomarker endpoints (cardiac lesion histological changes and serum cardiac troponin concentrations (cTnT), respectively) that have been recently used to measure effects of therapeutic compounds on cardiac damage [35]. Since the cardiac toxicity profile of 141 out of 203 samples analyzed was reported as

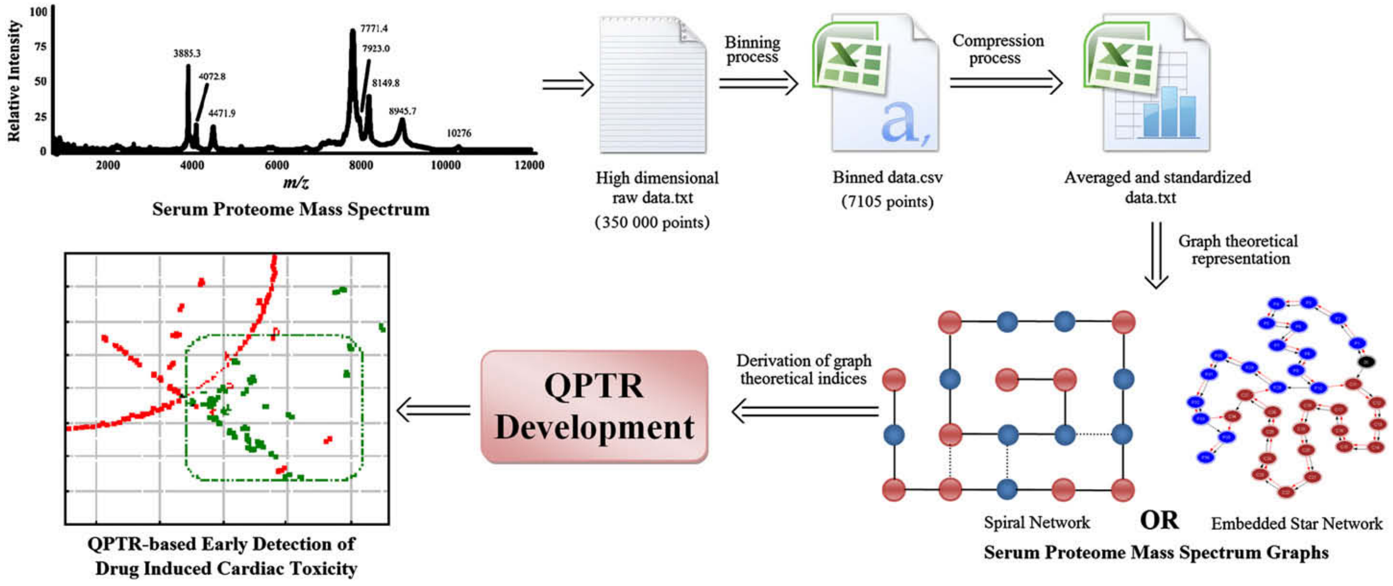

Relative Intensity
100
75
50
25
0
3885.3
4072.8
4471.9
7771.4
7923.0
8149.8
8945.7
10276
2000
4000
6000
8000
10000
12000
m/z
Serum Proteome Mass Spectrum
High dimensional raw data.txt
(350 000 points)
Binning process
Binned data.csv
(7105 points)
Compression process
Averaged and standardized data.txt
Graph theoretical representation
Spiral Network
OR
Embedded Star Network
Serum Proteome Mass Spectrum Graphs
Derivation of graph theoretical indices
QPTR Development
QPTR-based Early Detection of Drug Induced Cardiac Toxicity

Fig. 1. Schematic representation of the early detection of drug-induced cardiac toxicities.

unknown or with no definitive information about their cardiotoxic profile, only 62 samples were used in this work:

- Definitive Positive (34 samples with overt cardiotoxicity): Tab-delimited data files exported from serum proteome high-resolution spectra belonging to sera from SHR model with overt cardiotoxicity (cTnT $\geq$ 0.15 ng/ml and histologic lesion scores $\geq$ 1.0). We also included as positive the samples for rats with lower cTnT levels ($\geq$0.08 ng/ml) but as well with mild apparent pathologic changes as determined by histological score of lesion.
- Definitive Negative (28 samples without cardiotoxicity): Tab-delimited data files exported from serum proteome high-resolution spectra belonging to sera obtained from control SHR prior to treatment or following only 1–3 treatments with saline alone and whose cTnT = 0.

2.2. Blood proteome MS spiral networks

In order to generate the BPMSSNs we used MS binned data files derived from raw data files which were derived from MS of BP [27]. In addition, for graph representation the averaged and standardized m/z and I values were multiplied in order to obtain a value of the relationship between m/z and I that makes possible a graph representation. Such a Spiral graph is obtained in a similar way to the four-color maps introduced by Randic et al. for DNA sequences representation [31]. After that, a new averaged and standardized data file is generated consisting of 36 data points which can be used now in generating a blood proteome mass spectrum graph by using a Spiral representation. Since the values of m/z and I were standardized, the mean value of the m/z–I relationship of each sample is around 0.5. Consequently, a cut-off value of 0.5 is chosen for the values of m/z–I relationship related to each averaged data point. This cut-off value is used to codify each data point according to their respective average m/z–I relationship values, allowing their representation as a node on a 2D space. Specifically, in this work we represented the data points of the mass spectrum as a Spiral of nodes or vertices which are labelled differently; *i.e.* if the average m/z–I relationship absolute value <0.5, then the node is labelled with the letter C; otherwise is labelled with P. The Spiral begins with the first averaged and standardized mass spectrum data point (encoding information related to the spectrum's lower m/z region) and finishes in the last data point (spectrum's higher m/z region). If one connects the adjacent nodes labelled equally, then will obtain the Spiral network representing the serum proteome mass spectrum shown in Fig. 2. Two nodes are considered adjacent only if they are at one step away from each other in the Cartesian space. The only allowed are the connections in ordinates' and abscises' direction. Diagonal connections are not allowed because Euclidean distance of these nodes is different from 1 and consequently they are not considered adjacent. As a result, a segment of nodes with two different labels is obtained. Different labelling of nodes confers to each spectrum network a particular topology allowing their numerical (topological) characterization depending on the values of m/z and I of every MS of each sample. There is certain similarity between this type of BPMSSNs and other recently investigated by our group, but the processing of the MS and the procedure to link nodes in the Spiral network are different [36] (see next section).

2.3. Stochastic spectral moments of proteome MS spiral networks

By using the concept of Spectral Moments of a matrix we introduced the spectral proteomic stochastic moments as numerical indices of the BPMSSN [37], π_k(SN). In so doing, we used a Markov model (MM) to codify information about serum proteome mass spectral regions. Specifically, in this work the π_k(SN) was derived from the so-called MARCH-INSIDE (MARkov CHains INvariants for SImulation & DEsign) approach [38], which is used here for the first time to codify the information content encoded in a serum proteome mass spectrum. The MARCH-INSIDE approach has been applied previously to the field of proteins [37,39–42]. Here, the classic matrix MARCH-INSIDE approach [39] has been adapted to characterize the new Spiral networks. The method uses essentially two matrix magnitudes: the matrix $^{1}\prod$ and the zero order absolute initial probabilities vector $^{A}\boldsymbol{\pi}_{\mathbf{0}}$. The matrix $^{1}\prod$ is built up as a square matrix ($n \times n$) and contains the probabilities $^{1}p_{ij}$ to reach a node n_i moving throughout a walk of length $k = 1$ from a node n_j (see Eq. (1)).

$$^{1}p(a_{ij}, c_j) = a_{ij}{}^{*}c_j / \sum a_{ij}{}^{*}c_j, \qquad (1)$$

where, $\alpha_{ij} = 1$ if and only if the two nodes n_i and n_j are neighbours, placed at a topological distance $k = 1$ in the Spiral network, $\alpha_{ij} = 0$

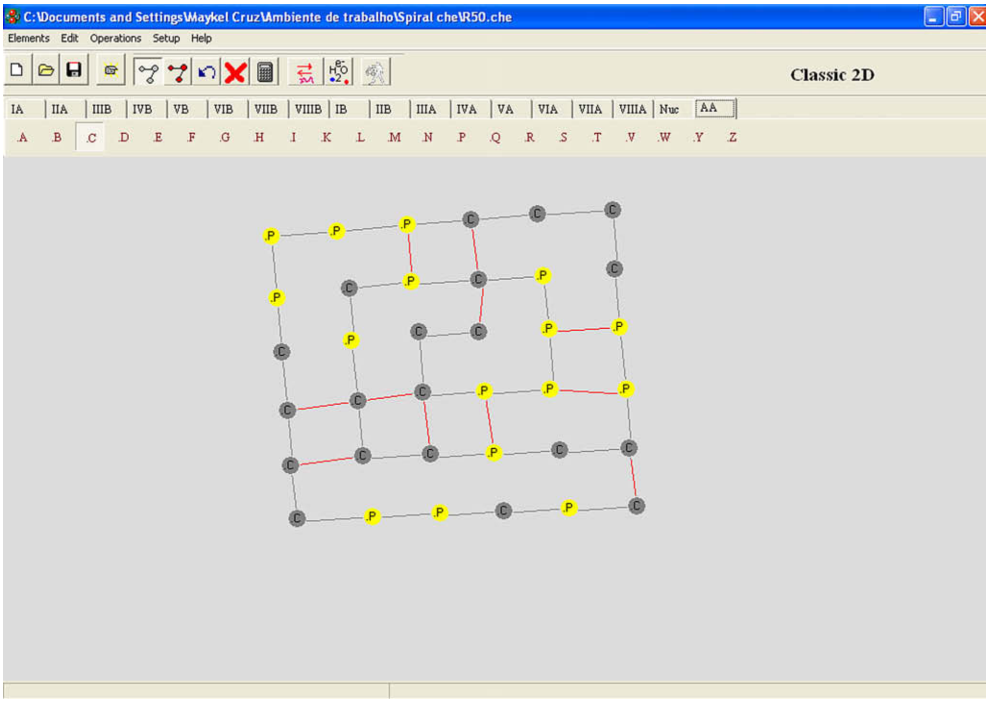

C:\Documents and Settings\Maykel Cruz\Ambiente de trabalho\Spiral che\R50.che
Elements Edit Operations Setup Help
Classic 2D
IA IIA IIIB IVB VB VIB VIIB VIIIB IB IIB IIIA IVA VA VIA VIIA VIIIA Nuc AA
A B C D E F G H I K L M N P Q R S T V W Y Z
P
C

Fig. 2. Blood proteome MS Spiral Network representation on MARCH-INSIDE interface.

otherwise. On the other hand, the parameter $c_j = 1$ if the m/z–I relationship > 0.5 for the MS region j represented by the node n_j, otherwise $c_j = 0.5$. This algorithm is different to other we used before processing MS and link nodes in the Spiral network [36]. On the other hand, the vector $^{\mathbf{A}}\boldsymbol{\pi}_\mathbf{0}$ lists the absolute initial probabilities $^{A}p_k(j)$ to reach a node n_i from a randomly selected node n_j (see Eq. (2)).

$$^{A}p_0(j) = 1/N, \tag{2}$$

where, the value N represents the number of nodes (spectral regions) in the Spiral network. Due to the particularities of the graph representation used here the $^{A}p_k(j)$ only depends on the total number of the data points or spectral regions on the graph. Consequently, all the nodes in the graph have the same and constant value of $^{A}p_k$. Since the elements of the matrices $^{k}\prod$ (which are the k natural powers of the matrix $^{1}\prod$) depend on the adjacency relationships between the nodes on the graph, the use of Markov chains (MCH) theory thus allows calculating the spectral proteomic stochastic moments (π_k) for any node n_j that one can reach in the Spiral network by moving from any node n_i throughout the entire network using walks of length k:

$$\pi_k(\mathrm{SN}) = \mathrm{Tr}\left(\left({}^{1}\prod\right)^{k}\right) = \sum_{j=1}^{n} {}^{k}p_{ij}, \tag{3}$$

where, Tr is the trace operator indicating the operation of summing up all the probabilities $^{k}p_{ii}$, within the main diagonal of these matrices. Finally, the MARCH-INSIDE software was used to compute the $\pi_k(\mathrm{SN})$ indices of order $k = 0, 1, 2\ldots, 10$). The π_0 was not used for the derivation of the QPTR models since the zero order index is constant by definition and only gives information related to the number of nodes or spectral regions in the spectrum graph.

2.4. LDA-based classification model of SN indices

Using the MARCH-INSIDE methodology as defined previously, we can develop a Linear Discriminant Analysis (LDA)-QPTR based. In the QPTR study the $\pi_k(\mathrm{SN})$ values are the TIs that play the role of independent or predictive variables. We selected LDA [43] in order to fit the discriminant function. The QPTR model classifies the rat's serum proteome spectrum into two general groups: cardiotoxic-risk (CT = 1 for positive samples) and non-cardiotoxic-risk (NCT = −1 for negative samples). In Eq. (4), b_k represents the coefficients of the classification function, determined by the least square method as implemented in the General Discriminant Analysis (GDA), a module of the STATISTICA 6.0 software package [44]. The general form of the QPTR model is described by the following equation:

$$\begin{aligned}\mathrm{CTR} &= b + b_1{*}\pi_1(\mathrm{SN}) + b_2{*}\pi_2(\mathrm{SN}) + \ldots + b_k{*}\pi_k(\mathrm{SN}) \\ &= b + \sum b_k{*}\pi_k(\mathrm{SN}),\end{aligned} \tag{4}$$

The best subset selection algorithm implemented on the GDA module was the method used to find the best combination of predictors producing the lowest percentage of misclassified instances on training and test sets, respectively [45,46]. The statistical significance of the LDA model was determined by Fisher's test by examining Fisher ratio (F) and the respective p-level (p). At the same time, the square Mahalanobis's distance (D^2) between the centroids of each one of the two groups (CT and NCT groups) and Wilks' U statistic were examined to test the discriminatory power of the function developed [47]. All the variables included in the model were standardized in order to bring them into the same scale. Subsequently, a standardized linear discriminant equation that allows to compare their coefficients is obtained [48]. We also inspected the cases/variables ratios (ρ parameter), and the number of variables to be explored in order to avoid over-fitting or chance correlation [45].

The most frequent cross-validation methods are the following: the independent dataset test, subsampling test, and jackknife test [49]. Chou and Shen have shown that only the jackknife test has the least arbitrariness [50,51]. Therefore, the jackknife test has been increasingly used by investigators to examine the accuracy of various predictors [52–57]. The Spiral QPTR model was trained by using the randomly selected 75% (47 out of 62) of the samples available. To test the predictive ability of the model we used the remaining 15 samples not used for training. The performance of the model on training and validation sets was verified by their respective Accuracies (Ac; it refers to the percentage of samples, which the model classifies correctly), Sensitivities (Se; percentage of cardiotoxic samples, which the model predicts to be cardiotoxic), and Specificities (Sp; percentage of non-cardiotoxic samples, which the model predicts to be non-cardiotoxic).

2.5. Machine learning classification algorithms

The data was analyzed with several different data mining algorithms for classification implemented in Weka data mining system [58,59]. We used three classification algorithms: a simple classification rule (OneR) and two decision trees (the random decision tree and the J48 decision tree). OneR, short for “One Rule”, is a simple machine learning classification algorithm that generates a one-level decision tree. OneR is able to infer typically simple classification rules from a set of instances that are straightforward for humans to interpret. This algorithm creates one rule for each attribute in the training data, and then selects the best rule as its “one rule”. The most frequent class for each attribute value must be determined to create a rule for a specific attribute. A rule is thus simply a set of attribute values bound to their majority class and it is based on such binding for each value of the attribute. OneR can also define the error rate of a rule as the number of misclassified training data instances by using the rule [59]. Typically, the machine learning algorithms select the rule with the lowest error rate as the best one. If two or more rules have the same error rate, the rule is chosen at random. In opposition, the OneR algorithm in Weka picks the rule with the highest number of correct instances, not the lowest error rate, and does not randomly select a rule when error rates are identical. Decision trees predict the value of a discrete dependent variable with a finite set of values (called class) from the values of a set of independent variables (called attributes), which may be either continuous or discrete. In this study, the class and the attributes or predictor variables remain the same that for the LDA model.

Decision tree algorithms are based on a divide-and-conquer approach and are also referred to as the so-called top-down induction of decision trees [60]. The algorithms work top-down, seeking at each stage an attribute that best separates the different classes, and then repeating recursively the process for each subset that results from the split. The most informative attribute is selected by introducing a function that assigns a value of the quality of the partition obtained by a specific attribute. Regarding continuous attributes, a threshold is determined, which splits the (sub)-tree into two ‘branches’, while with regard to discrete variables, ‘branches’ are created for each possible value of an attribute. The final subsets are called the ‘leaves’ and are labelled with a class.

Specifically, we used two classification trees implemented in Weka for the SN stochastic moment results. A random decision tree, which considers k randomly chosen attributes at each node and performs no pruning and a J48 decision tree. J48 algorithm [59] is an implementation of the C4.5 decision tree learner [61]. The algorithm for induction of decision trees uses the greedy search technique to induce decision trees for classification. C4.5 builds decision trees on the basis of the training data and then refines the tree to infer rules. Such an algorithm builds the complete tree and then refines it. J48 is similar to the C4.5 algorithm. However, once the tree is refined and pruned, no rules are inferred. We used the J48 and Random Tree default configurations which can be summarized as follows. For the J48 classification tree pruning is on and the technique used is sub-tree raising. The confidence threshold for pruning is 0.25, the minimum instances-per-leaf parameter is set to 2 and the number of folds, for the reduced error pruning value is set to 3. Regarding the random tree, the number of randomly chosen attributes is set to 1, the maximum depth of the tree is set to 0 for unlimited, and the minimum total weight of the instances in a leaf and the random number seed used for selecting attributes are set to 1.

The same training and external validation sets used for the LDA model were used in developing these classifiers. We also applied a 4-fold cross-validation based on training data to each classifier in order to consider another criterion in selecting the most predictive classifier. Here, a dataset having $n(47)$ instances is divided into $k(4)$ folds, where each fold has approximately n/k instances. The training and testing are done iteratively, in k iterations. In ith iteration, instances in all folds except the ith fold are used in the training phase and the instances from the ith fold are used for testing. Hence every instance from the dataset, is used exactly *once* as a training instance and $k-1$ times as a testing instance.

As used for the LDA model, the Accuracy, Sensitivity and Specificity values as well as the area under the receiving operating characteristic (ROC) curve were used as measures of performance. Other measures of performance were used additionally:

- *Number of True Positives* (TP): the number of examples (samples) classified as positive (cardiotoxic) and which are actually positive (correct classification);
- *Number of False Positives* (FP): the number of examples classified as positive and which are actually negative (non-cardiotoxic);
- *Number of True Negatives* (TN): number of examples classified as negative and which are actually negative (correct classification);
- *Number of False Negatives* (FN): number of examples classified as negative and which are actually positive.

Using the above-observed information, the following statistical measures were calculated:

- *Precision* [TP/(TP + FP)] - it gives what percentage of the examples predicted to be positive are actually positive.
- *Recall* [TP/(TP + FN)] - it gives what percentage of all positive examples were actually predicted to be positive by the algorithm.
- *F-Measure* [2*Precision*Recall/(Precision + Recall)] - it represents the relation that precision and recall should have. A high value (≈ 1) for precision with a low value (≈ 0) for recall is not suggestive. Similarly, high value for recall with a low value for precision does not mean much. For a perfect classifier, both precision and recall should be 1. A high value (≈ 1) of *F*-Measure implies a good classification.
- *Kappa Index* [TP/(TP + FN + FP)] - it is used to evaluate the agreement between predicted and observed nominal values in one dataset, while correcting for agreement that occurs by chance [62].

2.6. Proteome MS embedded star networks

The MS results have been transformed in the previous sections as a sequence of P and C letters. This sequence, that corresponds to the MS experiments, is similar to a protein sequence containing only two amino acid types and can be analyzed with the Star Network methods considering that each node is a letter (P or C) and the connections are between the nodes from the same group of characters (P or C). The star graph is the abstract representation of the network and this is a special case of trees with N vertices where one has got $N-1$ degrees of freedom and the remaining $N-1$ vertices have got one single degree of freedom [63]. For proteins, each of the 20 possible branches (“rays”) of the star contains the same amino acid type and the star centre is a non-amino acid vertex. In our case, the graphs will contain only two branches corresponding to the types of sequence character types (P and C). If the initial connectivity in the MS sequence is included, the graph is embedded (Fig. 3). In order to compare the graphs, it is necessary to transform the graphical representation in connectivity matrix, distance matrix and degree matrix. In the case of the embedded graph, the matrices of the connectivity in the sequence and in the star graph are combined. These matrices and the normalized ones are the base for the topological indices calculation.

2.7. Proteome MS embedded star networks TIs

The MS single-letter sequences are transformed into eSG TIs using S2SNet (Sequence to Star Networks) [64]. These calculations are characterized by non-weights, Markov normalization matrices and power of matrices/indices ($n = 0$–5). The summary file contains the following topological indices [65]:

Shannon Entropy of the n powered Markov Matrices [π_n(eSG)]:

$$\pi_n(\mathrm{eSG}) = \sum_i p_i{}^*\log(p_i), \tag{5}$$

where p_i are the n_i elements of the p vector, resulted from the matrix multiplication of the powered Markov normalized matrix ($n_i \times n_i$) and a vector ($n_i \times 1$) with each element equal to $1/n_i$.

The trace of the n connectivity matrices [π_n(eSG)]:

$$\pi_n(\mathrm{eSG}) = \sum_i (M^n)_{ii}, \tag{6}$$

where $n = 0$ – power limit, M = graph and sequence connectivity matrix (i^*i dimension); ii = ith diagonal element.

Harary number (H):

$$H = \sum_{i<j} (1/d_{ij}), \tag{7}$$

where d_{ij} are the elements of the distance matrix.

Wiener index (W):

$$W = \sum_{i<j} d_{ij}, \tag{8}$$

Gutman topological index (S_6):

$$S_6 = \sum_{ij} \mathrm{deg}_i{}^*\mathrm{deg}_j/d_{ij}, \tag{9}$$

where deg_i are the elements of the degree matrix.

Schultz topological index (non-trivial part) (S):

$$S = \sum_{i<j} \left(\mathrm{deg}_i + \mathrm{deg}_j\right)^* d_{ij}, \tag{10}$$

Balaban distance connectivity index (J):

$$J = (\mathrm{edges} - \mathrm{nodes} + 2)^* \sum_{i<j} m_{ij}{}^* \sqrt{\left(\sum_k d_{ik}{}^* \sum_k d_{kj}\right)}, \tag{11}$$

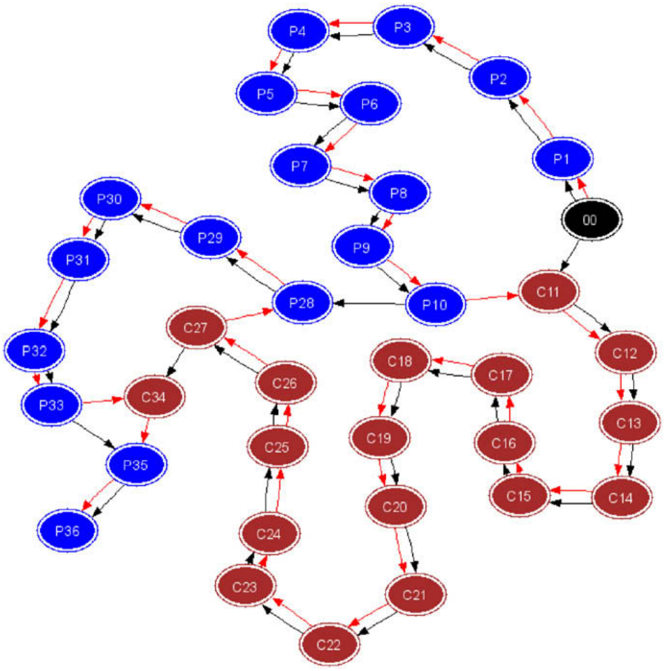


Fig. 3. Blood proteome MS embedded Star Network representation.

where $nodes + 1$ = AA numbers/node number in the Star Graph + origin, $\sum_k d_{ik}$ is the node distance degree.

Kier–Hall connectivity indices (^nX):

$$^0X = \sum_i 1/\sqrt{(\mathrm{deg}_i)}, \tag{12}$$

$$^2X = \sum_{i<j<k} m_{ij}{}^* m_{jk}/\sqrt{\left(\mathrm{deg}_i{}^*\mathrm{deg}_j{}^*\mathrm{deg}_k\right)}, \tag{13}$$

$$^3X = \sum_{i<j<k<m} m_{ij}{}^* m_{jk}{}^* m_{km}/\sqrt{\left(\mathrm{deg}_i{}^*\mathrm{deg}_j{}^*\mathrm{deg}_k{}^*\mathrm{deg}_m\right)}, \tag{14}$$

$$^4X = \sum_{i<j<k<m<o} m_{ij}{}^* m_{jk}{}^* m_{km}{}^* m_{mo}/\sqrt{\left(\mathrm{deg}_i{}^*\mathrm{deg}_j{}^*\mathrm{deg}_k{}^*\mathrm{deg}_m{}^*\mathrm{deg}_o\right)}, \tag{15}$$

$$^5X = \sum_{i<j<k<m<o<q} m_{ij}{}^* m_{jk}{}^* m_{km}{}^* m_{mo}{}^* m_{oq}/\sqrt{\left(\mathrm{deg}_i{}^*\mathrm{deg}_j{}^*\mathrm{deg}_k{}^*\mathrm{deg}_m{}^*\mathrm{deg}_o{}^*\mathrm{deg}_q\right)}, \tag{16}$$

Randic connectivity index (^1X):

$$^1X = \sum_{ij} m_{ij}/\sqrt{\left(\mathrm{deg}_i{}^*\mathrm{deg}_j\right)}, \tag{17}$$

These topological indices are used to construct LDA-QPTR based classification models in order to compare them with the same type of model, obtained with SN stochastic moments in this work and with the previous Lattice Network (LN) and SN Shannon entropy results [30]. We chose S2SNet in order to analyse sequences of transformed MS, in the same way that DRAGON [66] can analyse the small molecules, but cannot the mass spectra. Thus, we calculate Dragon like indices such are the Shannon entropies (π_n(eSG),

$n = 0–5$), the topological indices (W, H, S, S_6, J) and the connectivity indices (^{n}X, $n = 0–5$).

2.8. LDA-based classification model of eSG TIs

The same GDA method from STATISTICA has been chosen as the simplest and fastest method. In order to decide if a MS is NCT or CT, we added an extra dummy variable named *NCTorCT* (binary values of 0/1) and a cross-validation variable (CV). In the actual model, based on eSGs, the independent data test is used by splitting the data at random in a training series (*T*, 75%) used for a model construction and a prediction one (*P*, 25%); for model validation (the CV column is filled by repeating 3 *T* and 1 *P*). All independent variables are standardized prior to model construction. Using S2SNet methodology, as defined previously we can attempt to develop a simple linear QPTR, with the general formula, similar with Eq. (4):

$$\text{NCT/CT-score} = c_0 + \sum_{i=1 \to n} c_i{*}T_i, \tag{18}$$

where *NCT/CT-score* is the continue score value for the *NCT/CT* classification, T_i = all the eSG TIs described above, $c_1 - c_n$ = eSG TIs coefficients, n is the number for the indices and c_0 is the independent term. The GDA models' quality was determined by examining the same Wilks' statistic (U), Fisher ratio (F), p-level (p) and square Mahalanobis's distance (D^2). The *Best subset* model type was tested for the embedded TIs.

3. Results and discussion

3.1. Quantitative proteome–toxicity relationships

In the present work we propose the use of the graph theory combined with high-throughput mass spectrometry to the field of toxicoproteomics. In order to illustrate the potentialities of this approach, on the early detection of drug-induced cardiac toxicities, in the first step we decided to develop a QPTR based on π_k(SN), used here as numerical indices of the blood proteome mass spectrum Spiral network. The best LDA-based QPTR equation founded is described below:

$$\begin{aligned}\text{NCT/CT-score} = {} & 0.31 - 187.58{*}\pi_4(\text{SN}) + 940.64{*}\pi_6(\text{SN}) \\ & - 666.77{*}\pi_7(\text{SN}) - 380.05{*}\pi_8(\text{SN}) \\ & + 292.27{*}\pi_{10}(\text{SN})\end{aligned} \tag{19}$$

$$N = 47, F = 4.05, D^2 = 1.92, U = 0.67, p = 0.004.$$

This prediction model demonstrated an accuracy of 87.23% in classifying spectra coming from rat serum with overt cardiotoxicity and "non-cardiotoxic spectra". Specifically, 23 out of 26 CT samples and 18 out of 21 NCT samples were classified correctly, respectively (see Table 1 for details). The statistical significance of the model was evaluated through Fisher's test where F is the Fisher ratio and p represents the overall significance of the variables included in the model. Parsimony was tested by ρ value which is the ratio between the number of cases and the adjustable parameters. A value of ρ should be around 4 to discard any possibility of over-fitting. The same types of models and parameters have been used in previous works for the classical QSAR/QSPR models of small molecules [67] or for the QSPR studies of polymers [68]. Eq. (19) contains only π_4(SN), π_6(SN), π_7(SN), π_8(SN), π_{10}(SN) and not use π_0(SN), π_1(SN), π_2(SN), π_3(SN), π_5(SN), π_{10}(SN) as a result of the *Best subset* analysis algorithm, which select as more significant those predictors with the lowest CV misclassification rates (Table 2). All predictors used here have rates lower than 30%, which demonstrates the high significance of the predictors for all data subsets variations in CV with best subset analysis and high tolerance of the model to predictor variation.

Table 1
Classification matrices and performance of the LDA-based classification model on training and validation sets.

Model training			Model validation		
	NCT	CT		NCT	CT
NCT	18	3	NCT	5	2
CT	3	23	CT	2	6
Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
87.23%	88.46%	85.71%	73.33%	71.43%	75.00%

Additionally, the square of Mahalanobis's distance (D^2) and Wilks' U statistic provide a measure of the model's discriminatory power expressed through the relation between the intra- and inter-class variabilities and the separation between the centroids of each group, respectively. The small variation between the m/z and I values in data files generated from serum proteome mass spectra of cardiotoxic and non-cardiotoxic samples could be the cause of the non-ideal separation (100%) between the two groups. This is a logic result for an MS of a BP sample since the number of protein related to a toxic event is presumed to be insignificant in relation to the total number of serum proteins. However, we have discussed possible problems inherent to data generated with the present MS methodology [69,70]. They could affect the final validity of any kind of interpretation derived from it, independently of the discriminant power of the type of CIs, network, or statistical method used to fit the QPPR model. This fact, may also explain the existence of some misclassified cases. The non-ideal separation between the two groups should not be considered the consequence of the MS method because of the influence of other factors such as the type of graph representation, the class of topological indices, the classification algorithm and the animal population. This is a reasonable limitation of this kind of analysis but does not invalidate the results. In addition, the ROC curve [71] obtained, indicates that the model is not at random, but a statistically significant, classifier (see Fig. 4).

Table 2
The significance of the predictors for all data subsets variations in the LDA model (SN).

Subset no	CV Misclassification rate	No. of effects	π_4(SN)	π_6(SN)	π_7(SN)	π_8(SN)	π_{10}(SN)
1	23.40	4	in	in	in		in
2	25.53	5	in	in	in	in	in
3	25.53	3		in	in		in
4	29.79	2			in		in
5	29.79	2				in	in
6	29.79	2	in	in			
7	29.79	2	in		in		
8	29.79	2		in	in		
9	29.79	2	in			in	
10	29.79	2		in		in	

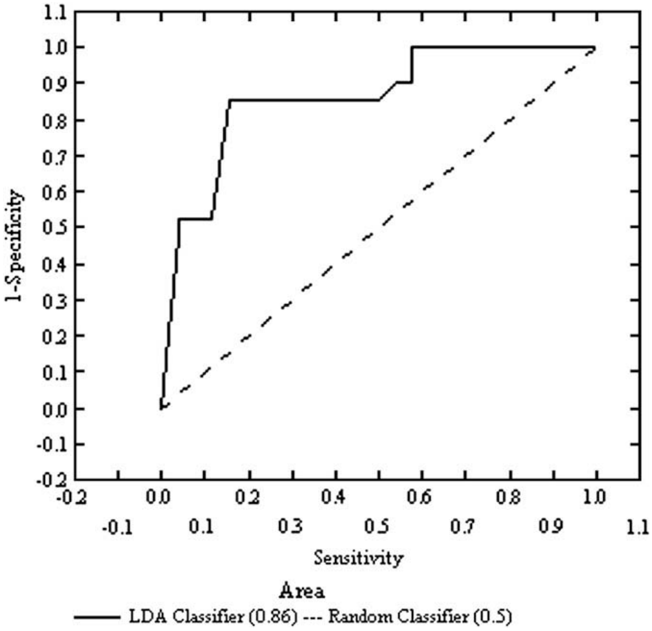


Fig. 4. Receiver operating characteristic curve (ROC curve) related to LDA model.

After all, the predictive ability of the model was assessed by using 15 samples never used for training. The proposed model was able to classify correctly 11 out of 15 samples (global predictability = 73.33%). In particular, 6 out of 8 CT samples (Sensitivity = 75%) and 5 out of 7 NCT samples (Specificity = 71.43%) were classified correctly. Next, it is necessary to find out if the basic assumptions of LDA are fulfilled [45,47] because in the case of severe violations, the reliability of the model's predictions could be compromised. The details about the statistical assumption can be found in Supplementary information section and in Fig. SM4 and Table SM5 from the supplementary material.

Finally, due to the limited data used in this work we must alert that the QPTR model developed is intended to prove the usefulness of the BPMSSNs and the numerical indices $\pi_k(\mathrm{SN})$ derived from this representation for toxicoproteomics studies. A dataset containing a higher number of samples could lead to really improved models, with a wider applicability domain. Specifically, the applicability domain of our model is limited by the number of instances (samples) used for training. A simple method to determine the applicability domain of a model is by plotting the standardized residuals *vs.* the leverages of the training instances [72]. The leverage (h) of an instance in the original variable space measures its influence on the model. The leverage of an instance h_i (see Eq. (21)) can be obtained from the respective diagonal elements of the hat matrix **H** (see Eq. (20)) [73].

$$\mathbf{H} = \mathbf{X}\cdot\left(\mathbf{X}\cdot\mathbf{X}^{\mathbf{T}}\right)^{-1}\cdot\mathbf{X}^{\mathbf{T}} \tag{20}$$

or

$$h_i = \vec{x}_i^{\,T}\left(\mathbf{X}\cdot\mathbf{X}^{\mathbf{T}}\right)^{-1}\vec{x}_i \quad (i = 1,\ldots,n) \tag{21}$$

where $\vec{x}_i$ is the descriptor vector of the considered instance and **X** is the model matrix derived from the training set descriptor values. The warning leverage h^* is defined as follows:

$$h^* = 3\times p'/n \tag{22}$$

where n is the number of training instances and p' is the number of model adjusting table parameters.

Fig. 5 shows the applicability domain of the LDA model, which is determined by training instances with h values lower than $h^* = 0.383$. The 15 new instances used for validation are also represented in the leverage plot shown in Fig. 6 in order to check whether they lie or not within the applicability domain of the model and consequently how reliable are the predictions. New instances such as that coming from the positive (cardiotoxic) sample 473 with an h value (0.392) higher than h^* and/or a value of standardized residual higher than 2 are out of the applicability domain of the model and consequently their predictions must be considered with caution.

In order to compare the values obtained with the Spiral network, we constructed several QPTR classification models using different classes of eSG TIs (Shannon entropies, spectral moments, topological indices and connectivity indices), all eSG TIs and the set of SN and eSG TIs (see Table 3). The best LDA-based QPTR equation based on all eSG TIs is described below:

$$\begin{aligned}\text{NCT/CT-score} = {} & 574.43 - 31979.38{*}\mathrm{Sh}_0 - 37576.67{*}\mathrm{Sh}_3 \\ & - 20431.66{*}\mathrm{Sh}_4 - 9337.56{*}\mathrm{Sh}_5 \\ & - 66125.67{*}\pi_4(\mathrm{eSG}) - 33353.95{*}\pi_5(\mathrm{eSG}) \\ & + 636.27{*}S_6 + 136.78{*}S + 51989.63{*}{}^0X \\ & + 12228.30{*}{}^2X - 2123.53{*}{}^4X\end{aligned} \tag{23}$$

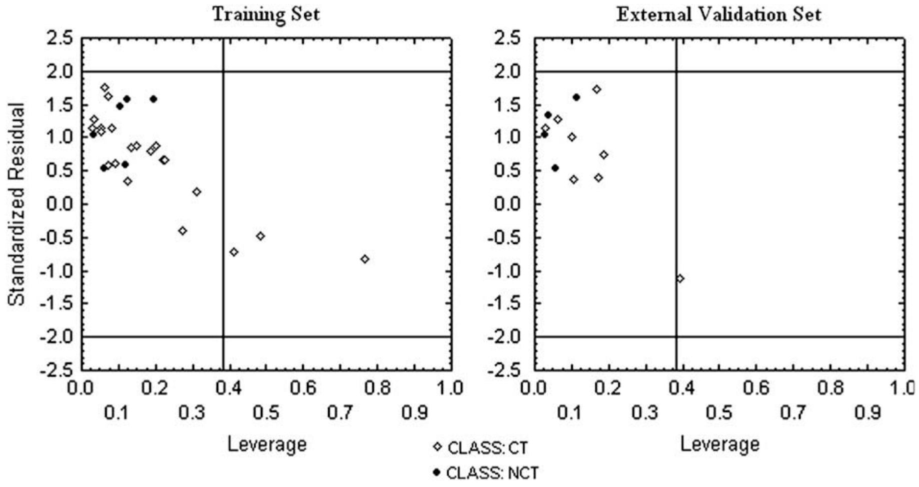


Fig. 5. Analysis of the Domain of Applicability of the LDA model.

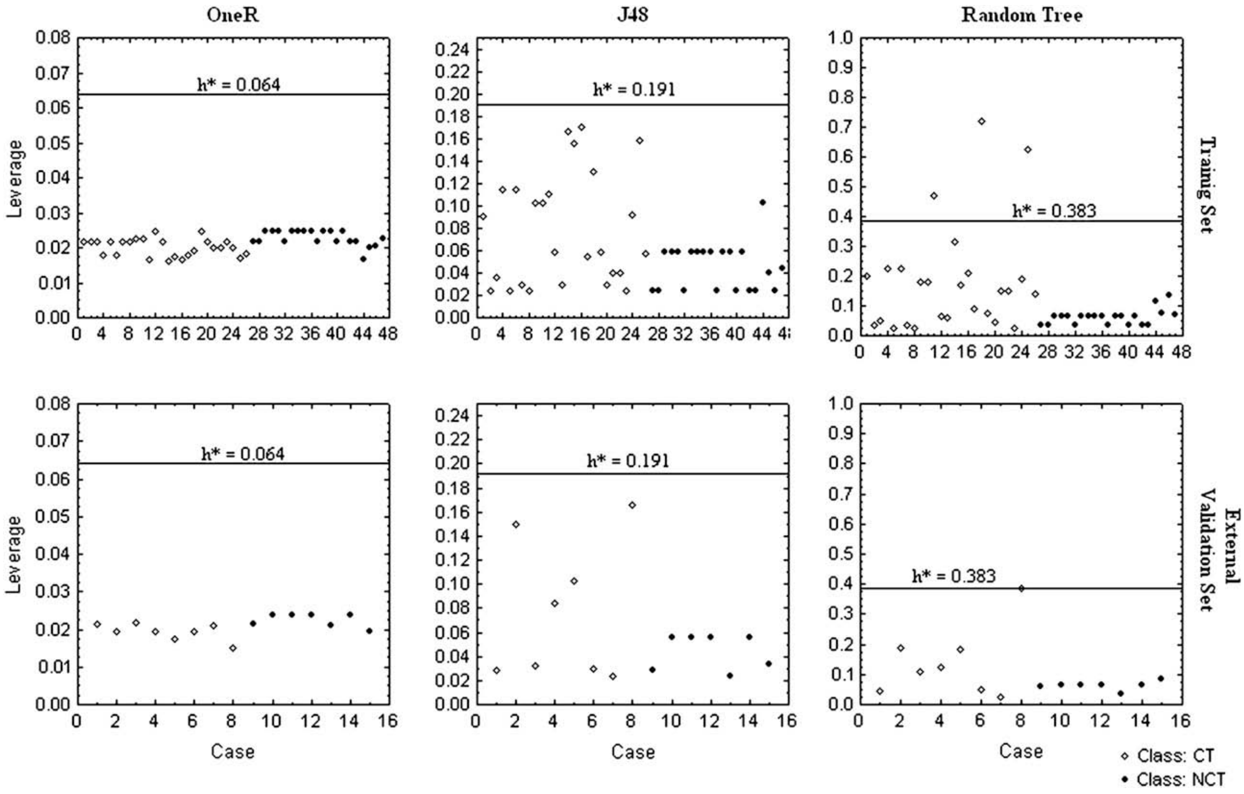


Fig. 6. Analysis of the Domain of Applicability of the three machine learning classification algorithms.

$$N = 62, F = 2.29, D^2 = 2.79, U = 0.58, p = 0.029$$

The model showed an Ac of 76.60%/86.67%, a Se of 65.38%/75.00% and an Sp of 90.48%/100% for the training/validation sets. Thus, the eSG models based on any TI class or all TIs have less quality (values less than 70%) than the SN best model, but better quality than the previous LN models (see Table 3) [30]. If we consider only the Shannon entropy models, the quality of the models increases in the following order: LN, eSG and SN. Using the spectral moments, the SN model is characterized by reasonable Ac/Se/Sp values (greater than 70%) opposite to the corresponding eSG model (less than 70%, even one value less than 40%). The combination of the SN spectral moments with all the eSG TIs produces a model of a better quality than in the case we are using only all the eSG TIs, but still with no reasonable Ac/Se/Sp values as in the case of the SN spectral moment model. These results proposed the SN spectral moments for the LDA models.

Until now we have demonstrated that the proposed SN LDA model is not a random classifier, is sufficiently accurate, has a satisfactory parsimony, and displays satisfactory robustness and predictivity. However, some LDA parametrical assumptions (normality and non-collinearity) are violated. Although the LDA is a robust multivariate technique and usually it is not affected by slight violations of these assumptions, we prefer to test the usefulness of the π_k(SN) numerical indices derived from blood proteome mass spectrum Spiral network for toxicoproteomics studies, by using other non-parametrical techniques like machine learning classification algorithms. In this sense, we tested three classification algorithms implemented in WEKA data mining system [58,59]. The first classification algorithm derived, and at the same time, the simplest was a OneR algorithm. The classification rule derived from several statistics of the performance of this rule is shown in Table 4. As it can be noted, this classification rule shows good performance measures in training, 4-fold cross-validation and external validation based on only one predictor variable, π_4(SN).

Table 3
Comparison of the LDA-based QPTR models for the LN, SN and eSG.

Class	Calculated indices	Graph	Calculation software	Dragon like indices	Model training			Model validation		
					Sp %	Se %	Ac %	Sp %	Se %	Ac %
Shannon entropies	π_k(LN), $k = 0$–10[a]	LN	MARCH-INSIDE	+	89.29	32.35	58.06	85.71	29.41	53.22
Shannon entropies	π_k(SN), $k = 0$–10[a]	SN	MARCH-INSIDE	+	100.00	70.59	83.87	89.29	67.65	77.42
Shannon Entropies	π_n(eSG), $n = 0$–5	eSG	S2SNet	+	90.48	61.54	74.47	85.71	37.50	60.00
Spectral moments	π_k(SN), $k = 0$–10	SN	MARCH-INSIDE	–	85.71	88.46	87.23	75.00	71.43	73.33
Spectral moments	π_n(eSG), $n = 0$–5	eSG	S2SNet	–	90.48	61.54	74.47	85.71	37.50	60.00
Topological indices	W, H, S, S_6, J	eSG	S2SNet	+	90.48	61.54	74.47	85.71	25.00	53.33
Connectivity indices	${}^nX$, $n = 0$–5	eSG	S2SNet	+	90.48	61.54	74.47	85.71	37.50	60.00
All eSG indices	π_n(eSG), π_n(eSG), W, H, S, S_6, J, $n = 0$–5	eSG	S2SNet	±	90.48	65.38	76.60	100.00	75.00	86.67
SN and eSG indices	π_k(SN), $k = 0$–10; π_n(eSG), π_n(eSG), W, H, S, S_6, J, $n = 0$–5	SN and eSG	MARCH-INSIDE and S2SNet	±	90.48	80.77	85.11	85.71	62.50	73.33

Note: LN = lattice network; SN = spiral network; eSG = embedded star graph; Sp = specificity; Se = selectivity; Ac = accuracy.
[a] Ref. [30].

Table 4
OneR classification algorithm and their respective measures of performance (SN).

One rule classification algorithm											
CLASS = CT IF π_4(SN) < 6.35825 or 6.62405 < π_4(SN) < 6.77775											
CLASS = NCT IF 6.35825 < π_4(SN) < 6.62405 or π_4(SN) ≥ 6.77775											
Performance details											
Training				*4-Fold cross-validation*				*External validation*			
Confusion matrix				Confusion matrix				Confusion matrix			
		CT	NCT			CT	NCT			CT	NCT
CT		20	6	CT		20	6	CT		7	1
NCT		1	20	NCT		6	15	NCT		2	5
Performance				Performance				Performance			
Accur.	Sensit.	Spec.	ROC area	Accur.	Sensit.	Spec.	ROC area	Accur.	Sensit.	Spec.	ROC area
85.1%	76.9%	95.2%	0.86	74.5%	76.9%	71.4%	0.74	80%	87.5%	71.4%	0.80
Prec.	Recall	*F*	Kappa	Prec.	Recall	*F*	Kappa	Prec.	Recall	*F*	Kappa
0.95	0.77	0.85	0.71	0.77	0.77	0.77	0.48	0.78	0.875	0.82	0.59

The second classification algorithm derived was based on a J48 decision tree. This is also a very simple tree with only three predictors variables [π_1(SN), π_4(SN) and π_7(SN)] used for the derivation of the classification rules. The J48 tree, constructed with only four nodes and five leaves which correspond to five points of classification decision (see Table 5), showed very good statistics on training, 4-fold cross-validation and external validation. Specifically, the sensitivity of this algorithm is excellent, especially on the external test set (100%). However, the specificity does not rise above the 71% on the external test set (see Table 5 for details).

Finally, the random tree shown in Table 6 demonstrates the best specificity among the three machine learning algorithms derived (100% on training and 85.7% on 4-fold cross-validation and external validation). Conversely, this tree is much more complex than the J48 and evidently than the OneR algorithm. Additionally, the performance on external test set contrasted to the performance on the training set is also much lower than the J48 and OneR algorithms, suggesting some overtraining. As shown, the three classification algorithms prove a good global performance (see Tables 4–6 for details). However, each one of the three classifiers stands out from the other two due to a particular property of the performance. The OneR algorithm is the simplest classifier, the best sensitivity is shown by the J48 classification tree, and the random tree is the most specific. Depending on the final goal, the user can take advantage of this disparity and use the most suitable classifier. We have to bear in mind the same considerations in the case of the applicability domain used for the LDA model.

Fig. 6 shows the leverage plot for the three classifiers for training and test instances carried out to establish the respective applicability domains. In this sense, the random tree is the only classifier not fully applicable to the external test set. As in the LDA model, the instance coming from the cardiotoxic sample 473 is slightly out of the applicability domain of the random tree classifier ($h =$ 0.387). Since the prediction of this instance should be considered cautiously, the global performance is affected. However, the instance out of the domain is positive and consequently the specificity is not affected.

Finally, the goal here is the early detection of drug-induced cardiac toxicity. Hence, the most important feature to consider is the sensitivity of the classifier since the consequences of

Table 5
J48 classification tree and their respective measures of performance (SN).

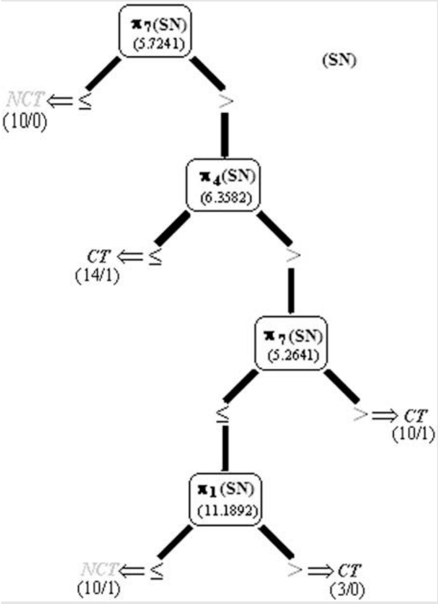
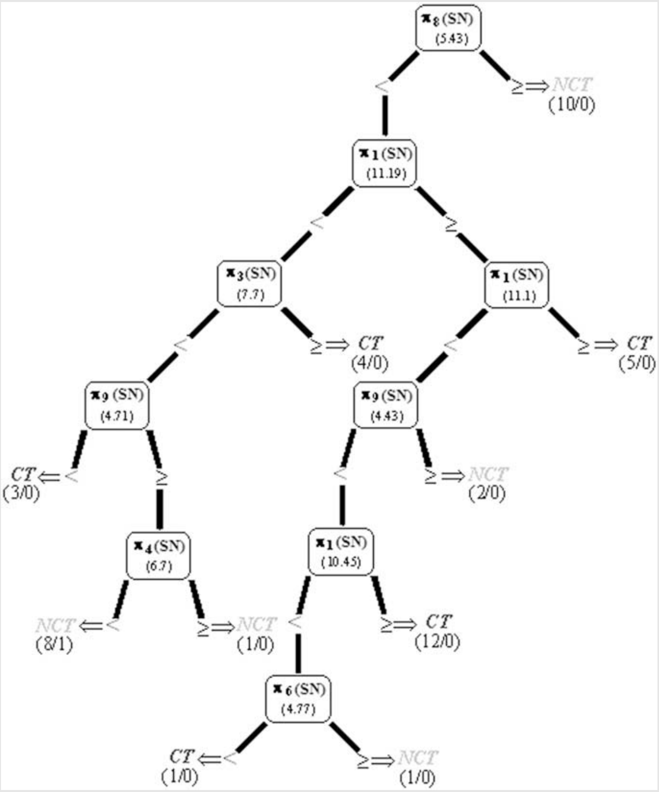
J48 classification tree	Performance details			
 π_7(SN) (5.7241) (SN) NCT ⇐ ≤ (10/0) > π_4(SN) (6.3582) CT ⇐ ≤ (14/1) > π_7(SN) (5.2641) ≤ > ⇒ CT (10/1) π_1(SN) (11.1892) NCT ⇐ ≤ (10/1) > ⇒ CT (3/0)	*Training*			
	Confusion matrix			
			CT	NCT
	CT		25	1
	NCT		2	19
	Performance			
	Accur.	Sensit.	Spec.	ROC area
	93.7%	96.2%	90.5%	0.95
	Prec.	Recall	*F*	Kappa
	0.93	0.96	0.94	0.87
	4-Fold cross-validation			
	Confusion matrix			
			CT	NCT
	CT		21	5
	NCT		5	16
	Performance			
	Accur.	Sensit.	Spec.	ROC area
	78.7%	80.8%	76.2%	0.78
	Prec.	Recall	*F*	Kappa
	0.81	0.81	0.81	0.57
	External validation			
	Confusion matrix			
			CT	NCT
	CT		8	0
	NCT		2	5
	Performance			
	Accur.	Sensit.	Spec.	ROC area
	86.7%	100%	71.4%	0.90
	Prec.	Recall	*F*	Kappa
	0.80	1.00	0.89	0.73

Table 6
Random classification tree and their respective measures of performance.

Random classification tree



Performance details											
Training				*4-Fold cross-validation*				*External validation*			
Confusion matrix				Confusion matrix				Confusion matrix			
		CT	NCT			CT	NCT			CT	NCT
CT		25	1	CT		21	5	CT		6	2
NCT		0	21	NCT		3	18	NCT		1	6
Performance				Performance				Performance			
Accur.	Sensit.	Spec.	ROC area	Accur.	Sensit.	Spec.	ROC area	Accur.	Sensit.	Spec.	ROC area
97.9%	96.2%	100%	0.99	83%	80.8%	85.7%	0.81	80%	75%	85.7%	0.79
Prec.	Recall	*F*	Kappa	Prec.	Recall	*F*	Kappa	Prec.	Recall	*F*	Kappa
1.00	0.96	0.98	0.96	0.88	0.81	0.84	0.66	0.86	0.75	0.80	0.60

misclassifying a CT sample (it classifies a CT sample as NCT) can be more devastating due to the potential risk to human life than in the opposite case. Accordingly, the J48 classifier is beginning to look like the most likely candidate. Other facts in support of this choice are the simplicity of the J48 tree contrasted to the random tree as well as the much better performance on the external test set contrasted to the random tree and the OneR classification algorithms. An exhaustive comparison of the measures of performance (Ac, Se, Sp, Kappa Index) of the three classifiers on training and validation experiments is shown in Table 7. The results of the comparison shown in this table fully justify the choice of the J48 classifier as the best option for the early detection of the drug-induced cardiac toxicity.

4. Concluding remarks

In this work, we defined new TIs, which are the stochastic spectral moments of the BPMSSNs, π_k(SN), and we compared them

Table 7
Comparison of the performance of the One Rule (OneR), J48 tree (J48) and Random tree (RT) classification algorithms on training, 4-fold cross-validation and external validation in the case of SN.

Performance measure	Training	4-Fold cross-validation	External validation
Accuracy	**RT** > J48 > OneR	**RT** > J48 > OneR	**J48** > OneR = RT
Sensitivity	**J48** = **RT** > OneR	**J48** = **RT** > OneR	**J48** > OneR > RT
Specificity	**RT** > OneR > J48	**RT** > J48 > OneR	**RT** > OneR = J48
Precision	**RT** > OneR > J48	**RT** > J48 > OneR	**RT** > J48 > OneR
Recall	**J48** = **RT** > OneR	**J48** = **RT** > OneR	**J48** > OneR > RT
F-Measure	**RT** > J48 > OneR	**RT** > J48 > OneR	**J48** > OneR > RT
Kappa index	**RT** > J48 > OneR	**RT** > J48 > OneR	**J48** > RT > OneR
ROC area	**RT** > J48 > OneR	**RT** > J48 > OneR	**J48** > OneR > RT
Simplicity	**OneR** > J48 > RT		

with several eSG TIs. In addition, the actual work results are compared with the LN and SN Shannon entropy models presented in a previous work. We showed the low quality results obtained with the eSG despite the good results in the case of protein sequences in the previous works. The comparison between the TI classes and network types shows the promotion of the SN stochastic moments for the graphical analyses of the blood proteome mass spectra, opposite to the eSG stochastic moments, eSG/SN/LN Shannon entropies, eSG topological indices and eSG connectivity indices. Thus, we demonstrated the potentialities of using the BPMSSNs proposed and the π_k(SN) indices derived from these graphs to the study of complex mixtures of biopolymers such as BP with special relevance to the field of toxicoproteomics. The method could be in principle extended to other mixtures of biopolymers.

Acknowledgements

The authors thank the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) (SFRH/BD/30698/2006 and SFRH/BPD/24997/2005) and the Xunta de Galicia (projects PXIB20304PR and BTF20302PR), the Ministerio de Sanidad y Consumo (project PI061457) for financial support. Gonzalez-Díaz H. acknowledges tenure-track research position funded by the Program Isidro Parga Pondal of the Xunta de Galicia. Gonzalez-Díaz H. also acknowledges additional financial support of Xunta de Galicia for one-year post-doctoral training under supervision of Prof. Kuo-Chen Chou at the Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA.

Appendix. Supplementary data

The computed values of the five π_k(SN) predictor variables included in the LDA-based QPTR model; observed and predicted classifications according to the LDA model; posterior probabilities to be classified as CT or NCT according to the LDA model; respective values of leverage and standardized residual (Std. Res.); and the distribution of the 62 samples used for training or model validation are depicted in Table SM1 of the supplementary material related to this paper. The computed values of the seven π_k(SN) predictor variables used for the three Machine Learning classification algorithms; observed and predicted classifications according to the OneR, J48 decision tree and Random decision tree classification algorithms, respectively; and the distribution of the 62 samples used for training or model validation are depicted in Table SM2. The values of leverage related to the OneR, J48 decision tree and Random decision tree classification algorithms, respectively for the 62 samples used are depicted in Table SM3.

As the names imply, LDA establishes a linear, additive relationship between the predictive variables and the response variable and indeed, this is the simplest functional form to adopt with no prior information. Visual inspection of the distribution of the standardized residuals for all drugs (standardized residuals *vs.* cases; see section A in Fig. SM4) supports this choice, as no systematic pattern is seen [74]. When we checked the parametric assumption of multivariate normality of residuals, it was found that the residuals do not exhibit adequate values of skewness and kurtosis [47], which is a sign of deviation from normal distribution. The hypothesis of multivariate normality of residuals is rejected according to the three Kolmogorov–Smirnov, Shapiro–Wilks and Lilliefors hypothesis tests applied (statistic values in Table SM5). The frequency histograms shown in section B of figure. SM4 confirms visually the violation of the normality assumption. In addition, as the term related to the error (represented by residuals) is not included in the LDA equation, the mean must be 0. Actually, the residual mean value for our model is close to the assumed value of 0 (see Table SM5).

Moving on to the next important parametric assumption of LDA, *i.e.* homocedasticity (*i.e.*: homogeneity of variance of the variables) was also checked by simply plotting the square standardized residuals for each predictor variable [47]. The plots in section C of Fig. SM4 reveal an adequate scatter on the points, without any consistent pattern, validating the assumption of homocedasticity. In addition, the variables included in the model exhibit a high collinearity (pair correlation between one or more than one variables higher than 0.7). As a consequence, the common interpretation of a regression coefficient, as measuring the change in the expected value of the response variable when the given predictor variable is increased by one unit while all other predictor variables are held constant, is not fully applicable when multicollinearity exists. However, the fact that some or all predictor variables are correlated among themselves neither, in general, inhibits the model's ability to obtain a good fit nor it tends to affect inferences about the mean responses or predictions of new observations [75]. Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.polymer.2008.09.070.

References

[1] Kholodovych V, Gubskaya AV, Bohrer M, Harris N, Knight D, Kohn J, et al. Polymer 2008;49:2435–9.
[2] Morrill JA, Jensen RE, Madison PH, Chabalowski CF. J Chem Inf Comput Sci 2004;44(3):912–20.
[3] Cao C, Lin Y. J Chem Inf Comput Sci 2003;43(2):643–50.
[4] Mattioni BE, Jurs PC. J Chem Inf Comput Sci 2002;42(2):232–40.
[5] (a) Cruz VL, Martinez J, Martinez-Salazar J, Ramos J, Reyes ML, Toro-Labbe A, et al. Polymer 2007;48:7672–8;
(b) Chou, KC; J Biol Chem 1989;264:12074–9;
(c) Chou, KC. Biophysical Chemistry 1990;35:1–24.
[6] Mandloi M, Sikarwar A, Sapre NS, Karmarkar S, Khadikar PV. J Chem Inf Comput Sci 2000;40(1):57–62.
[7] Marrero-Ponce Y, Medina-Marrero R, Castillo-Garit JA, Romero-Zaldivar V, Torrens F, Castro EA. Bioorg Med Chem 2005;13(8):3003–15.
[8] Marrero-Ponce Y, Nodarse D, González-Díaz H, Ramos de Armas R, Romero-Zaldivar V, Torrens F, et al. Int J Mol Sci 2004;5:276–93.
[9] Gonzalez-Diaz H, Gonzalez-Diaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics 2008;8(4):750–78.
[10] González-Díaz H, Vilar S, Santana L, Uriarte E. Curr Top Med Chem 2007; 7(10):1025–39.
[11] Bonchev D, Buck GA. J Chem Inf Model 2007;47(3):909–17.
[12] González-Díaz H, Molina RR, Uriarte E. Polymer 2003;45:3845–53.
[13] González-Díaz H, Saíz-Urra L, Molina R, Uriarte E. Polymer 2005;46: 2791–8.
[14] González-Díaz H, Pérez-Bello A, Uriarte E. Polymer 2005;46:6461–73.
[15] Liotta LA, Ferrari M, Petricoin E. Nature 2003;425(6961):905.
[16] Hu S, Loo JA, Wong DT. Proteomics 2006;6(23):6326–53.
[17] McDonald WH, Yates 3rd JR. Dis Markers 2002;18(2):99–105.
[18] Bartels C. Biomed Environ Mass Spectrom 1990;19:363–8.
[19] Fernandez-de-Cossio J, Gonzalez J, Besada V. Comput Appl Biosci 1995;11(4): 427–34.
[20] Taylor JA, Johnson RS. Rapid Commun Mass Spectrom 1997;11:1067–75.
[21] Dancík V, Addona TA, Clauser KR, Vath JE, Pevzner PA. J Comput Biol 1999;6 (3/4):327–42.
[22] Frank A, Pevzner P. Anal Chem 2005;77:964–73.
[23] van Dalen EC, van den Brug M, Caron HN, Kremer LC. Eur J Cancer 2006;42(18):3199–205.
[24] Jones RL, Ewer MS. Expert Rev Anticancer Ther 2006;6(9):1249–69.
[25] Urbanova D, Urban L, Carter A, Maasova D, Mladosievicova B. Neoplasma 2006;53(3):183–90.
[26] Anderson NL, Anderson NG. Mol Cell Proteomics 2002;1(11):845–67.
[27] Petricoin EF, Rajapaske V, Herman EH, Arekani AM, Ross S, Johann D, et al. Toxicol Pathol 2004;32(Suppl. 1):122–30.
[28] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Lancet 2002;359(9306):572–7.
[29] Petricoin 3rd EF, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. J Natl Cancer Inst 2002;94(20):1576–8.
[30] Cruz-Monteagudo M, Gonzalez-Diaz H, Borges F, Dominguez ER, Cordeiro MN. Chem Res Toxicol 2008;21(3):619–32.
[31] Randic M, Lers N, Plavsic D, Basak SC, Balaban AT. Chem Phys Lett 2005;407(1):205–8.
[32] Randic M, Zupan J, Vikic-Topic D. J Mol Graphics Modell 2007:290–305.
[33] Munteanu CR, González-Díaz H, Borges F, Magalhães AL. J Theor Biol 2008; 254(4):775–83.
[34] Munteanu CR, Gonzalez-Diaz H, Magalhaes AL. J Theor Biol 2008;254(2):476–82.

[35] Zhang J, Herman EH, Knapton A, Chadwick DP, Whitehurst VE, Koerner JE, et al. Toxicol Pathol 2002;30(1):28–40.

[36] Ferino G, Gonzalez-Diaz H, Delogu G, Podda G, Uriarte E. Biochem Biophys Res Commun 2008. doi:10.1016/j.bbrc.2008.1005.1071.

[37] González-Díaz H, Uriarte E. Biopolymers 2005;77(5):296–303.

[38] González-Díaz H, Molina-Ruiz R, Hernandez I. **MARCH-INSIDE** version 3.0 (**MAR**kov **CH**ains **IN**variants for **SI**mulation and **DE**sign); 2007. email: gonzalezdiazh@yahoo.es.

[39] Ramos de Armas R, González-Díaz H, Molina R, Uriarte E. Proteins 2004; 56(4):715–23.

[40] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A. J Comput Chem 2007;28(6):1042–8.

[41] Gonzalez-Diaz H, Saiz-Urra L, Molina R, Santana L, Uriarte E. J Proteome Res 2007;6(2):904–8.

[42] Gonzalez-Diaz H, Perez-Castillo Y, Podda G, Uriarte E. J Comput Chem 2007;28(12):1990–5.

[43] Van Waterbeemd H. In: Van Waterbeemd H, editor. Discriminant analysis for activity prediction. Chemometric methods in molecular design, vol. 2. New York: Wiley-VCH; 1995. p. 265–82.

[44] STATISTICA. Statsoft Inc.; 2001.

[45] Van Waterbeemd H. Chemometric methods in molecular design. New York: Wiley-VCH; 1995.

[46] Cruz-Monteagudo M, Gonzalez-Diaz H, Aguero-Chapin G, Santana L, Borges F, Dominguez ER, et al. J Comput Chem 2007;28(11):1909–23.

[47] Stewart J, Gill L, editors. Econometrics. 2nd ed. London: Prentice Hall; 1998.

[48] Kutner MH, Nachtsheim CJ, Neter J, Li W. Standardized multiple regression model. Applied linear statistical models. New York: McGraw Hill; 2005. p. 271–7.

[49] Chou KC, Zhang CT. Crit Rev Biochem Mol Biol 1995;30(4):275–349.

[50] Chou KC, Shen HB. Anal Biochem 2007;370(1):1–16.

[51] Chou KC, Shen HB. Nat Protoc 2008;3:153–62.

[52] Chen YL, Li QZ. J Theor Biol 2007;248:377–81.

[53] Chen YL, Li QZ. J Theor Biol 2007;245(4):775–83.

[54] Diao Y, Li M, Feng Z, Yin J, Pan Y. J Theor Biol 2007;247(4):608–15.

[55] Lin H. J Theor Biol 2008;252(2):350–6.

[56] Niu B, Cai YD, Lu WC, Li GZ, Chou KC. Protein Pept Lett 2006;13(5):489–92.

[57] Xiao X, Chou KC. Protein Pept Lett 2007;14(9):871–5.

[58] Waikato environment for knowledge analysis (WEKA). New Zealand: University of Waikato; 2005.

[59] Witten IH, Frank E. Data mining: practical machine learning tools and techniques. 2nd ed. San Francisco; 2005.

[60] Quinlan JR. Mach Learn 1986;1(1):81–106.

[61] Quinlan JR. C4.5: programs for machine learning. San Francisco: Morgan Kaufmann; 1993.

[62] Cohen J. Educ Psychol Meas 1960;20(1):37–46.

[63] Harary F. Graph theory. MA; 1969.

[64] Munteanu CR, Gonzáles-Diáz H. S2SNet – Sequence to Star Network, Reg. No. 03/2008/1338. Santiago de Compostela, Spain; 2008.

[65] Todeschini R, Consonni V. Handbook of molecular descriptors. Wiley-VCH; 2002.

[66] Todeschini R, Consonni V, Pavan M. Dragon software version 2.1; 2002.

[67] Morales AH, Rodríguez-Borges JE, García-Mera X, Fernández F, Dias-Sueiro-Cordeiro MN. J Med Chem 2007;50:1537–45.

[68] Marrero-Ponce Y, Medina-Marrero R, Castro AE, Ramos de Armas R, González-Díaz H, Romero-Zaldivar V, et al. Molecules 2004;9:1124–47.

[69] Ransohoff DF. J Natl Cancer Inst 2005;97:315–9.

[70] Baggerly KA, Morris JS, Edmonson SR, Coombes KR. J Natl Cancer Inst 2005; 97:307–9.

[71] Zweig MH, Broste SK, Reinhart RA. Clin Chem 1992;38(8 Pt 1):1425–8.

[72] Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Environ Health Perspect 2003;111(10):1361–75.

[73] Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models. 5th ed. New York: McGraw Hill; 2005.

[74] Stewart J, Gill L. Econometrics. London; 1998.

[75] Kutner MH, Nachtsheim CJ, Neter J, Li W. Multicollinearity and its effects. Applied linear statistical models. New York: McGraw Hill; 2005. p. 278–89.