Bioorganic & Medicinal Chemistry 16 (2008) 9684–9693

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

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

Quantitative Proteome–Property Relationships (QPPRs). Part 1: Finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra

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

[a] Unit of Bioinformatics & Connectivity Analysis (UBICA), Institute of Industrial Pharmacy, Faculty of Pharmacy, Department of Organic Chemistry, 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, University of Porto, 4150-047, Porto, Portugal

[d] REQUIMTE/Science Faculty, Chemistry Department, University of Porto, 4169-007, Porto, Portugal

ARTICLE INFO

Article history:
Received 15 April 2008
Revised 29 September 2008
Accepted 2 October 2008
Available online 5 October 2008

Keywords:
QSAR
Complex network theory
Markov model
Mass spectroscopy
Biomarker
Proteomics
Toxicoproteomics
Quantitative Proteome–Property Relationship
Blood proteome
Drug toxicity

ABSTRACT

Numerical parameters of the molecular networks, also referred as Topological Indices or Connectivity Indices (CIs), have been used in Bioorganic and Medicinal Chemistry to find Quantitative Structure–Activity, Property or Toxicity Relationship (QSAR, QSPR and QSTR) models. QSPR models generally use CIs as inputs to predict the biological activity of compounds. However, the literature does not evidence a great effort to find QSAR-like models for other biologically and chemically relevant systems. For instance, blood proteome constitutes a protein-rich information reservoir, since the serum proteome Mass Spectra (MS) represents a potential information source for the early detection of Biomarkers for diseases and/or drug-induced toxicities. The concept of mass spectrum network (MS network) for a single protein is already well-known. However, there are no reported results on the use of CIs for a MS network of a whole proteome to explore MS patterns. In this work, we introduced for the first time a novel network representation and the CIs for the MS of blood proteome samples. The new network bases on Randic's Spiral network have been previously introduced for protein sequences. The new MS CIs, called here Spiral Markov Connectivity (SMC_k) of the MS Spiral graph can be calculated with the software MARCH-INSIDE, combining network and Markov model theory. The SMC_k values could be used to seek QSAR-like models, called in this work Quantitative Proteome–Property Relationships (QPPRs). We calculate the SMC_k values for 62 blood samples and fit a QPPR model by discriminating proteome MS, typical of individuals susceptible to suffer drug-induced cardiotoxicity from control samples. The accuracy, sensitivity, and specificity values of the QPPR model were between 73.08% and 87.5% in training and validation series. This work points to QPPR models as a powerful tool for MS detection of biomarkers in proteomics.

1. Introduction

A very convenient and natural way of representing the relationships between objects is the use of networks or graphs, where objects are represented by nodes and the relationship between them by edges. Structures that can be represented as networks are ubiquitous, and many problems of practical interest can be represented by networks.[1] In this sense, the networks or the graph theory have been widely used in Bioorganic Medicinal Chemistry and Bioinformatics to study chemical and biological systems. Specifically, the studies of the Quantitative Structure–Activity, Property or Toxicology Relationships (QSAR, QSPR or QSTR) constitute one of the most popular and representative methods based on the network theory.[2] In QSAR-like methods, the molecules under study are represented as networks where atoms play the role of nodes and chemical bonds act as edges.[3,4] Next, different numerical indices or parameters representative of the structure (connectivity) of the network (molecule) can be calculated. These parameters are often referred as Topological Indices (TIs) or Connectivity Indices (CIs). There are numerous works developed until today related to the use of QSAR methods, which intended to confer rationality to the drug development process.[5,6] However, most of these works are mainly based on TIs or CIs of small-sized molecules.[7] Many authors, such as Balaban, Basak, Caballero, Cai, Chou, Estrada, Fernández, Nandy, Liao, Leong, Randic, Vilar and others have recently worked very hard on the definition, implementation of computing algorithms, and the practical use of indices and network representations for proteins, RNA and DNA sequences at the macro-molecular and supra-molecular level.[8–32]

[*] Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912.
E-mail addresses: humberto.gonzalez@usc.es, gonzalezdiazh@yahoo.es (H. González-Díaz).

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.10.004

In any case, TIs or CIs derived from networks/graphs of systems larger than small-sized molecules are also a representation of molecular structures. In this sense, it is of major relevance the extremely important work that Randic and others have carried out in order to define CIs for Proteomics maps.[33–43] This fact evidences even less effort in finding CIs and QSPR-like relationships between other biologically and chemically relevant sources of information, such as whole proteome Mass Spectra (MS), DNA microarrays, Proteomics electrophoresis maps and others. This point has been recently discussed by Bonchev in a very interesting work.[44] The integration of these different experimental techniques, with network theory and other theoretical tools, opens an interesting way for the search of Biomarkers for disease prognosis.[45–53] In two recent reviews we have made a detailed discussion on the uses of network CIs at the molecular, macro-molecular, and supra-molecular level, with applications in Bioorganic and Medicinal Chemistry and Proteomics.[54,55] Nevertheless, specifically the definition and search of whole proteome MS CIs and QSAR-like models even promising has remained as a field not very explored yet.

The application of network theory to mass spectrometry was first proposed by Bartels for peptide sequencing.[56] The basic idea consists in transforming a mass spectrum into a network called "*spectrum graph*". Basically, each peak in the experimental spectrum is represented as a vertex (or several nodes) in the spectrum network and a directed edge is established between two nodes if the mass difference of the two nodes equals the mass of one or several amino acids. Several algorithms that make use of spectrum networks have been designed for de novo peptide sequencing. Among the most popular are: "SeqMS"ra[57], "Lutefisk"ra[58], "Sherenga"ra[59] and more recently "PepNovo"ra.[60] The construction of the spectrum network of all these algorithms shares the basic idea proposed by Bartels, with its respective particularities, but it is limited to a single protein structure. Consequently, they are also molecular structure networks and, in their present form, they cannot constitute whole proteome networks, but a study of whole proteome with MS networks is initiated.

There are many proteome samples susceptible to be studied with whole proteome MS networks and CIs. More, blood proteome is of major interest due to different reasons. 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.[61,62] The blood proteome is changing constantly as a consequence of the perfusion of the organ undergoing drug-induced damage and then, this process then 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.[63] Consequently, a blood proteome 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[63] and the optimal performance in the low-mass range exhibited by mass spectroscopy,[64,65] this method applied to Proteomics may offer the best chance to discover these early stage changes.

However, due to thousands of intact and cleaved proteins in the human serum proteome, finding the single disease-related protein biomarker 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 has occurred. For these reasons, identifying patterns by using the serum proteome spectrum instead of directly identifying a single marker candidate represents a more attractive and realistic choice for this purpose. In this sense, Petricoin et al. successfully identified patterns of biomarkers of low-molecular weight as ion peak features within the MS, and used these patterns as the diagnostic endpoint itself for the early detection of drug-induced cardiac toxicities,[66] ovarian[67] and prostate cancer.[68]

On the other hand, González-Díaz et al. have introduced a QSAR-like methodology referred to as MARCH-INSIDE (*MAR*kovian *CH*emicals *IN*silico *DE*sign). In previous, work we have extended MARCH-INSIDE to calculate CIs not only for small molecules or biopolymer structural networks, but also for more complex networks of larger biological systems including whole proteomes. It determined us to rename this methodology as *MAR*kov *CH*ains *IN*variants for *SI*mulation & *DE*sign, which allows us to keep the same acronym MARCH-INSIDE, but adapting it better to the new applications. The new MARCH-INSIDE also focuses on CIs calculated as numerical parameters of a stochastic matrix associated with the network of the system. The elements of this matrix are the probabilities that two parts of the system (network nodes) participate in some kind of interaction or connection (network edges). The systems studied with MARCH-INSIDE included small molecules, DNA sequences, RNA secondary structures, protein sequences and 3D structures, and viral surfaces. In these cases, the atoms, nucleotides, or amino acids played the role of nodes and the nodes pair-wise connections (edges) were represented by chemical bonds, electrostatic, van der Waals, vibrations, virus surface adjacency and others. MARCH-INSIDE is based on Markov Chains theory so we can use the Chapman–Kolgomorov equations to calculate the stochastic matrices of long-range interactions in the system. Next, we can derive from these matrices many different CIs of the network, which numerically characterize the structure of the system. Despite the high flexibility of MARCH-INSIDE, it has not been applied yet to the characterization of whole proteomes using MS Spiral networks and searching for QPPRs.[55]

In the present work we have decided to identify serum proteome cardiac toxicity patterns and apply them to find a predictive QSPR-like model using a network theoretical representation and CIs codification of MS, instead of directly identifying patterns within the MS. Specifically, we propose an alternative network theoretical representation of the blood proteome MS, based on Spiral networks introduced by Randic et al. for DNA sequences.[69] Next, using the MARCH-INSIDE approach we derive whole proteome MS stochastic CIs from the new representation. These numerical indices were named here the Spiral Markov Connectivity (SMC_k or Γ) indices of order k. They account for connectivity in Spiral networks using Markov Chains to codify long-range connectivity patterns between nodes placed at distance k. The SMC_k indices are then used as inputs in the derivation of a Quantitative Proteome–Toxicity Relationship (QPTR) model, which may be classified into a broader class referred as Quantitative Proteome–Property Relationships (QPPRs). QPPRs based on Spiral MS CIs are proposed as a new alternative for the early detection of drug-induced cardiac toxicities. A schematic representation of the approach proposed in this work is shown in Figure 1.

2. Methods

2.1. Serum proteome mass spectrum data set

For the generation of the serum proteome MS Spiral networks 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.[66] According to these authors, the data files are generated by first exporting the raw data file generated from the QSTAR mass spectra into tab-delimited files that generated approximately 350,000 data points per spectrum. The binning process condenses the number of data

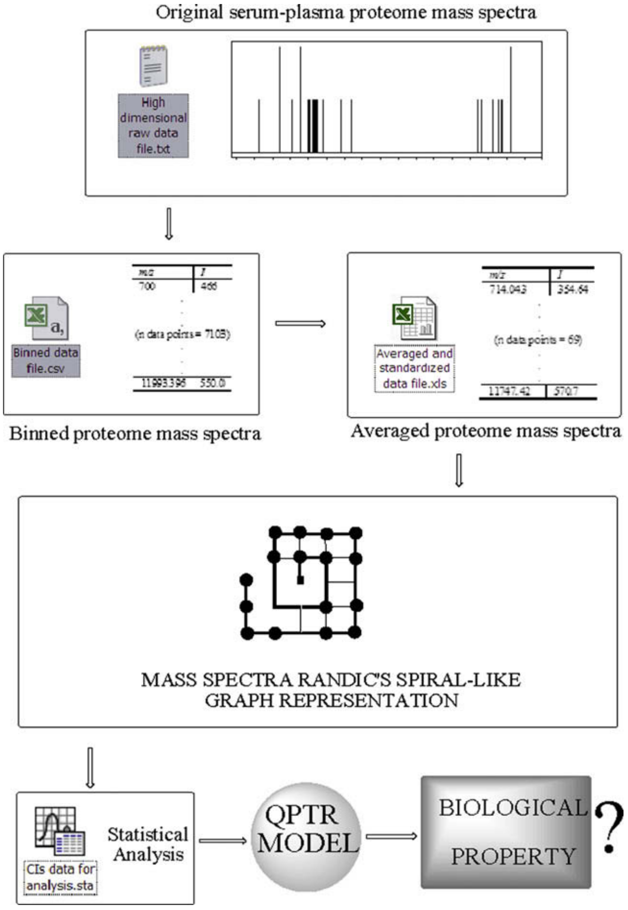
Original serum-plasma proteome mass spectra
High dimensional raw data file.txt
m/z
I
700
466
(n data points = 7103)
11993.396
550.0
Binned data file.csv
Binned proteome mass spectra
m/z
I
714.043
354.64
(n data points = 69)
11747.42
570.7
Averaged and standardized data file.xls
Averaged proteome mass spectra
MASS SPECTRA RANDIC'S SPIRAL-LIKE GRAPH REPRESENTATION
CIs data for analysis.sta
Statistical Analysis
QPTR MODEL
BIOLOGICAL PROPERTY
?

Figure 1. Schematic representation of the steps given in this work.

points to 7105 points per sample. The high-resolution spectra are binned using a function of 400 parts per million (ppm) such 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).[66]

Using the Spontaneously Hypertensive Rat (SHR) model, in which animals were challenged with doxorubicin or with mitoxantone ± 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.[70–73] Since the cardiac toxicity profile of 141 out of 203 samples analyzed was reported as unknown or with no definitive information about their cardiotoxic profile, only 62 samples were used in this work:

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

2.2. Spiral Markov Connectivity (SMC_k) parameters of serum proteome MS-based networks

In order to generate MS Spiral networks, we used high-dimensional data produced by high-throughput mass spectrometry consisting of binned data files derived from raw data files, which were generated from serum proteome mass spectra.[66] Although the binned process reduces efficiently the number of data points, it is still unmanageable for network generation. Hence, the number of data points in the binned data files was condensed to 36 by including in each new data point the averaged m/z and I values of 200 consecutive data points. Each new data point condenses now the information encoded on 200 binned data points. Due to the number of data points in the binned data files, the last data point was generated by using the last 205. Considering the successive transformations applied to the raw data (binning and averaging processes) all the averaged m/z and I values were replaced by their respective standardized values. The values were standardized in order to bring all of them (regardless of their distributions and original units of measurement) to compatible units from a distribution with a mean of 0 and a standard deviation of 1. Standardization also makes the results entirely independent of the ranges of values or the units of measurements. Finally, the averaged and standardized m/z and I values were multiplied in order to obtain a value for the relationship between m/z and I that makes possible a network representation.

Next, a new averaged and standardized data file is generated consisting of 36 data points which can be used now to generate a serum proteome mass spectrum network by using a spiral representation. Since the values of m/z and I were standardized, the mean value of the $<I_i> = m/z \cdot I$ of each sample is around 0.5. Consequently, a cutoff value of 0.5 is chosen for the $<I_i>$ values. This cutoff value is used to codify each data point according to their respective average $<I_i>$ values allowing their representation as a node on a two-dimensional (2D) space. Such a spiral network is obtained in an analogue way to the four-color maps introduced by Randic et al. for DNA sequences representation.[69] Specifically, in this work we represented the MS data points as a Spiral composed by nodes which are differently labeled; that is, if $<I_i>$ is lower than 0.5 then the node is labeled with the letter C; otherwise is labeled with a P. The MS Spiral network begins with the first averaged and standardized MS data point related to the lower m/z region, runs clockwise, and finishes with the higher m/z region. Next, by connecting the adjacent nodes equally labeled we obtain networks similar to the MS Spiral networks shown in Figure 2. This network picture can be drawn directly from the interface of the software MARCH-INSIDE, used to calculate the CIs values.[74] Two nodes are considered adjacent only if they are at one step away one from each other in the Cartesian space (Euclidean distance is equal to 1). Connections are allowed only in ordinate and abscissa directions are allowed. Diagonal connections are not allowed and consequently, diagonal nodes are not considered adjacent. As a result, a segment of nodes labeled with two different labels is obtained. Different labeling of nodes confers to each MS network a particular topology or connectivity allowing their numerical (topological) characterization with CIs depending on the values of m/z and I of the specific MS for each blood proteome sample.

In the following step, we used a Markov model (MM) to codify the information about serum proteome MS regions. Specifically, in this work we introduced the CIs for the MS Spiral network denoted SMC_k. The SMC_k values were derived from the MARCH-INSIDE approach mentioned above.[74] The MARCH-INSIDE approach has been previously applied to the field of proteins.[75,76] MARCH-INSIDE has been used here for the first time to codify the information content encoded in a serum proteome MS. The classic matrix MARCH-INSIDE approach[77] has been adapted to characterize the new spiral networks. The method uses essentially two matrix magnitudes:

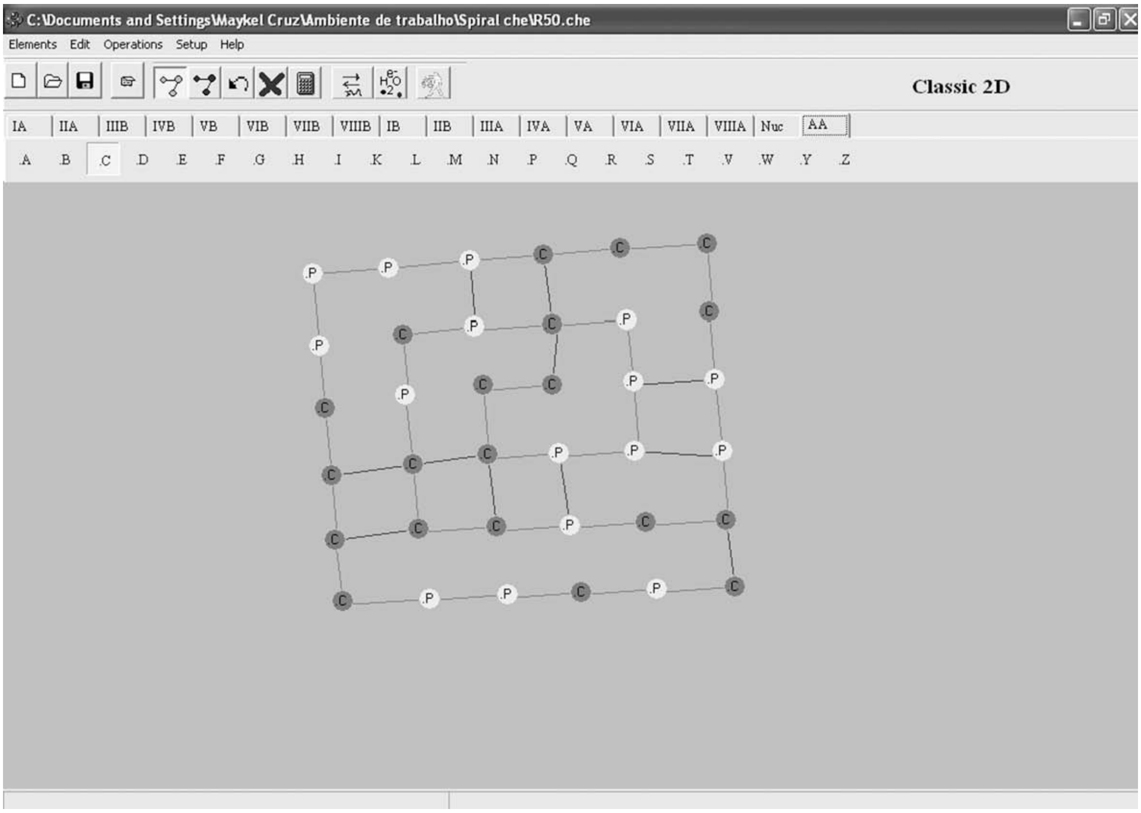
C:\Documents and Settings\Maykel Cruz\Ambiente de trabalho\Spiral che\R50.che
Elements Edit Operations Setup Help
Classic 2D
IA IIA IIIB IVB VB VIB VIIB VIIIB IB IIB IIIA IVA VA VIA VIIA VIIIA Nuc AA
A B C D E F G H I K L M N P Q R S T V W Y Z
P
C

Figure 2. MARCH-INSIDE software visualization of a blood proteome MS Spiral network.

(a) The matrix $^{1}\mathbf{\Pi}$ (see Eq. 1). This matrix is built up as a square matrix ($n \times n$). The matrix $^{1}\mathbf{\Pi}$ contains the probabilities $^{1}p_{\mathrm{ij}}$ to reach a node $\mathrm{n_i}$ moving throughout a walk of length $k = 1$ from a node n_{j}:

$$1p_{ij} = \frac{\alpha_{ij} \cdot o_{nj}}{\sum_{l}^{n} \alpha_{il} \cdot o_{nl}} \tag{1}$$

Where, α_{ij} or α_{il} represents the adjacency relationships between nodes. The $\alpha_{ij} = 1$ if and only if the two nodes n_i and n_j are neighbours in the spiral reticule placed at a topological distance $k = 1$ and equally labeled or they are consecutive neighbours in the spectral sequence (spiral backbone of nodes); otherwise $\alpha_{ij} = 0$. The elements o_j are the discrete forms of $<I_j> = m/z_j \cdot I_j$. It means that $o_{nj} = 0.5$ when $<I_j>$ is higher than 0.5 and $o_{nj} = 1$ otherwise.

(b) The zeroth-order absolute initial probabilities vector $\mathbf{^{A}\pi_0}$ (see Eq. 2). This vector lists the absolute initial probabilities $^{A}\mathrm{p}_k(j)$ to reach a node n_i from a randomly selected node $\mathrm{n_j}$ in the Spiral network with N nodes (spectral regions):

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

Due to the particularities of the network representation used here, the $^{A}p_k(j)$ only depends on the total number of data points N or spectral regions of the network. Consequently, all the nodes in the network have the same constant value of $^{A}p_{k.}$

As the elements of the matrices $^{k}\mathbf{\Pi}$ (which are k natural powers of the matrix $^{1}\mathbf{\Pi}$) depend on the adjacency relationships between the nodes on the network, the use of Markov chains (MCH) theory thus allows to calculate the $\mathrm{SMC}_k(j)$ for any node n_j that can reach in the spiral network by moving from any node n_i throughout the entire network using walks of length k:

$$\mathrm{SMC}_k = \sum_{j=1}^{n} \mathrm{SMC}_k(j) = \sum_{j=1}^{n} {}^{A}p_k(j) \cdot o_j = {}^{A}\pi_0 \cdot ({}^{\mathbf{1}}\Pi)^k \cdot \boldsymbol{o} \tag{3}$$

Where, $\boldsymbol{o}$ is a vector whose elements o_j are the discrete forms of $<I_j> = m/z_j \cdot I_j$. It means that $o_j = 0.5$ when $<I_j>$ o is higher than 0.5 and $o_j = 1$ otherwise. It is remarkable that $^{A}p_k$ can be written using a MM as the product of $\mathbf{^{A}\pi_0}$ and the natural powers of the matrix $^{1}\Pi$ based on the Chapman–Kolgomorov equations[78] (see right member of Eq. 3 above). Thus, the SMC_k values encode in a stochastic manner the information content related to spectral properties (m/z and I) of all the nodes (spectral regions) placed at different distances in the spiral spectrum network. The evaluation of Eq. 4 for $k = 0$ gives the order zero 2D MS Spiral network node overlapping parameter (SMC_0); for $k = 1$ the first order (SMC_1) and so on. This expansion is illustrated for a linear fragment n_1–n_2–n_3 of a certain spiral network:

$$\begin{aligned}\mathrm{SMC}_0 &= \left[{}^{A}p_0(n1),\quad {}^{A}p_0(n2),\quad {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} o_{n1} \\ o_{n2} \\ o_{n3} \end{bmatrix} \\ &= {}^{A}p_0(n1) \cdot o_{n1} + {}^{A}p_0(n2) \cdot o_{n2} + {}^{A}p_0(n3) \cdot o_{n3} \\ &= \sum_{ni} {}^{A}p_0(ni) \cdot o_{ni} \end{aligned} \tag{4a}$$

$$\begin{aligned}\mathrm{SMC}_1 &= \left[{}^{A}p_0(n1),\quad {}^{A}p_0(n2),\quad {}^{A}p_0(n3)\right] \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} o_{n1} \\ o_{n2} \\ o_{n3} \end{bmatrix} \\ &= {}^{A}p_1(n1) \cdot o_{n1} + {}^{A}p_1(n2) \cdot o_{n2} + {}^{A}p_1(n3) \cdot o_{n3} \\ &= \sum_{ni} {}^{A}p_1(ni) \cdot o_{ni} \end{aligned} \tag{4b}$$

$$
\begin{aligned}
\mathrm{SMC}_2 &= \left[{}^{A}p_0(n1),\quad {}^{A}p_0(n2),\quad {}^{A}p_0(n3)\right] \\
&\quad \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix} \cdot \begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix} \\
&= p_2(n1)\cdot o_{n1} + {}^{A}p_2(n2)\cdot o_{n2} + {}^{A}p_2(n3)\cdot o_{n3} \\
&= \sum_{ni} {}^{A}p_2(ni)\cdot o_{ni}
\end{aligned}
\tag{4c}
$$

$$
\begin{aligned}
\mathrm{SMC}_k &= \left[{}^{A}p_0(n1),\quad {}^{A}p_0(n2),\quad {}^{A}p_0(n3)\right] \\
&\quad \cdot \left(\begin{bmatrix} {}^{1}p_{11} & {}^{1}p_{12} & 0 \\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23} \\ 0 & {}^{1}p_{32} & {}^{1}p_{33} \end{bmatrix}\right)^{k} \cdot \begin{bmatrix} o_1 \\ o_2 \\ o_3 \end{bmatrix} \\
&= p_k(n1)\cdot o_{n1} + {}^{A}p_k(n2)\cdot o_{n2} + {}^{A}p_k(n3)\cdot o_{n3} \\
&= \sum_{ni} {}^{A}p_k(ni)\cdot o_{ni}
\end{aligned}
\tag{4d}
$$

2.3. Statistical analysis

Using the MARCH-INSIDE methodology, we can attempt to develop a simple linear QPPR with the following general formula:

$$
\begin{aligned}
\mathrm{CT} &= b_0 + b_1\cdot \mathrm{SMC}_1 + b_2\cdot \mathrm{SMC}_2 + \ldots + b_k\cdot \mathrm{SMC}_k \\
&= \sum_{k=1}^{v} b_k\cdot \mathrm{SMC}_k + b_0
\end{aligned}
\tag{5}
$$

Here, SMC_k are CIs of whole proteome MS Spiral networks and act as independent or predictive variables. The definition of SMC_k can be found below in the Results section of this work. We selected linear discriminant analysis (LDA)[79–84] to fit the discriminant function. The QPPR model classifies the rat's serum proteome spectrum into two general groups; namely, cardiotoxic-risk (CT = 1 for positive samples) and non-cardiotoxic-risk (NCT = −1 for negative samples). In Eq. 5, b_k represents the coefficients of the classification function, determined by the least-squares method, as implemented in the general discriminant analysis (GDA) module of the STATISTICA 6.0 software package.[85]

We used the Best Subset selection algorithm to search for an adequate combination of predictors, producing the lowest percentage of misclassified instances on training and test sets, respectively.[86,87] The statistical significance of the LDA model was determined by Fisher's test by examining Fisher ratio (F) and the corresponding p-level (p). At the same time, the square Mahalanobis distance (D^2) between the centroids of each one of the two groups (CT and NCT groups) and the Wilk's U statistic were examined to test discriminatory power of the function developed.[80–82,84,88,89] 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 the comparison between their coefficients, is obtained.[90] We also inspected the percentage of good classification, case/variable ratios (ρ parameter), and a number of variables to be explored in order to avoid over-fitting or chance correlation.[86,87] The most frequent cross-validation methods are the following: the independent dataset test, subsampling test, and jackknife test. Chou and Shen have shown that only the jackknife test has the least arbitrariness. Therefore, the jackknife test has been increasingly used by investigators to examine the accuracy of various predictors. The QPPR model was trained by using independent training series selecting at random 75% (47 out of 62) of the samples available. In order to test the predictive ability of the model we used the remaining 15 samples not used for training, which constitute an independent cross-validation data set (CV0). In addition, we used a four-folded re-sampling for CV. It means that we interchanged the samples in training and CV0 randomly, generating up to four different CV series (CV0, CV1, CV2, CV3). The average value of accuracy for these CV series was used as a measure of model stability.[91–99]

3. Results and discussion

3.1. Quantitative Proteome–Property Relationships (QPPRs)

In the present work we propose the use of the network theory combined with high-throughput MS in Toxicoproteomics. In order to illustrate the potentialities of this approach on the early detection of drug-induced cardiac toxicities, we decided to develop a QPPR based on SMC_k values. The best QPPR equation founded is described below:

$$
\begin{aligned}
\mathrm{CT} = {}& -15.09\cdot \mathrm{SMC}_1 + 161.69\cdot \mathrm{SMC}_5 - 425.12\cdot \mathrm{SMC}_7 \\
& + 277.16\cdot \mathrm{SMC}_8 - 0.34 \\
N = 47 \quad F &= 4.37 \quad D^2 = 1.61 \quad U = 0.71 \quad p = 0.005 \quad \rho = 4.7
\end{aligned}
\tag{6}
$$

This QPPR model demonstrated an accuracy of 78.72% classifying MS into two different groups. One group contains MS characteristic of blood proteome samples of an individual that will develop drug-induced cardiotoxicity in the future (CT = 1). The other group is composed by MS control samples with non-cardiotoxic patterns. Specifically, 19 out of 26 CT samples and 18 out of 21 NCT samples were correctly classified. 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 higher than 4 discards any possibility of over-fitting. The square of Mahalanobis's distance (D^2) and Wilk's U statistic provided a measure of the model's discriminatory power expressed through the separation between the centroids of each group and the relation between the intra and inter class variability, respectively. 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 13 out of 15 samples (global predictability = 86.67%). In particular, 7 out of 8 CT samples (sensitivity = 87.50%) and 6 out of 7 NCT samples (specificity = 85.71%) were classified correctly. In addition to the independent cross-validation data set (CV0), the predictive ability of the model was tested by calculating a four-folded re-sampling for CV. The four models are characterized by the following cross-validation accuracy values: 86.67% (CV0), 80.00% (CV1), 76.08% (CV2), and 87.5% (CV3). These values and the correspondent average of 82.56% demonstrate the stability of the model to the change of the training and cross-validation sets.

Following these evaluation criteria also used by other authors in the literature, we can conclude that these results are typical of good discriminant functions obtained with LDA.[82,100–107] In addi-

Table 1
Classification matrices and performance of the LDA-based classification model Eq. 6 on training and validation sets.

Model training			Model validation		
	NCT	CT		NCT	CT
NCT	18	3	NCT	6	1
CT	7	19	CT	1	**7**
Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
78.72%	73.08%	85.71%	86.67%	87.50%	85.71%

tion, the obtained receiver operating characteristic curve (ROC curve), indicates that the model is not random, but a statistically significant classifier (see Fig. 3). A ROC curve plots the sensitivity versus one minus the specificity. An ideal classifier hugs the left side and top side of the network, and the area under the curve is 1.0. A random classifier should achieve approximately 0.5.[108,109]

We decided to test the influence of some different functional forms of the predictors included in the linear model. This test is important to determine if we are using the optimal form of the CIs in the QPPR model or if we can obtain the best model with a simple transformation. In this sense, we developed other three alternative models with three different functional forms (quadratic, inverse and logarithmic). A fourth mixed model including the best subset found of the linear, quadratic, inverse and logarithmic terms, previously analyzed, was also developed. As it can be noted in Table 2, the linear, inverse and mixed models showed the highest accuracy, sensitivity and specificity on the training set. However, the best values of accuracy, sensitivity and specificity on the prediction set are achieved by the linear one, overcoming the predictive performance of the other two models. Although the mixed model resulted to be slightly more statistically significant (see values of F and p in Table 2) and achieved better separation between groups (judged from U and D^2 statistics) compared to the rest of the alternative models, the linear model is still considered the best option due to a lower number of variables used as well as to a better performance on validation sets. Once selected the linear option as the best functional form, it is necessary to find out if the basic assumptions of LDA are fulfilled because in case of severe violations, the reliability of the model's predictions could be compromised.[87,88,110]

As the names implies, LDA establishes a linear, additive relationship between the predictive variables and the response vari-

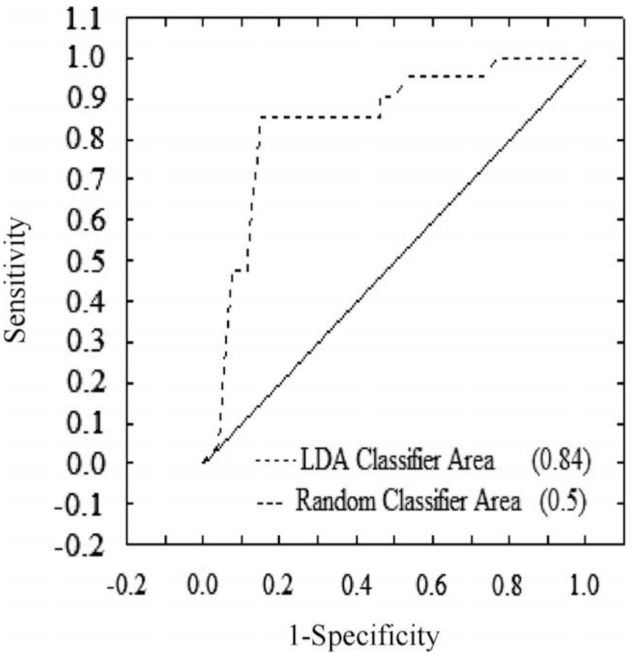

Figure 3. Receiver operating characteristic curve (ROC curve).

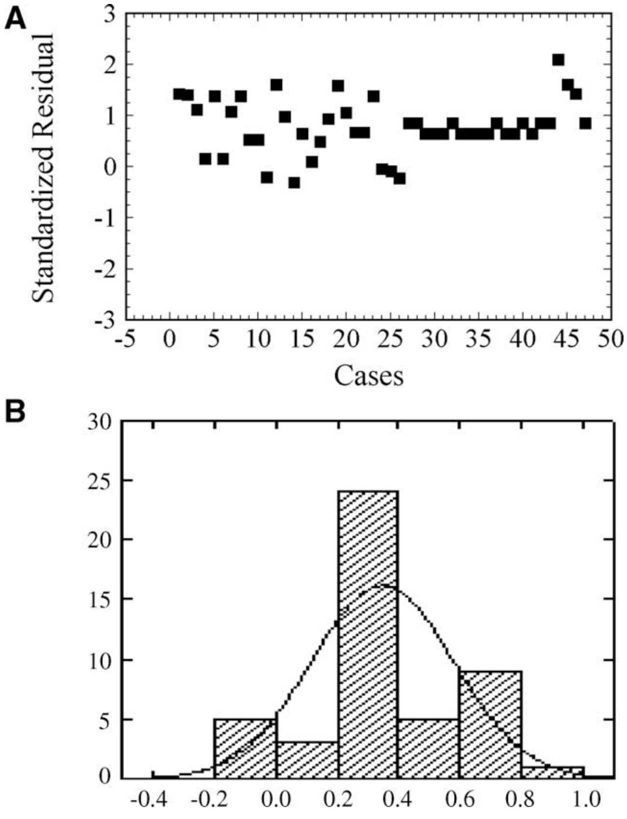

Figure 4. (A) Scatter plot of standardized residual vs cases. (B) Histogram of residuals.

Table 2
Testing the influence of some functional forms of the variables Γ_k (SMC$_k$) on the performance of the model.

Model	Funct. form	Parameters	F	p	U	D^2	ρ
1	Linear	SMC$_1$, SMC$_5$, SMC$_7$, SMC$_8$	4.37	0.005	0.71	1.61	4.70
2	Quadratic	(SMC$_1$)^2, (SMC$_5$)^2, (SMC$_7$)^2, (SMC$_8$)^2	4.13	0.007	0.72	1.53	4.70
3	Inverse	1/SMC$_1$, 1/SMC$_5$, 1/SMC$_7$, 1/SMC$_8$	5.21	0.002	0.67	1.92	4.70
4	Logarithmic	Log(SMC$_1$), Log(SMC$_5$), Log(SMC$_7$), Log(SMC$_8$)	4.78	0.003	0.69	1.76	4.70
5	Mixed	SMC$_5$, (SMC$_7$)^2, (SMC$_8$)^2, 1/SMC$_1$, Log(SMC$_7$)	5.19	0.0009	0.61	2.45	3.92
		Performance					
		Training			Validation		
		Ac	Se	Sp	Ac	Se	Sp
1	Linear	78.72	73.08	85.71	86.67	87.50	85.71
2	Quadratic	74.47	65.38	85.71	80.00	75.00	85.71
3	Inverse	78.72	73.08	85.71	80.00	87.50	71.43
4	Logarithmic	76.60	69.23	85.71	80.00	87.50	71.43
5	Mixed	78.72	73.08	85.71	80.00	87.50	71.43

Note: Ac, accuracy; Se, sensitivity; Sp, specificity.

Table 3
Some important descriptive statistics in residual analysis and normality tests.

Residual's descriptive statistics				
N	Mean	Skewness	Kurtosis	
47	0.35	0.001	0.11	
Residual's normality tests				
Kolmogorov–Smirnov		Lilliefors	Shapiro–Wilk	
D	*p*	*p*	*W*	*p*
0.16	<0.2	<0.01	0.96	0.08

able 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 Figure 4) supports this choice as no systematic pattern is seen.[88] When we checked the parametric assumption of multivariate normality of residuals, it was found that the residuals exhibit adequate values of skewness and kurtosis,[88] which is a sign of normal distribution fitting. However, the hypothesis of multivariate normality of residuals cannot be confirmed since the results of applying the Kolmogorov–Smirnov and Shapiro–Wilks test do not reject this hypothesis, but Lilliefors hypothesis test provides opposite results evidencing slight deviations from the normal distribution (statistic values in Table 3). This can be noted through the frequency histogram shown in section B of Figure 4. In addition, as the term related to the error (represented by residuals) is not included in the LDA equation; the mean must be ideally 0. Actually, the residual mean value for our model is close to the assumed value of 0 (see Table 3).[111]

Moving on to the next important parametric assumption of LDA, that is, homocedasticity (i.e., homogeneity of variance of the variables) was also checked by simply plotting the square standardized residuals for each predictor variable.[88] The plots in Figure 5 reveal an adequate scatter on the points, without any consistent pattern. Hence, there is no reason to reject the pre-adopted assumption of homocedasticity. The main violation of our model relies on the high collinearity exhibited by the variables included in the model (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

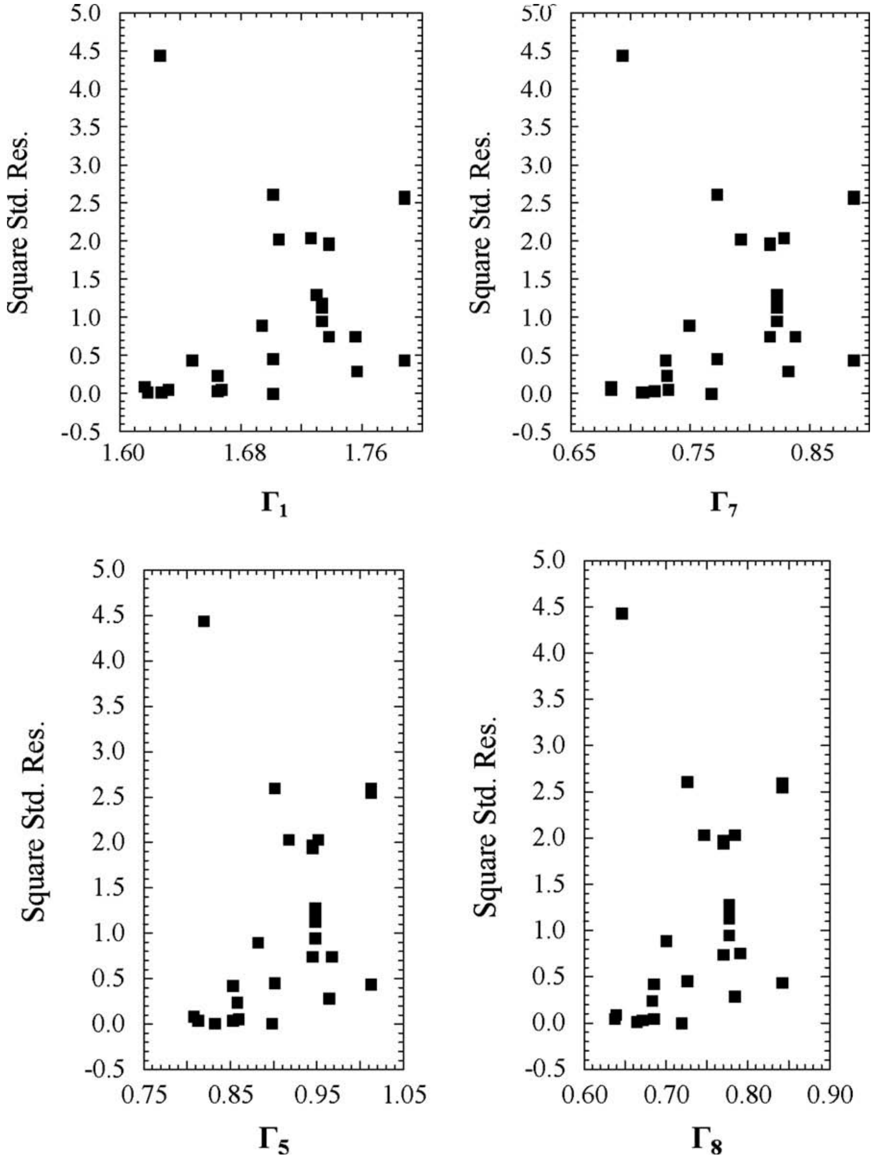

Figure 5. Plots of square standardized residual vs the values of the SMC_k included in the QPPR model.

constant is not fully applicable when multiple co-linearity exists. However, the fact that some or all predictor variables are correlated among themselves does not, in general, inhibit the model's ability to obtain a good fit, nor does it tend to affect inferences about the mean responses or predictions of new observations.[112] In any case, using Randic's orthogonalization procedure, we were able to obtain a model with orthogonal variables (Eq. 7) and the same statistical parameters of Eq. 6:

$$\mathrm{CT} = 1.32 \cdot {}^{1}\mathrm{O}_1 + 1.03 \cdot {}^{2}\mathrm{O}_5 + 36.46 \cdot {}^{3}\mathrm{O}_7 + 110.41 \cdot {}^{4}\mathrm{O}_8 - 0.14 \qquad (7)$$

In the orthogonalized equation, we re-write all terms using the following notation: ${}^{i}O_k$, where O is an orthogonal variable, k is the order of the original SMC$_k$ variable and i is the orthogonalization order.[113–116]

Finally, due to natural limitations inherent to QSPR-like models caused by data conformation it is interesting to study the Domain of Applicability (DA) of the model. The DA may be reduced because of the low number of instances (samples) used for training. A simple method to determine the DA of QSPR-like models is the visual inspection of a leverage plot (i.e., plot of residuals vs the leverages of the training instances).[117,118] The leverage (h) of a sample in the original variable space measures its influence on the model and it is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad (i = 1, \ldots, n) \qquad (8)$$

Where, x_i is the CIs or descriptor vector of the considered instance (Γ_k values in this work). $\boldsymbol{X}$ is the model matrix derived from the training set descriptor values. The warning leverage h^* is defined as follows:

$$h^* = 3 \times p'/n \qquad (9)$$

Where, n is the number of training instances and p' is the number of the model adjustable parameters. Figure 6 shows the results for the DA analysis of the final QPPR model, which was determined by training instances with h values lower than h^* = 0.32. New samples with an h value higher than h^* and/or a value of standardized residual higher than 2 or lower than −2 are out of the DA bandwidth of the model and consequently cannot be reliably predicted.[119–121]

4. Conclusions

In the present, work we introduced for the first time a Spiral network representation of the serum proteome MS. Additionally, we defined CIs of the whole blood proteome MS, called the Spiral Markov Connectivity (SMC$_k$) parameters. Finally, we investigated the applications of the SMC$_k$ indices in Toxicoproteomics. In so doing, we seek a QPPR model for the early detection of drug-induced cardiac toxicities based on the SMC$_k$ values for the MS of a

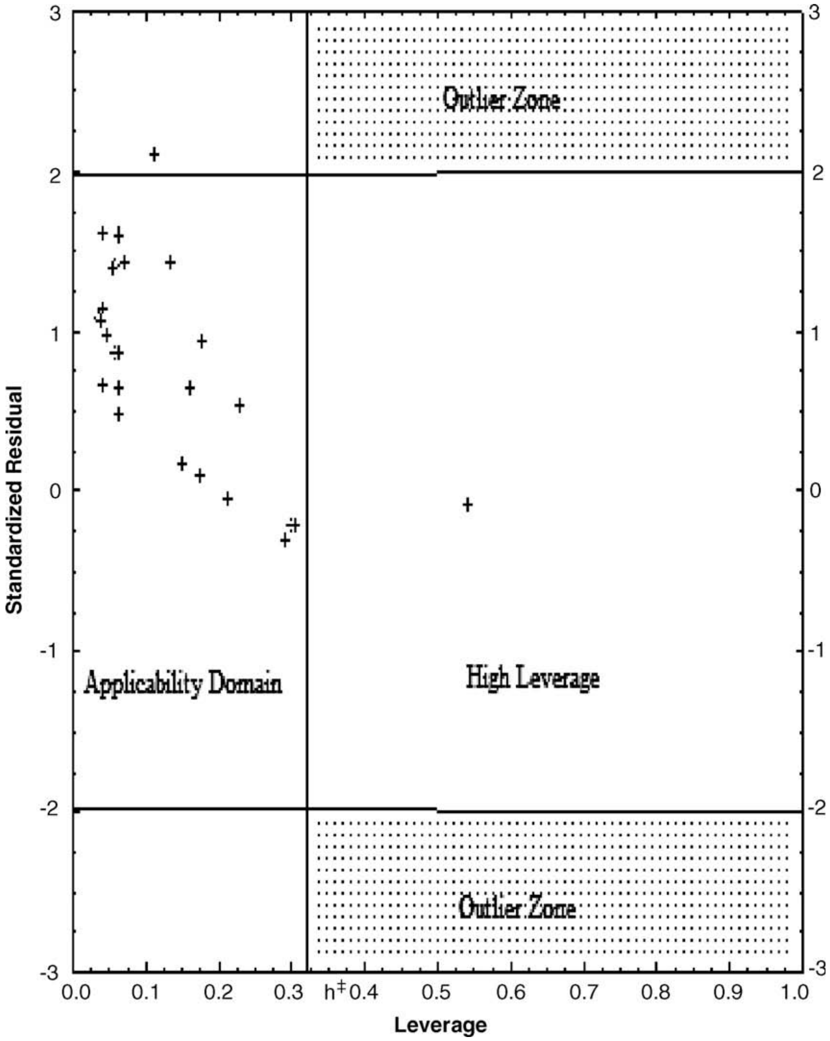

Figure 6. Analysis of the Domain Applicability of the QPPR model.

blood sample. We can conclude that the connectivity indices (SMC_k) derived from presented Spiral network representation of the blood proteome MS allow to capture important features of protein Biomarkers for the early detection of drug-induced cardiac toxicities. This kind of study could become an interesting alternative and/or complementary technique to the direct search of Biomarker patterns in clinical proteomics using LDA and other data analysis techniques (see Fig. 6).[122]

Acknowledgments

We are grateful to Editor Prof. Herbert Waldmann and the two unknown referees for their kind attention and useful comments. The authors thank the Portuguese Fundação para a Ciência e a Tecnologia (FCT) (SFRH/BD/30698/2006 and SFRH/BPD/24997/2005). The corresponding author González-Díaz H. acknowledges contract/grant sponsorship for the research position funded by the Program Isidro Parga Pondal of the "Xunta de Galicia".

Supplementary data

The computed values of the five predictor variables included in the QPPR model for the 62 samples used, as well as their respective observed and predicted classifications are included in the Supplementary material related to this work. This supplementary information also contains the distribution of 62 samples for training or model validation and the posterior probabilities to be classified as CT or NCT (Table SM1). In addition, we presented the matrix correlations of the original and orthogonalized variables in Table SM2. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2008.10.004.

References and notes

1. Chartrand, G. *Introductory Graph Theory*; Dover: New York, 1985.
2. Vilar, S.; Santana, L.; Uriarte, E. *J. Med. Chem.* **2006**, *49*, 1118.
3. Bonchev, D. *J. Mol. Graphics Modell.* **2001**, *20*, 65.
4. Vilar, S.; Quezada, E.; Santana, L.; Uriarte, E.; Yanez, M.; Fraiz, N.; Alcaide, C.; Cano, E.; Orallo, F. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 257.
5. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2002.
6. Ivanciuc, O. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412.
7. Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573.
8. Randić, M. *Chem. Phys. Lett.* **2004**, 468.
9. Randic, M.; Zupan, J. *SAR QSAR Environ. Res.* **2004**, *15*, 191.
10. Randic, M.; Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 532.
11. Randič, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.
12. Chou, K. C. *Proteins* **2001,** *(Erratum: ibid., 2001, Vol. 44, 60) 43*, 246.
13. Chou, K. C.; Cai, Y. D. *Proteins* **2003**, *53*, 282.
14. Chou, K. C.; Cai, Y. D. *J. Cell Biochem.* **2004**, *91*, 1197.
15. Cai, Y. D.; Chou, K. C. *J. Proteome Res.* **2005**, *4*, 967.
16. Gao, Y.; Shao, S.; Xiao, X.; Ding, Y.; Huang, Y.; Huang, Z.; Chou, K. C. *Amino Acids* **2005**, *28*, 373.
17. Liu, H.; Yang, J.; Wang, M.; Xue, L.; Chou, K. C. *Protein J.* **2005**, *24*, 385.
18. Shen, H.; Chou, K. C. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 288.
19. Cai, Y. D.; Chou, K. C. *J. Theor. Biol.* **2006**, *238*, 395.
20. Wang, S. Q.; Yang, J.; Chou, K. C. *J. Theor. Biol.* **2006**, *242*, 941.
21. Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Chou, K. C. *Amino Acids* **2006**, *30*, 49.
22. Liao, B.; Ding, K. *J. Comput. Chem.* **2005**, *26*, 1519.
23. Liao, B.; Ding, K.; Wang, T. *J. Biomol. Struct. Dynamics* **2005**, *22*, 455.
24. Liao, B.; Luo, J.; Li, R.; Zhu, W. *Int. J. Quantum Chem.* **2006**, *106*, 1749.
25. Yu-Hua, Y.; Liao, B.; Tian-Ming, W. *J. Mol. Struct. THEOCHEM* **2005**, *755*, 131.
26. Zhu, W.; Liao, B.; Ding, K. *J. Mol. Struct. THEOCHEM* **2005**, *757*, 193.
27. Caballero, J.; Fernandez, L.; Abreu, J. I.; Fernandez, M. *J. Chem. Inf. Modell.* **2006**, *46*, 1255.
28. Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. *Proteins* **2007**, *67*, 834.
29. Cui, G.; Chen, Y.; Huang, D. S.; Han, K. *J. Biomed. Biotechnol.* **2008**, *2008*, 860270.
30. Zhang, G. Z.; Han, K. *Comput. Biol. Chem.* **2007**, *31*, 233.
31. Han, K.; Nepal, C. *FEBS Lett.* **2007**, *581*, 1881.
32. Byun, Y.; Han, K. *Nucleic Acids Res.* **2006**, *34*, W416.
33. Randic, M.; Witzmann, F. A.; Kodali, V.; Basak, S. C. *J. Chem. Inf. Modell.* **2006**, *46*, 116.
34. Randic, M. *J. Proteome Res.* **2006**, *5*, 1575.
35. Randic, M.; Novic, M.; Vracko, M. *J. Chem. Inf. Modell.* **2005**, *45*, 1205.
36. Randic, M.; Lers, N.; Vukicevic, D.; Plavsic, D.; Gute, B. D.; Basak, S. C. *J. Proteome Res.* **2005**, *4*, 1347.
37. Randic, M.; Estrada, E. *J. Proteome Res.* **2005**, *4*, 2133.
38. Bajzer, Z.; Randic, M.; Plavsic, D.; Basak, S. C. *J. Mol. Graphics Modell.* **2003**, *22*, 1.
39. Randič, M. *Int. J. Quantum Chem.* **2002**, *90*, 848.
40. Randic, M.; Zupan, J.; Novic, M.; Gute, B. D.; Basak, S. C. *SAR QSAR Environ. Res.* **2002**, *13*, 689.
41. Randic, M.; Novic, M.; Vracko, M. *J. Proteome Res.* **2002**, *1*, 217.
42. Randic, M.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 983.
43. Randic, M.; Zupan, J.; Novic, M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1339.
44. Bonchev, D.; Buck, G. A. *J. Chem. Inf. Model.* **2007**, *47*, 909.
45. Anderson, N. L.; Polanski, M.; Pieper, R.; Gatlin, T.; Tirumalai, R. S.; Conrads, T. P.; Veenstra, T. D.; Adkins, J. N.; Pounds, J. G.; Fagan, R.; Lobley, A. *Mol. Cell Proteomics* **2004**, *3*, 311.
46. Shakhnovich, B. E.; Harvey, J. M.; Comeau, S.; Lorenz, D.; DeLisi, C.; Shakhnovich, E. *BMC Bioinformatics* **2003**, *4*, 34.
47. Bensmail, H.; Golek, J.; Moody, M. M.; Semmes, J. O.; Haoudi, A. *Bioinformatics* **2005**, *21*, 2210.
48. Zhou, G.; Li, H.; Gong, Y.; Zhao, Y.; Cheng, J.; Lee, P.; Zhao, Y. *Proteomics* **2005**, *5*, 3814.
49. Anderson, K. S.; LaBaer, J. *J. Proteome Res.* **2005**, *4*, 1123.
50. Ruddat, V. C.; Whitman, S.; Klein, R. D.; Fischer, S. M.; Holman, T. R. *Prostate* **2005**, *64*, 128.
51. Yanagisawa, K.; Xu, B. J.; Carbone, D. P.; Caprioli, R. M. *Clin. Lung Cancer* **2003**, *5*, 113.
52. Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; Andrews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. *Proteomics* **2005**, *5*, 3226.
53. Ornstein, D. K.; Tyson, D. R. *Urol. Oncol.* **2006**, *24*, 231.
54. González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. *Proteomics* **2008**, *8*, 750.
55. González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E. *Curr. Top. Med. Chem.* **2007**, *7*, 1025.
56. Bartels, C. *Biomed. Environ. Mass Spectrom.* **1990**, *19*, 363.
57. Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V. *Comput. Appl. Biosci.* **1995**, *11*, 427.
58. Taylor, J. A.; Johnson, R. S. *Rapid Commun. Mass Spectrom.* **1997**, *11*, 1067.
59. Dancík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6*, 327.
60. Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77*, 964.
61. Liotta, L. A.; Ferrari, M.; Petricoin, E. *Nature (London)* **2003**, *425*, 905.
62. Mehta, A. I.; Ross, S.; Lowenthal, M. S.; Fusaro, V.; Fishman, D. A.; Petricoin, E. F., 3rd; Liotta, L. A. *Dis. Markers* **2003**, *19*, 1.
63. Hu, S.; Loo, J. A.; Wong, D. T. *Proteomics* **2006**, *6*, 6326.
64. Kantor, A. B. *Dis. Markers* **2002**, *18*, 91.
65. McDonald, W. H.; Yates, J. R., 3rd *Dis. Markers* **2002**, *18*, 99.
66. Petricoin, E. F.; Rajapaske, V.; Herman, E. H.; Arekani, A. M.; Ross, S.; Johann, D.; Knapton, A.; Zhang, J.; Hitt, B. A.; Conrads, T. P.; Veenstra, T. D.; Liotta, L. A.; Sistare, F. D. *Toxicol. Pathol.* **2004**, *32*(Suppl. 1), 122.
67. Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, *359*, 572.
68. Petricoin, E. F., 3rd; Ornstein, D. K.; Paweletz, C. P.; Ardekani, A.; Hackett, P. S.; Hitt, B. A.; Velassco, A.; Trucco, C.; Wiegand, L.; Wood, K.; Simone, C. B.; Levine, P. J.; Linehan, W. M.; Emmert-Buck, M. R.; Steinberg, S. M.; Kohn, E. C.; Liotta, L. A. *J. Natl. Cancer Inst.* **2002**, *94*, 1576.
69. Randic, M.; Lers, N.; Plavsic, D.; Basak, S. C.; Balaban, A. T. *Chem. Phys. Lett.* **2005**, *407*, 205.
70. Lambertenghi-Deliliers, G.; Zanon, P. L.; Pozzoli, E. F.; Bellini, O. *Tumori* **1976**, *62*, 517.
71. Zhang, J.; Herman, E. H.; Ferrans, V. J. *Am. J. Pathol.* **1993**, *142*, 1916.
72. Herman, E. H.; Zhang, J.; Rifai, N.; Lipshultz, S. E.; Hasinoff, B. B.; Chadwick, D. P.; Knapton, A.; Chai, J.; Ferrans, V. J. *Cancer Chemother. Pharmacol.* **2001**, *48*, 297.
73. Zhang, J.; Herman, E. H.; Knapton, A.; Chadwick, D. P.; Whitehurst, V. E.; Koerner, J. E.; Papoian, T.; Ferrans, V. J.; Sistare, F. D. *Toxicol. Pathol.* **2002**, *30*, 28.
74. González-Díaz, H.; Molina-Ruiz, R.; Hernandez, I. **2005,** MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design), gonzalezdiazh@yahoo.es.
75. Gonzalez-Diaz, H.; Saiz-Urra, L.; Molina, R.; Gonzalez-Diaz, Y.; Sanchez-Gonzalez, A. *J. Comput. Chem.* **2007**, *28*, 1042.
76. Gonzalez-Diaz, H.; Perez-Castillo, Y.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1990.
77. Ramos de Armas, R.; Gonzalez Diaz, H.; Molina, R.; Uriarte, E. *Proteins* **2004**, *56*, 715.
78. Gnedenko, B. *The Theory of Probability*; Mir Publishers: Moscow, 1978.
79. van de Waterbeemd, H.. In *Chemometric Methods in Molecular Design*; VCH: Weinheim, 1995; Vol. 2.
80. Murcia-Soler, M.; Perez-Gimenez, F.; Nalda-Molina, R.; Salabert-Salvador, M. T.; Garcia-March, F. J.; Cercos-del-Pozo, R. A.; Garrigues, T. M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1345.

81. Linding, R.; Jensen, L. J.; Diella, F.; Bork, P.; Gibson, T. J.; Russell, R. B. *Structure* **2003**, *11*, 1453.
82. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. *J. Antimicrob. Chemother.* **2004**, *53*, 65.
83. de Armas, R. R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Biopolymers* **2005**, *77*, 247.
84. Cercos-del-Pozo, R. A.; Perez-Gimenez, F.; Salabert-Salvador, M. T.; Garcia-March, F. J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 178.
85. StatSoft.Inc., STATISTICA (data analysis software system), ver. 6.0, www.statsoft.com, **2002**.
86. Kowalski, R. B.; Wold, S. In *Handbook of Statistics*; Krishnaiah, P. R., Kanal, L. N., Eds.; North Holland Publishing Company: Amsterdam, 1982; p 673.
87. Van de Waterbeemd, H. In *Chemometric Methods in Molecular Design*; Waterbeemd, Van de, Ed.; VCH: New York, 1995.
88. Stewart, J.; Gill, L. *Econometrics*; Prentice Hall: London, 1998.
89. Marrero-Ponce, Y.; Diaz, H. G.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
90. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. *Applied Linear Statistical Models*; McGraw Hill: New York, 2005.
91. Chou, K. C.; Zhang, C. T. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275.
92. Chou, K. C.; Shen, H. B. *Anal. Biochem.* **2007**, *370*, 1.
93. Chou, K. C.; Shen, H. B. *Nat. Protocols* **2008**, *3*, 153.
94. Chen, Y. L.; Li, Q. Z. *J. Theor. Biol.* **2007**, *248*, 377.
95. Chen, Y. L.; Li, Q. Z. *J. Theor. Biol.* **2007**, *245*, 775.
96. Diao, Y.; Li, M.; Feng, Z.; Yin, J.; Pan, Y. *J. Theor. Biol.* **2007**, *247*, 608.
97. Lin, H. *J. Theor. Biol.* **2008**, *252*, 350.
98. Niu, B.; Cai, Y. D.; Lu, W. C.; Li, G. Z.; Chou, K. C. *Protein Pept. Lett.* **2006**, *13*, 489.
99. Xiao, X.; Chou, K. C. *Protein Pept. Lett.* **2007**, *14*, 871.
100. Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E. *J. Med. Chem.* **2006**, *49*, 1149.
101. Ponce, Y. M.; Diaz, H. G.; Zaldivar, V. R.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
102. Patankar, S. J.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 885.
103. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Medina-Casamayor, P. *J. Mol. Graphics Modell.* **2003**, *21*, 375.
104. Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Pereira, D. M.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Aran, V. J.; Martinez-Fernandez, A. R.; Garcia Sanchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838.
105. McElroy, N. R.; Jurs, P. C.; Morisseau, C.; Hammock, B. D. *J. Med. Chem.* **2003**, *46*, 1066.
106. Mattioni, B. E.; Jurs, P. C. *J. Mol. Graphics Modell.* **2003**, *21*, 391.
107. Gozalbes, R.; Galvez, J.; Garcia-Domenech, R.; Derouin, F. *SAR QSAR Environ. Res.* **1999**, *10*, 47.
108. Zweig, M. H. *Arch. Pathol. Lab. Med.* **1994**, *118*, 141.
109. Zweig, M. H.; Broste, S. K.; Reinhart, R. A. *Clin. Chem.* **1992**, *38*, 1425.
110. Cruz-Monteagudo, M.; González-Díaz, H.; Agüero-Chapin, G.; Santana, L.; Borges, F.; Domínguez, R. E.; Podda, G.; Uriarte, E. *J. Comput. Chem.* **2007**, *28*, 1909.
111. González-Díaz, H.; Pérez-Bello, A.; Cruz-Monteagudo, M.; González-Díaz, Y.; Santana, L.; Uriarte, E. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 20.
[112]. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. *Applied Linear Statistical Models*; McGraw Hill: New York, 2005.
113. González-Díaz, H.; Marrero, Y.; Hernandez, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castanedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Perez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318.
114. Randić, M. *New J. Chem.* **1991**, *15*, 517.
115. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
116. Randić, M. *J. Comput. Chem.* **1993**, *14*, 363.
117. Atkinson, A. C.. In *Plots, Transformations and Regression*; Clarendon Press: Oxford, 1985.
118. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
119. Monari, G.; Dreyfus, G. *Neural Comput.* **2002**, *14*, 1481.
120. Meloun, M.; Syrovy, T.; Bordovska, S.; Vrana, A. *Anal. Bioanal. Chem.* **2007**, *387*, 941.
121. Meloun, M.; Hill, M.; Militky, J.; Vrbikova, J.; Stanicka, S.; Skrha, J. *Clin. Chem. Lab. Med.* **2004**, *42*, 311.
122. Lilien, R. H.; Farid, H.; Donald, B. R. *J. Comput. Biol.* **2003**, *10*, 925.