

[g013]

Predicting Proteome-Early Drug Induced Cardiac Toxicity Relationships (Pro-EDICToRs) with Node Overlapping Parameters (NOPs) of a new class of Blood Mass-Spectra graphs

Humberto González-Díaz^{a,*}, Maykel Cruz-Monteagudo^{b,c}, Fernanda Borges^c, Eugenio Uriarte^a

^a Unit of Bioinformatics & Connectivity Analysis, Institute of Industrial Pharmacy, and Department of Organic Chemistry, Faculty of Pharmacy, USC, 15782, Santiago de Compostela, Spain.

^b Applied Chemistry Research Center, Faculty of Chemistry and Pharmacy, UCLV, SC, 54830, Cuba.

^c Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, University of Porto 4050-047, Porto, Portugal.

ABSTRACT – Blood Serum Proteome-Mass Spectra (SP-MS) may allow detecting Proteome-Early Drug Induced Cardiac Toxicity Relationships (called here Pro-EDICToRs). However, due to the thousands of proteins in the SP identifying general Pro-EDICToRs patterns instead of a single protein marker may represent a more realistic alternative. In this sense, first we introduced a novel Cartesian 2D spectrum graph for SP-MS. Next, we introduced the graph node-overlapping parameters (nop_k) to numerically characterize SP-MS using them as inputs to seek a Quantitative Proteome-Toxicity Relationship (QPTR) classifier for Pro-EDICToRs with accuracy higher than 80%. Principal Component Analysis (PCA) on the nop_k values present in the QPTR model explains with one factor (F_1) the 82.7% of variance. Next, these nop_k values were used to construct by the first time a Pro-EDICToRs Complex Network having nodes (samples) linked by edges (similarity between two samples). We compared the topology of two sub-networks (cardiac toxicity and control samples); finding extreme relative differences for the re-linking (P) and Zagreb (M2) indices (9.5 and 54.2 % respectively) out of 11 parameters. We also compared sub-networks with well known ideal random networks including Barabasi-Albert, Kleinberg Small World, Erdos-Renyi, and Epsstein Power Law models. Finally, we proposed Partial Order (PO) schemes of the 115 samples based on LDA-probabilities, F_1 -scores and/or network node degrees. PCA-CN and LDA-PCA based POs with Tanimoto's coefficients equal or higher than 0.75 are promising for the study of Pro-EDICToRs. These results show that simple QPTRs models based on MS graph numerical parameters are an interesting tool for proteome research.

KEYWORDS: Toxicoproteomics, Drug-induced cardiac toxicities, Mass spectrometry, Mass Spectrum graph, Markov model, Quantitative Proteome-Toxicity Relationships, Complex Networks, Principal Components Analysis, and Partial Order.

CORRESPONDING AUTHOR FOOTNOTE: To whom correspondence should be addressed: Humberto González-Díaz. Faculty of Pharmacy, University of Santiago de Compostela 15782, Spain. Tel: +34-981-563100. Fax: +34-981 594912. Email: gonzalezdiazh@yahoo.es or qohumbe@usc.es

1. Introduction

The main basis for early detection of disease and drug induced toxicity nowadays remains finding translational and safety biomarkers that can predict or anticipate toxic manifestation and detect damage earlier in human trials. Specifically, cardiotoxicity is a serious adverse effect of chemotherapy that encompasses a spectrum of disorders, ranging from relatively benign arrhythmias to potentially lethal conditions such as myocardial ischemia/infarction and cardiomyopathy. 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^{1,2}. The extent and severity of the necrosis can be monitored by the levels of bioactive markers³. However, the number of new biomarkers reaching routine clinical use remains unacceptably low^{4,5}. At the same time, 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^{6,7}. Consequently, a blood proteome represents a potential target for the early detection of diseases and drug induced toxicities.

The blood proteome is changing constantly as a consequence of the perfusion of the organ undergoing drug-induced damage and 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 occurs⁸. 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 body⁸ and the optimal performance in the low mass range exhibited by mass spectroscopy^{9,10}, the use of this method applied to proteomics may offer the best chance of discovering these early stage changes.

However, due to the thousands of intact and cleaved proteins in the SP, 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 has occurred. For these reasons, to identify patterns by using the SP-Mass Spectra (SP-MS) 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 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 toxicities¹¹, ovarian¹² and prostate cancer¹³. Consequently, we can state that SP-MS may allow detecting Proteome-Early Drug Induced Cardiac Toxicity Relationships (called here Pro-EDICToRs) at the first stages.

In the present work we decided to identify SP Pro-EDICToRs parameters and use it in generating a prediction model by using a graph theoretical approach instead of directly identify patterns within the high-throughput MS. The application of graph theory to MS was first proposed by Bartels for peptide sequencing¹⁴. The basic idea consists in transforming a mass spectrum into a graph called the spectrum graph. Basically, each peak in the experimental spectrum is represented as a node (or several nodes) in the spectrum graph and a directed edge (or arc) is established between two vertices if the mass difference of the two vertices equals the mass of one or several aminoacids. Several algorithms that make use of spectrum graphs have been designed for *de novo* peptide sequencing. Among the most popular are “SeqMS”¹⁵, “Lutefisk”¹⁶, “Sherenga”¹⁷ and more recently “PepNovo”¹⁸.

The construction of the spectrum graph of all these algorithms share the basic idea proposed by Bartels with their respective particularities. The SeqMS algorithm first assumes a list of possible ion types with corresponding probabilities. The list is then used to transform the experimental MS into a spectrum graph. Each peak will correspond to a set of nodes in the spectrum graph, according to the list of ion types. A graph is then obtained by linking all pairs of vertices that differ by the mass of an amino acid or the combination of several amino acids¹⁵. In Lutefisk algorithm the experimental spectrum data is first reduced to a list of significant fragment ions. The N- and C-terminal evidence list, which reveals the possible N- and C-terminal ions respectively, is then determined

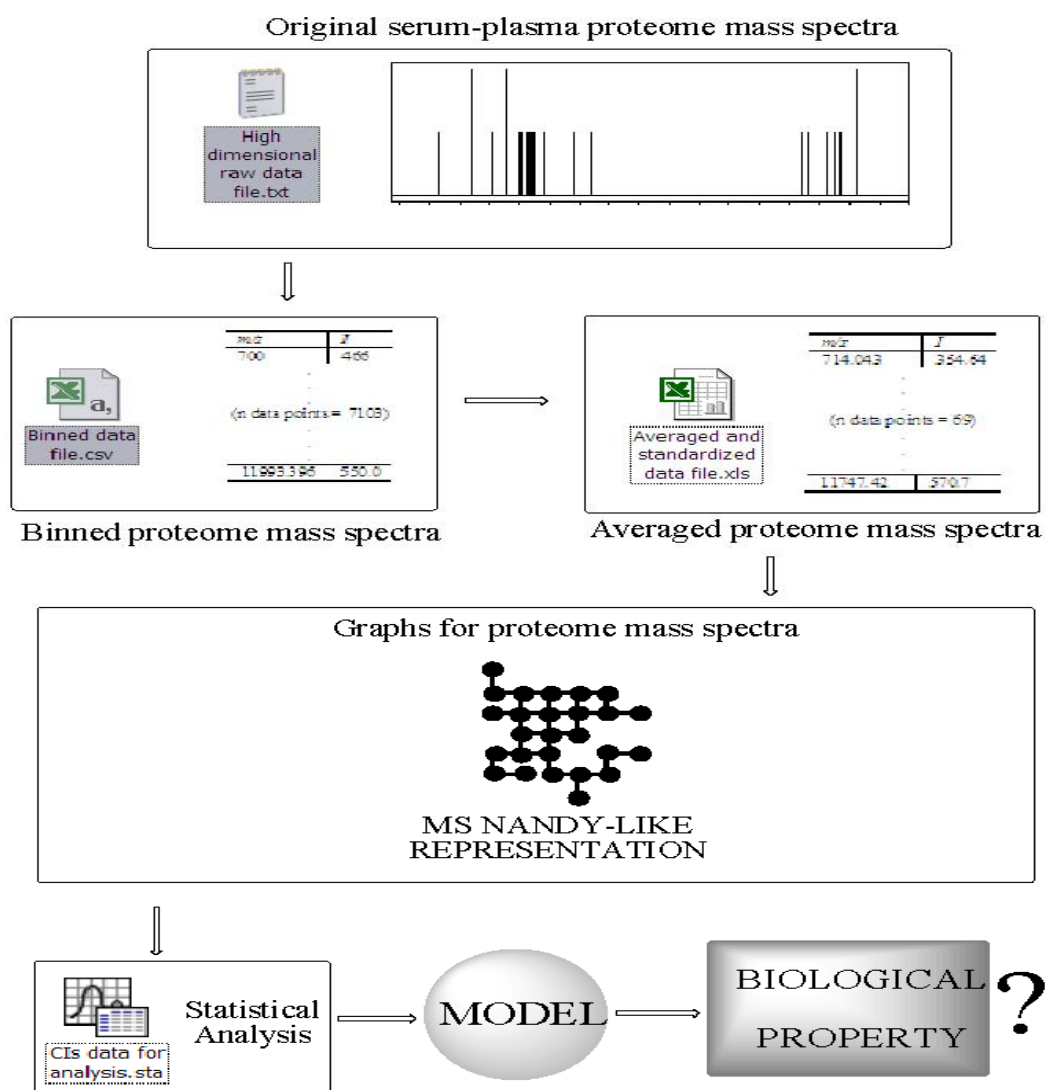
and a “sequence spectrum” (a term proposed by Bartels) is derived where the x ordinate consists on mass/charge (m/z) values for the b -ions and the y ordinate consists on the probability of cleavage of each site¹⁶. The Sherenga algorithm is also based on ion types. Because of it, a method is designed to automatically learn ion types from a training set of experimental spectra of known sequences, without knowing a priori the fragmentation patterns. After the ion types are learned, the experimental MS is transformed into a spectrum graph¹⁷. The nodes in the spectrum graph used in the PepNovo algorithm are assigned by creating for each mass in the experimental spectrum a set of nodes at different masses. Nodes having similar masses are merged (since it is likely that they are created by different ion fragments from the same cleavage site). Here the nodes are scored according to a probability-based score that gives premiums for present fragment ions, and penalties for missing ones¹⁸.

On the other hand, many graph based representations have been introduced for biological data different from MS as for example: single DNA, RNA and protein sequences or even for 2D proteomics maps. In several cases we can calculate from these graphs numerical indices, sometimes called Topological Indices (TIs) or Connectivity Indices (CIs), encoding important biological information.^{19, 20} In this sense, we can call the attention of readers on the works after Randić, Liao, Nandy, Basak, and many others, which developed some of these graph theory based representations with applications in proteome research in general including Toxicoproteomics.²¹⁻⁴¹ A novel 2D representation for proteins sequence similar to the proposed by Nandy for DNA sequences⁴²⁻⁴⁴ was introduced by our group to study protein sequences. This 2D graph embedded in a Cartesian space assigns each one of the four aminoacid groups to each axis direction according to the physicochemical nature of the aminoacids (polar, non-polar, acid, or basic)⁴⁵. The numerical parameters that characterize the previous representations can be used to seek a Quantitative Structure-Activity Relationship (QSAR) models to predict systems function. System herein in the more broad sense in proteomics and bioinformatics including drugs activity, protein function, proteome-disease relationships, proteins structure NMR, gene microarray data, proteomic electrophoresis 2D maps or proteome MS. In general, these sequence, graph, and/or higher dimension numerical indices can be combined with data analysis methods such as Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Genetic Algorithms (GA), k-Nearest Neighbors (KNN) and/or other machine learning classifiers in bioinformatics, medicinal chemistry, and proteomics.^{20, 46-60}

Among the multiple applications of the above mentioned numerical indices we can list: data dimensions reduction, clustering, and/or ordering, in addition. With these goals in mind one can perform different studies apart from and/or complementary to LDA, SVM, ANN, GA, or KNN including: Principal Component Analysis (PCA) and/or Complex Network construction. PCA, have been largely used alone or combined with these methods to reduce data, construct parameters spaces, and rank cases or samples in proteins drug inhibitors search, protein characterization, and proteome research with important applications in Toxicogenomics and Toxicoproteomics too.⁶¹⁻⁷⁴ In addition, Complex Network construction based on the above mentioned parameters is also very useful in genomics, proteomics, medical-social research or science in general. These networks are large objects composed at least by nodes and edges or arcs. Drugs, genes, RNAs, proteins, organisms, brain cortex regions, diseases, patients or environmental systems to name only a few examples that may play the role of nodes. Otherwise, edges represent some kind of similarity/dissimilarity relationships between nodes often.^{20, 75-99} As mentioned above, after using system structural numeric parameters as inputs for LDA, SVM, ANN, KNN, PCA analysis or Complex network prediction we can employ the outputs of these methods or direct experimental data to construct sample orders. Orders based on only one ranking attribute are simple but may fail in describing all remarkable sample features for complex cases or samples such as DNA sequence, proteins structure, proteomics maps or gene microarray data. In this cases, may be more useful the construction of a Partial Order (PO) based on more than one ranking parameters (x_i). These POs can be represented with the so called Hasse diagrams, which are also graphs or network like representations. The nodes of Hasse diagrams are the samples or cases (as above expressed: chemicals, proteins, proteomics maps, organisms, environmental systems) and edges herein express ordering instead of similarity/dissimilarity relationships.¹⁰⁰⁻¹¹⁷

Despite the proved efficacy of graph/network representations based numerical parameters few works have been reported integrating Toxicoproteomics MS data with QPTR models, PCA, Complex networks and PO analysis based on graphs other than the classic fragmentation MS graph. This field of study appears to be promising in the study of Pro-EDICToRs and proteome research in general. In the present work, we propose an

alternative Cartesian 2D graph representation to the MS graph based on an adaptation of representations similar those proposed by Nandy, Randic, Liao and others, for protein and DNA sequences respectively. Next, we derived from the new graphical representation numerical indices based on Graph and Markov chain theory. These numerical indices, called node overlapping parameters (nop_k), are then used to find a QPTR model for early detection of drug-induced cardiac toxicities. A visual representation of the approach proposed in this work is shown in Scheme 1. In addition, we perform alternative Principal Components Analysis (PCA) and construct by the first time a Complex Network of Pro-EDICToRs. The outputs of the three analysis LDA, PCA, and Complex Networks were used to propose alternative Partial Orders (PO) of the samples.



Scheme 1. Schematic representation of the SP-MS Cartesian graph-based early detection of drug-induced cardiac toxicities

2. Methods

SP-MS data set. For the generation of the SP-MS based QTPR model we used tab-delimited data files containing mass/charge (m/z) and peak intensity (I) values taken from serum proteome high-resolution spectra reported by Petricoin *et al.*¹¹. According to Petricoin *et al.*, the high resolution mass spectrometer used in generating the respective mass spectra is a hybrid quadrupole time-of-flight mass spectrometer (QSTAR pulsar I, Applied Biosystems Inc., Framingham, Massachusetts). 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 points to 7105 points per sample. The high-resolution spectra is 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)¹¹. 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, respectively) that have been used recently to measure effects of therapeutic compounds on cardiac damage¹¹⁸⁻¹²¹. Since the cardiac toxicity profile of 88 out of 203 samples analyzed was reported as unknown, only 115 samples were used in this work:

1. Definitive Positive (34 samples with overt cardiotoxicity): Tab-delimited data files exported from SP 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). Also included as positive were rats with lower cTnT levels (\geq 0.08 ng/mL) but also with mild apparent pathologic changes determined by the histological lesion scores.
2. Probable Positive (10 samples with probable cardiotoxicity): Tab-delimited data files exported from SP high-resolution spectra belonging to sera from SHR model with low serum cTnT (between 0.08 and 0.15 ng/mL).
3. Definitive Negative (28 samples without cardiotoxicity): Tab-delimited data files exported from SP 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.
4. Probable Negative (43 samples without clear evidences of carditoxicity): Tab-delimited data files exported from rats serum that were expected to be classified as negatives (histological score = 0 or not taken) but were older as they were on long-term (6-to 12-week dosing) saline alone or dexrazoxane. Because the animals were older and SHR develop hypertension and myopathy as they age, they had been considered as probable negatives.

Cartesian coordinates spectrum graphs. For the generation of the SP 2D Cartesian coordinates spectrum graphs we used high-dimensional data produced by high-throughput mass spectrometry consisting of binned data files derived from raw data files generated from SP-MS¹¹. Although the binned process reduces efficiently the number of data points, it is still unmanageable for graph generation. Hence, the number of data points in the binned data files was condensed to 71/36 by including in each new data point the averaged m/z and I values of 100/200 consecutive data points. Each new data point condenses now the information encoded on 100/200 binned data points making the search of Pro-EDICToRs a more tractable problem. Due to the number of data points in the binned data files, the last data point was generated by using the last 105/205. 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. After that, a new averaged and standardized data file is generated consisting of 71/36 data points which can be used now in generating a SP spectrum graph by using a Cartesian 2D/spiral representation. In so doing, a cut off value of 0.5 it is chosen for both m/z and I values related to each averaged data point. This cut off value is used to codify each data point according to their respective average m/z and I values allowing their representation as a node on a Cartesian 2D space. Each data point in the averaged data file is placed in a Cartesian 2D space starting with the first data point at the (0, 0) coordinate. The coordinates of the successive data points are calculated as follows in a similar manner to that for DNA spaces¹²²:

a) Increases in +1 (abscissas) if the absolute m/z value < 0.5 and the absolute I value > 0.5 for a data point (rightwards-step) or:

b) Decreases in -1 (abscissas) if the absolute m/z value > 0.5 and the absolute I value < 0.5 for a data point (leftwards-step) or:

c) Increases in +1 (ordinates) if the absolute m/z and I values > 0.5 for a data point (rightwards-step) or: (upwards-step):

d) Decreases in -1 (ordinates) if the absolute m/z and I values < 0.5 for a data point (downwards-step).

Once we applied the above mentioned transformations we obtained a SP spectrum graph in a Cartesian coordinates 2D space; where each node encode information related to m/z and I values of a condensed spectral region. The **Figure 1** illustrates the appearance of the SP spectrum graphs obtained within the graphical interface of MARCH-INSIDE software.

2D SP-MS node overlapping parameters (nop_k). We used a Markov model (MM) to codify information about SP-MS regions. Specifically, in this work we introduced the 2D SP-MS node overlapping parameters (nop_k) as numerical indices of the SP-MS Cartesian 2D graph. In this study, the stochastic matrix of the classic MARCH-INSIDE approach used for small molecules, RNAs, and proteins has been adapted to characterize the new Cartesian 2D graphs. The method uses essentially three matrix magnitudes:^{61, 123-131}

a) The matrix ${}^1\Pi$ (see Eq. 1). This matrix is built up as a square matrix ($n \times n$). Note that the number of nodes (n) in the graph may be equal or smaller than the number of data points in the MS averaged data files. The matrix ${}^1\Pi$ contains the probabilities ${}^1p_{ij}$ to reach a node n_i with n coincident spectral data points (dp_i) moving throughout a walk of length $k = 1$ from a node n_j with n overlapping spectral data points (dp_j):

$${}^1p_{ij} = \frac{\alpha_{ij} \cdot dp_j}{\sum_{G=1}^n \alpha_{ij} \cdot dp_G} \quad (1) \quad {}^A p_0(j) = \frac{dp_j}{\sum_{G=1}^n dp_G} \quad (2)$$

Where, α_{ij} represents the adjacency relationships between nodes (if n_j is adjacent to n_i then $\alpha_{ij} = 1$; otherwise $\alpha_{ij} = 0$).

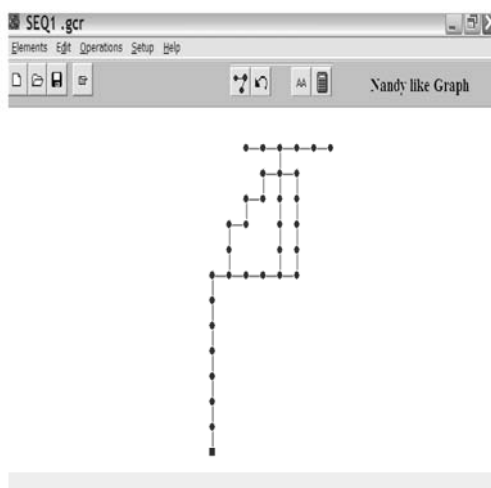


Figure 1. MARCH-INSIDE software view of a SP-MS Cartesian graph.

- b) The spectral data points vector ${}^0\Delta$. The method considers that a number of spectral data points or weight (dp_j) can be assigned to each node. The number of MS data points of the node is equal to the sum of the MS data points coinciding in the same node. So, to retain a more compact matrix notation all dp_j are arranged as a column vector ${}^0\Delta$.
- c) The zero order 2D average spectral proteomic information content vector ${}^0\omega$ (see Eq. 2). This vector lists the absolute initial probabilities ${}^A p_k(j)$ with which a node n_j selected at random encode a given MS proteomic information content. Due to the particularities of the graph representation used here one node can contain information related to the m/z and I of more than one averaged data points. Consequently, the initial absolute probability to encode a given information content of a node depends on the number of data points coinciding on the node n_j (dp_j) and the total number of data points on the spectrum graph (dp_G).

The use of MM theory thus allows calculating the node overlapping parameters (nop_k) for any node n_j that one can reach in the 2D Cartesian graph by moving from any node n_i throughout the entire graph using walks of length k . Considering that the nop_k values are discrete average values we determine them as the sum of two-terms products. The first term is the probability of reaching node n_j by moving from any node n_i throughout walks of length k and the second the number of MS data points coinciding on the node dp_j (see central member of Eq. 3 below).

$$nop_k = {}^0\omega^T \cdot ({}^1\Pi)^k \cdot {}^0\Delta = \sum_{j=1}^n {}^A p_k \cdot dp_j \quad (3)$$

The nop_k encode the properties (m/z and I) of the different MS regions (nodes) in the 2D Cartesian graph. It is remarkable that nop_k can be written using a MM as the product of ${}^0\Delta$ and the natural powers of the matrix ${}^1\Pi$ based on the Chapman–Kolgomorov equations (see right member of Eq. 3 above).¹³²

Statistical Analysis. The nop_k values were calculated with the MARCH-INSIDE (MARKovian CHemicals IN Silico DESign) software¹³³, which is used here for the first time to codify the information content encoded in a SP mass spectrum. MARCH-INSIDE software, specifically the sub-routine BIOMARKS¹³⁴ has been applied intensely to the field of proteins^{61, 123, 124, 135-143}. Using this methodology we can attempt to develop a linear QPTR model to find Pro-EDICToRs with the general formula:

$$CT = b + b_1 \cdot nop_1 + b_2 \cdot nop_2 + \dots + b_k \cdot nop_k = b + \sum b_k \cdot nop_k \quad (4)$$

Here, nop_k values act as the independent or predictive variables. We selected linear discriminant analysis (LDA)¹⁴⁴ to fit the Pro-EDICToRs discriminant function. The QPTR model classifies the rat's SP spectrum into two general groups based on cardiotoxic risk (CTR) indicator. The groups are cardio-toxic risk group (CTR = 1 for definitive and probable positive samples, respectively) and no-cardio-toxic risk or NCTR group (CTR = -1 for definitive and probable negative samples, respectively). In QPTR model, b_k represents the coefficients of the classification function, determined by the least square method as implemented in the LDA module of the STATISTICA 6.0 software package¹⁴⁵.

Best subset method was used for variable selection¹⁴⁶⁻¹⁴⁸. 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 (CTR and NCTR groups) was examined to test discriminatory power of the function developed¹⁴⁹. All the variables included in the QPTR model were standardized in order to bring it into the same scale. Subsequently, a standardized linear discriminant equation that allows to compare their coefficients is obtained¹⁵⁰.

We also inspected the percentage of good classification, samples/variables ratios (ρ parameter), and number of variables to be explored to avoid over-fitting or chance correlation^{146, 147}. Training of the QPTR model was carried out by selecting at random 86 (75%) out of 115 available samples. Specifically, in the training set were included

53 negative samples (21 definitive and 32 probable) and 33 positive samples (25 definitive and 8 probable). The remaining 29 samples (25%), never used for training, were employed to test the Pro-EDICToRs predictive ability of the QPTR model. This prediction set was composed by 18 negative samples (7 definitive and 11 probable) and 11 positive samples (9 definitive and 2 probable).

Principal Component Analysis (PCA). We used as inputs for the PCA the five values of nop_k (nop_0 , nop_1 , nop_3 , nop_4 , nop_5) selected as the more important in the LDA analysis (see Results section). We selected specifically these values because they proved to efficiently detect Pro-EDICToRs separating CRT from non-CRT samples. In total we used as input for PCA analysis a dataset composed by the five values of nop_k for the 115 samples, which represents 575 data points. All the analysis was developed with the software STATISTICA, using the default parameters. The higher variance explained with the lower possible number of components was the criteria used to stop the process of inclusion of new components. The amount of variance explained and the absolute value for the respective eigenvalue were used to rank the importance of a component. The loadings of each sample for the main principal components or factor (F_1) were used to rank the samples.¹⁵¹⁻¹⁵³

Complex Network Analysis. In order to construct a Complex Network of MS blood proteome samples for the study of Pro-EDICToRs we carried out the following steps:

1. First, we selected as inputs the five values of nop_k (nop_0 , nop_1 , nop_3 , nop_4 , nop_5) selected as the more important in the LDA analysis (see Results section).¹³³

2. We calculated the contributions of each nop_k to the CRT probability for each one of the 115 samples by substituting the molecular descriptors into the QPTR equation using the Microsoft Excel application.¹⁵⁴ In means that, we calculated the result of the multiplication of each nop_k for every sample by its own coefficient (b_k) in the QPTR model (see above). Consequently, in so doing we transformed the previous data set of 575 data points (five nop_k for each one of the 115 samples) into another data with the same dimensions but containing weighted $*nop_k = b_k \cdot nop_k$ values instead of the original nop_k values.

3. All the contributions $*nop_k$ predicted were used as input for the software STATISTICA employed to calculate the blood proteome i^{th} -sample/ j^{th} -sample Regression coefficients (R_{ij}). These coefficients were used as pair similarity measures and represented actually as a 115×115 sample-sample R-matrix. This matrix was derived using the Basic statistics module of STATISTICA.^{153, 155}

4. Using Microsoft excel¹⁵⁴ again we transformed the matrix derived with STATISTICA into a Boolean matrix. The elements of this matrix are equal to 1 if two proteome samples have a sample-sample R_{ij} lower than certain cut-off. This cut-off or threshold value used was selected in such a way that it minimizes the average node degree given that we guarantee 0% of disconnected nodes in the total network including all CTR and NCTR cases.¹⁵⁵

5. The line command used in Excel to transform the distance matrix into a Boolean matrix was $f = \text{if}(A\$1 = \$B2, 0, \text{if}(B2 > 0.9999971, 0, 1))$. It allows transforming distance into Boolean values and equals the main diagonal elements to 0 avoiding loops in the future network. The Boolean matrix was saved as a txt format file.

6. After, renamed the .txt file as a .mat file we read it with the software CentiBin.^{156, 157} Using CentiBin we can either represent the network or highlight all samples (nodes) connected to a specific sample and calculating many parameters including node degree.

7. The large Complex Network derived was split into two sub-networks (CRT and non-CRT samples). Next, the Pajek software^{158, 159} was used to calculate different network topological indices (TIs) in order to compare the Pro-EDICToRs patterns in both networks. The TIs calculated with Pajek were: the number of nodes (n), the Total adjacency index or the same: number of edges (m), The Zagreb group index (M1), The Zagreb group index (M2), The Randic connectivity index (X_r), The Platt index (F), Index of relinking (P).²⁰ The comparison was based on the value of the relative difference ($D\%$) between the CTR with respect to NCTR sub-network for each TI. The values of $D\%(TI)_{CTR,NCTR}$ were calculated as follows: $D\%(TI)_{CTR,NCTR} = [TI(CTR) - TI(NCTR)] \cdot 100 / TI(NCTR)$.

8. CentiBin software was used to generate ideal random networks by four different algorithms with the aim of comparing these networks with the predicted ProEdictors sub-networks of CTR and non-CTR class. The ideal networks list include: Barabasi-Albert random network, Kleinberg small world network (SWN), Erdos-Renyi network and Epsstein power law network (PLN). All these ideal networks were generated with a number of

nodes between the number of nodes of the CTR and non-CTR sub-networks.¹⁵⁷ The comparison was based on the value of the relative difference (D%) between the actual network a (CTR or NCTR) with respect to the ideal i^{th} -sub-network for each TI. The values of $D\%(TI)_{a,i}$ were calculated as follows: $D\%(TI)_{a,i} = [TI(CTR) - TI(NCTR)] \cdot 100 / TI(NCTR)$.

Partial Order Analysis. We used three sample attributes as inputs for the construction of alternative two-attributes POs schemes: CTR posterior probabilities predicted with the LDA model, PCA F_1 scores, and Complex Network node degrees for each sample x_i (see the three previous sections). Different test and statistics were calculated to compare and assess the quality of these alternative POs, including the two more important: $T(g_1, g_2)$ index: Tanimoto's coefficient and χ^2 : Chi-square. The Chi-square is a classic statistic and the $T(g_1, g_2)$ indices for two alternative POs of two posets (A and B) can be calculated as follows:^{107-111, 113-116, 160-164}

$$T(g_1, g_2) \equiv \frac{A \cap B}{A \cup B} = \frac{\sum_{sr}}{\sum_{sr} + \sum_{rr} + g_1 \cdot \sum_{irA} + g_2 \cdot \sum_{irB}} \quad (5)$$

Where, g_1 and g_2 are weights that can take the values of either 0 or 1 and to explain the above equation Sorensen and Burggemann *et al.* introduced the following notations for comparable ($<_A, <_B, \leq_A, \leq_B$) or incomparable ($\|_A^A, \|_B^B$) elements in posets A and B. Being, $<$ and \leq the classic symbols of less than and less or equal than and $\|$ the symbol introduce to demark not-comparable features.

$$\sum_{sr} : \text{sum of pairs } x_i <_A x_j \text{ and } x_i <_B x_j \Leftrightarrow \text{same ranking} \quad (5.1)$$

$$\sum_{rr} : \text{sum of pairs } x_i <_A x_j \text{ and } x_j <_B x_i \Leftrightarrow \text{reverse ranking} \quad (5.2)$$

$$\sum_{irA} : \text{sum of pairs } x_i \leq_A x_j \text{ and } x_i \|_B^B x_j \Leftrightarrow \text{incomplete ranking in A} \quad (5.3)$$

$$\sum_{irB} : \text{sum of pairs } x_i \|_A^A x_j \text{ and } x_i \leq_B x_j \Leftrightarrow \text{incomplete ranking in B} \quad (5.4)$$

Other important statistics reported were the P(IB): Stability of ranking, d(N): Diversity, t(N): Selectivity, NL: Number of Levels, NEL: Number of Elements in the Largest Level, V(N): Comparability, U(N): Contradictions, K(N): Level of degeneracy, NEC: Number of equivalent classes, and C: Complexity, see references for details. We carried out the PO analysis using the software Hasse for Windows (WHASSE), which was kindly released by Prof. R. Bruggemann.^{107-110, 113, 115, 116}

3. Results

QPTR model for Pro-EDICToRs study. LDA and other predictors combined with a clustering technique such as PCA can be used for data processing in bioinformatics and proteomics including the compression and classification of large MS data.^{62, 165-167} For this kind of analysis the MS data points can be used directly, without transformation, or previously transformed. For instance, Lilien *et al.* have developed an algorithm called Q5 for probabilistic classification of healthy versus disease whole serum samples using MS. The algorithm employs PCA followed by LDA on whole spectrum surface-enhanced laser desorption/ionization time of flight (SELDI-TOF) MS data and was demonstrated on four real datasets from complete, complex SELDI spectra of human blood serum. Replicate experiments of different training/testing splits of each dataset were employed to verify robustness of the algorithm. The probabilistic classification method achieved excellent performance with sensitivity, specificity, and positive predictive values above 97% on three ovarian cancer datasets and one prostate cancer dataset. The Q5 method outperforms previous full-spectrum complex sample spectral classification techniques and can provide clues as to the molecular identities of differentially expressed proteins and peptides.¹⁶⁸

In the present work we are proposing the use of high-throughput MS graph theory parameters instead of PCA to reduce data dimension and later combine it with LDA in the field of Toxicoproteomics. To illustrate the

potentialities of this approach on the early detection of Pro-EDICToRs research we decided to develop a QPTR model based on graph theoretical indices derived from the Nandy like graph representation of the SP-MS proposed above, which has been used for DNA sequences.^{169, 170} The nop_k values proposed above are used here as numerical indices of the SP-MS Cartesian graph in the development of the QPTR equation described below:

$$CTR = 0.861 \cdot nop_0 - 17.394 \cdot nop_1 + 84.56 \cdot nop_3 - 80.24 \cdot nop_4 + 13.16 \cdot nop_5 - 0.67 \quad (6)$$

$$N = 86 \quad F = 5.54 \quad D^2 = 1.43 \quad U = 0.74 \quad p = 0.0002 \quad \rho = 7.17$$

As shown in **Table 1**, this model predicts correctly 80.23% of cases (69 out of 86 samples) used for training. Specifically, 24 out of 33 CTR samples (sensitivity = 72.73%) and 45 out of 53 NCTR samples (specificity = 84.91%) were classified correctly, respectively. The statistical significance of the model was evaluated through a 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 number of cases and adjustable parameters. A satisfactory ρ value ($7.17 > 4$) discard any possibility of over-fitting. On the other hand, the square of Mahalanobis's distance (D^2) and Wilk's U statistic provide a measure of the model's discriminatory power, respectively. These values indicate a statistically significant separation of the two groups (CTR and NCTR) by the LDA technique, despite of the complex nature of the discrimination problem under consideration.^{171, 172}

Table 1. QPTR Model's performance on training and test sets, respectively

Training Set classification matrix ^a				Parameter (%)	Test Set classification matrix ^a			
	NCTR	CTR			NCTR	CTR		
NCTR	45	8	84.91	Specificity	72.22	13	5	NCTR
CTR	9	24	72.73	Sensitivity	72.73	3	8	CTR
			80.23	Accuracy	72.41			

^a Row entries are the observed frequencies and column entries the predicted ones.

This is a logic result for a SP-MS since the number of protein related to a toxic event is presumed to be insignificant in relation to the total number of serum proteins. The values of the respective nop_k computed and included on Eq. (5) for all the cases belonging to both CTR (represented by an open square) and NCTR (represented by a solid square) groups were plotted to illustrate this point (**Figure 2** depicts one of these plots as matter of example).

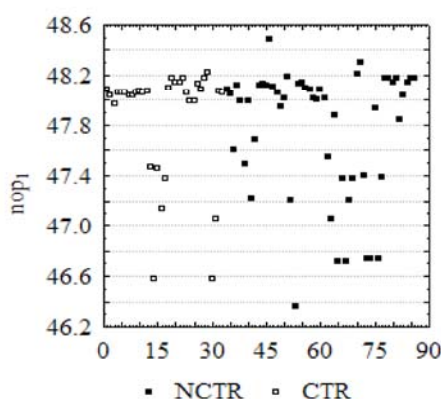


Figure 2. Example of scatter plots for one variable (nop_1) included on QPTR equation.

In addition, the receiver operating characteristic curve (ROC curve) obtained, indicate that the model is not a random, but a statistically significant Pro-EDICToRs' classifier (see **Figure 3**). A ROC curve plots the Sensitivity vs. one minus the Specificity. An ideal classifier hugs the left side and top side of the graph, and the area under the curve is 1.0. A random classifier should achieve approximately 0.5.^{173, 174}

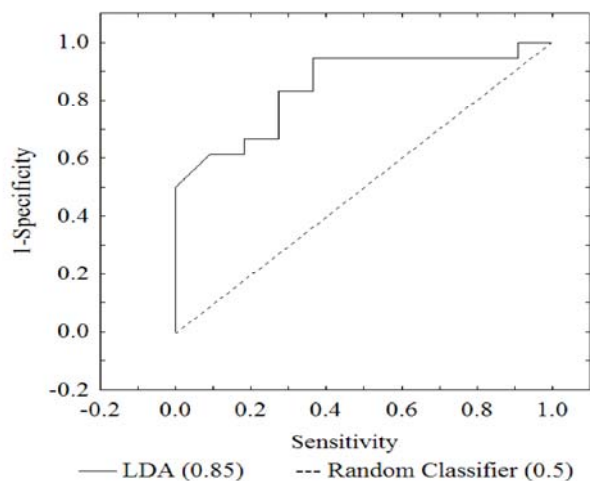


Figure 3. ROC curve related to QPTR equation

The next step is to find out whether or not the model fulfill the basic assumptions of LDA^{147, 149}. On the case of severe violations, the reliability of the Pro-EDICToRs predictions may be compromised. 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 **Figure 4**) supports this choice as no systematic pattern is seen¹⁴⁹.

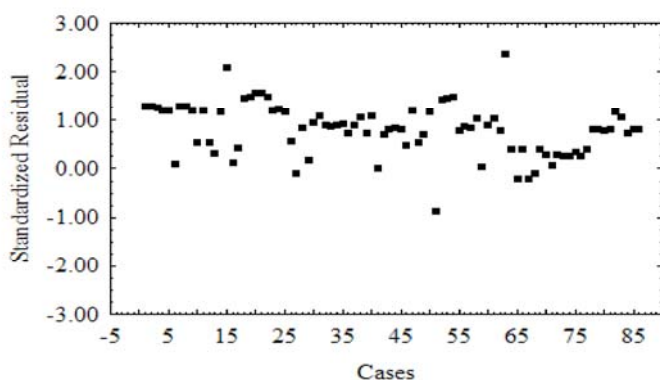


Figure 4. Plot of residuals vs. case number

The parametric assumption of homocedasticity (*i.e.*: homogeneity of variance of the variables) was also checked out by simply plotting the square standardized residuals for each predictor variable¹⁴⁹. These plots reveal an adequate scatter on the points, without any consistent pattern, validating *a posteriori* the pre-adopted assumption of homocedasticity (see **Figure 5** to visually inspect one of these plots). 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.

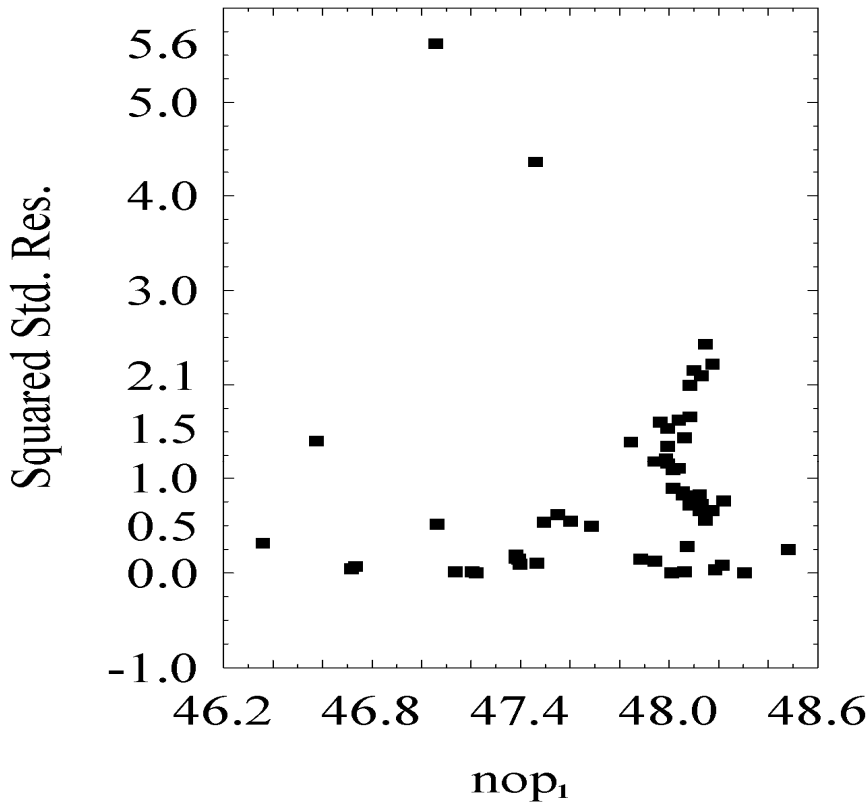


Figure 5. Example of plot for square residual vs. nop_k included on QPTR equation

Moving on to the next important parametric assumption of LDA, *i.e.*: normality of residuals, it was found that the residuals exhibit adequate values of skewness (-0.18) and kurtosis (0.86)¹⁴⁹, which is a characteristic of a normal distribution fitting. Additionally, the hypothesis of normality of residuals is confirmed by a Kolmogorov-Smirnov statistic ($D = 0.95$) with a p-level < 0.2 , a Shapiro-Wilk's statistic ($W = 0.98$) with $p = 0.16$, and Lilliefors hypothesis test with $p < 0.1$. See also the distribution histograms in **Figure 6**.

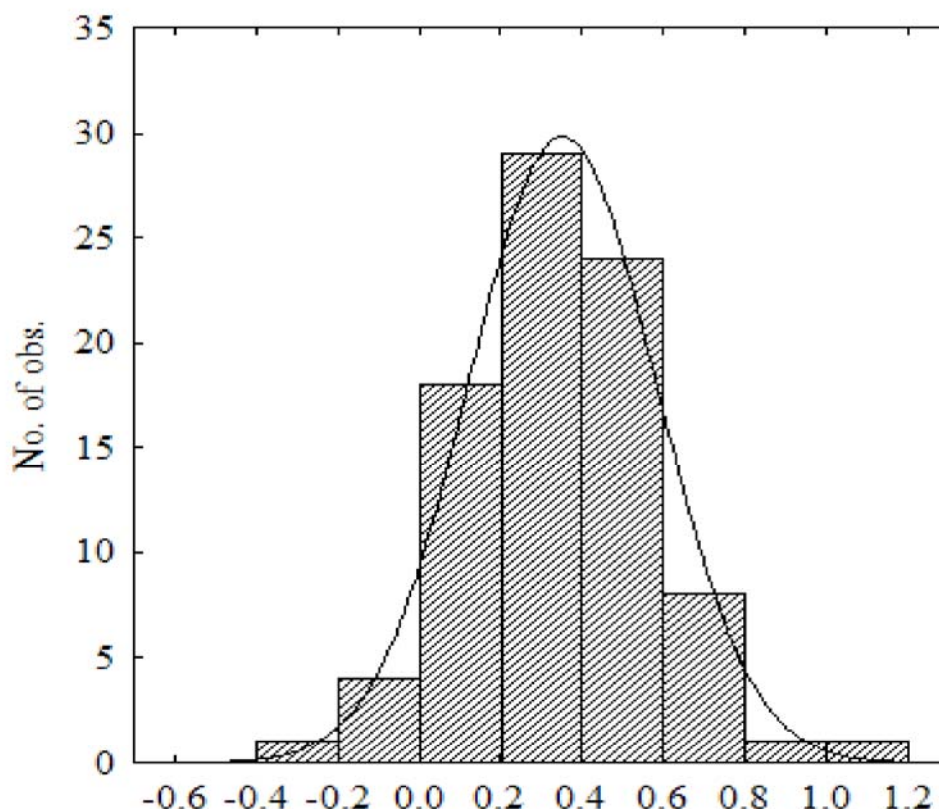


Figure 6. Histogram of QPTR equation residuals distribution

Finally, we detected certain co-linearity exhibited by the variables included in the model (pair correlation between some variables higher than 0.7). However, this fact in general do not inhibit the model's ability to make inferences about the mean responses or predictions of new observations¹⁷⁵. In this sense, we assessed the predictive ability of the model by using an external test set, never used to train the model. The proposed model was able to classify correctly 21 out of 29 test samples (global predictability = 72.41%). In particular, 8 out of 11 CTR samples (sensitivity = 72.73%) and 13 out of 18 NCTR samples (specificity = 72.22%) were classified correctly (see **Table 1**).

PCA study of ProEdictors. As explained above LDA combined with PCA can be used for data processing in bioinformatics and proteomics including the compression and classification of large MS data.^{62, 165, 168} We substituted above PCA by the graph theoretical parameters nop_k in order to codify MS information. In any case, we can still use PCA for other goals. For instance, a classic use of PCA is the construction of low dimension spaces for systems structural parameters. The PCA scores derived for each system, sample, or case (drugs, proteins, proteomes, and brain cortex regions) can be used to rank the importance of each sample with respect to the others. System can be understood here in the broader sense including drugs, proteins, or brain connectivity.^{61, 123, 151, 176-181} In this work we performed PCA on the $*nop_k$ values for the 115 blood proteome samples to study the Pro-EDICToRs patterns extracting the first four principal components (F_1, F_2, F_3, F_4 , and F_5). The **Figure 7**

illustrates the fade of the explained variance with respect to the number of order of the components (1, 2, 3, 4, and 5).

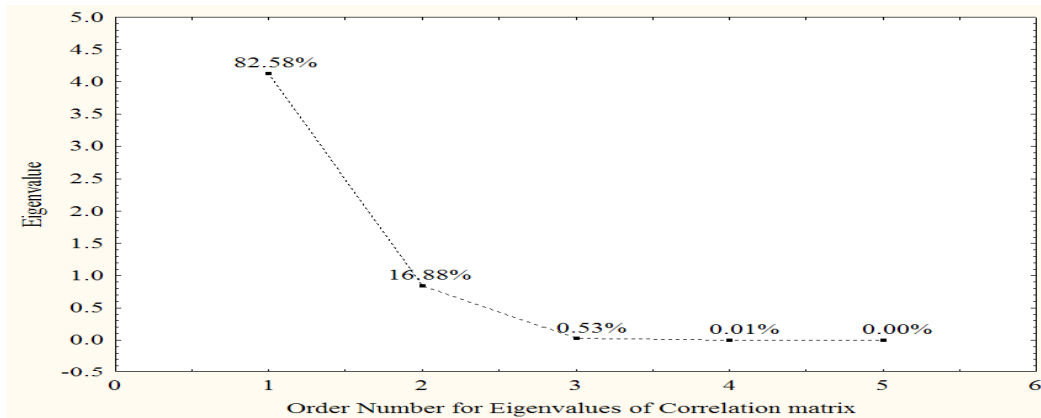


Figure 7. Plot of eigenvalues of factors *vs.* number of order (PCA variance explained is used as point label).

The first principal component or factor (F_1) with eigenvalue 4.1 explained the 82.7% of variance of all the data. The second principal component (F_2) explains 16.88% and the third only 0.53% of variance, which justify the use of only F_1 to rank the samples in a 1D order if necessary and only the two first principal components (F_1 and F_2) to construct a ProEDICToRs 2D-PCA space. The **Figure 8** depicts the 2D-PCA space, which illustrates the distribution of training, validation and/or CTR or non-CTR samples.

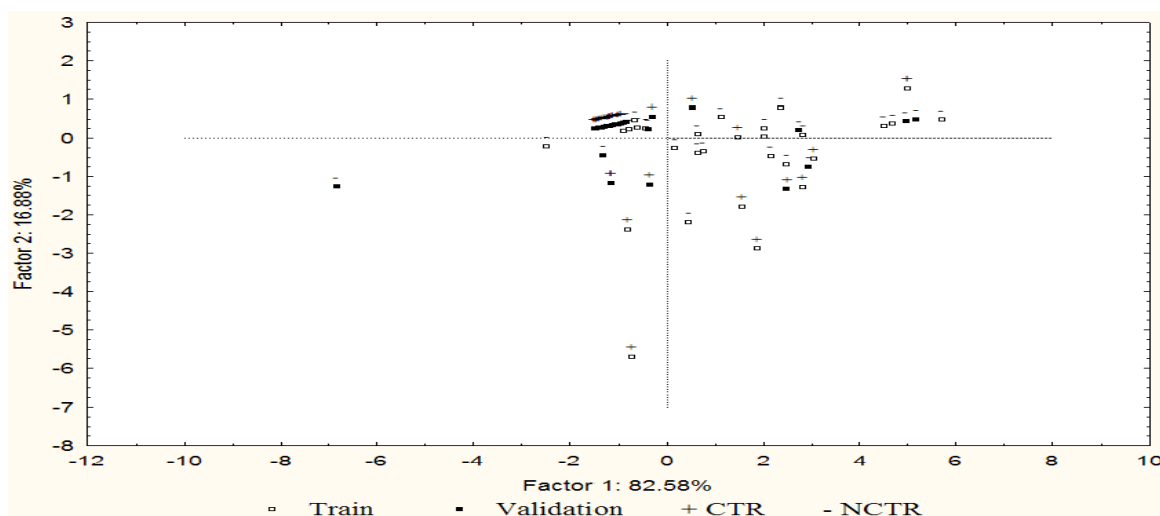


Figure 8. 2D-PCA space for the 115 samples.

Complex network study of ProEdictors. Next, all the nop_k values present in the QPTR model were used to construct by the first time a Pro-EDICToRs Complex Network (CN). As discussed above the nodes of the networks are the samples, which may vary from drugs, proteins, or tissues, to individuals in social networks; and the edges indicate high similarity between two samples or cases.¹⁸²⁻¹⁸⁵ First we constructed a general Pro-EDICToRs network using all the 115 samples. The average node degree was 47.48 with a threshold cutoff of $R = 0.9999971$, which guarantee 0% of disconnected nodes in the total network including all CTR and NCTR cases. Next, in order to perform a comparative study of overall connectivity patterns predicted for both contrasting groups (cardiac toxicity and control samples) we split the network into two sub-networks¹⁸⁶: the CTR and the non-CTR or NCTR sub-network. The **Figure 9** and the **Figure 10** graphically illustrates both sub-networks. We compared the TIs of both sub-networks; finding an average relative difference of $D\%(TI)_{CTR,NCTR} = 38.6\%$ for 7 TIs studied, including the small/higher differences of 9.5 and 54.2 % for the re-linking P and Zagreb M2 indices respectively.²⁰ The **Table 2** summarizes the results for this comparison. Between the TIs reported we can find n (the number of samples). Considering that we built the general network with a cutoff that guarantee 100% of connected nodes the difference $D\%(n)_{CTR,NCTR} = 38.0\%$ between both sub-networks with respect to n reflects only differences in the large of the number of samples used for the analysis. Consequently, for this specific case $D\%(n)_{CTR,NCTR}$ does not reflect any important topological similarity or dissimilarity between both sub-networks. Conversely, the number of links (blood proteome sample-sample pairs) in the network clearly depends in this case on the procedure used to construct the network (MS graph parameters, similarity measure, cutoff selected).¹⁵⁵

Table 2. Results of the Complex network study of CTR vs. NCTR cases

Parameter ^a	CTR sub-network	NCTR sub-network	D%(TI) _{CTR,NCTR} ^b
n	44	71	38.0
m	533	837	36.3
M1	33808	61380	44.9
M2	540686	1187725	54.5
Xr	20.00	34.22	41.6
F	32742	59706	45.2
P	11.29545	10.32	9.5

^aThe number of nodes (n), Total adjacency index or the same: number of edges (m), The Zagreb group index (M1), The Zagreb group index (M2), The Randic connectivity index (Xr), The Platt index (F), Index of relinking (P). ^bD% is the relative difference between the Topological Index (TI) for the sub-network CTR with respect to NCTR sub-network and was calculated as follows: $[TI(CTR) - TI(NCTR)] \cdot 100 / TI(NCTR)$. The average node degree was 47.48 with a threshold cutoff of $R = 0.9999971$ to guarantee 0% of disconnected nodes in the total network including all CTR and NCTR cases.

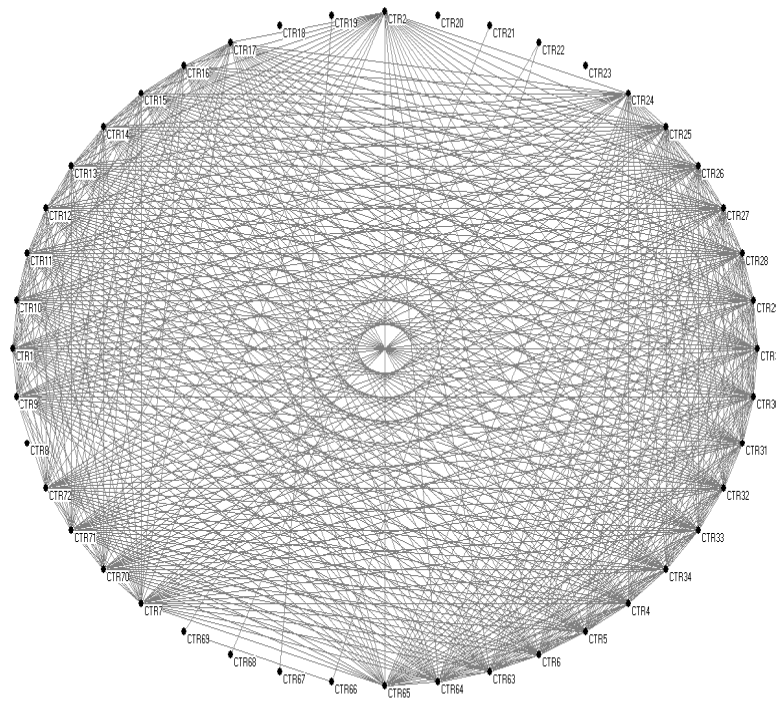


Figure 9. Complex Pro-EDICToRs sub-network for CTR samples

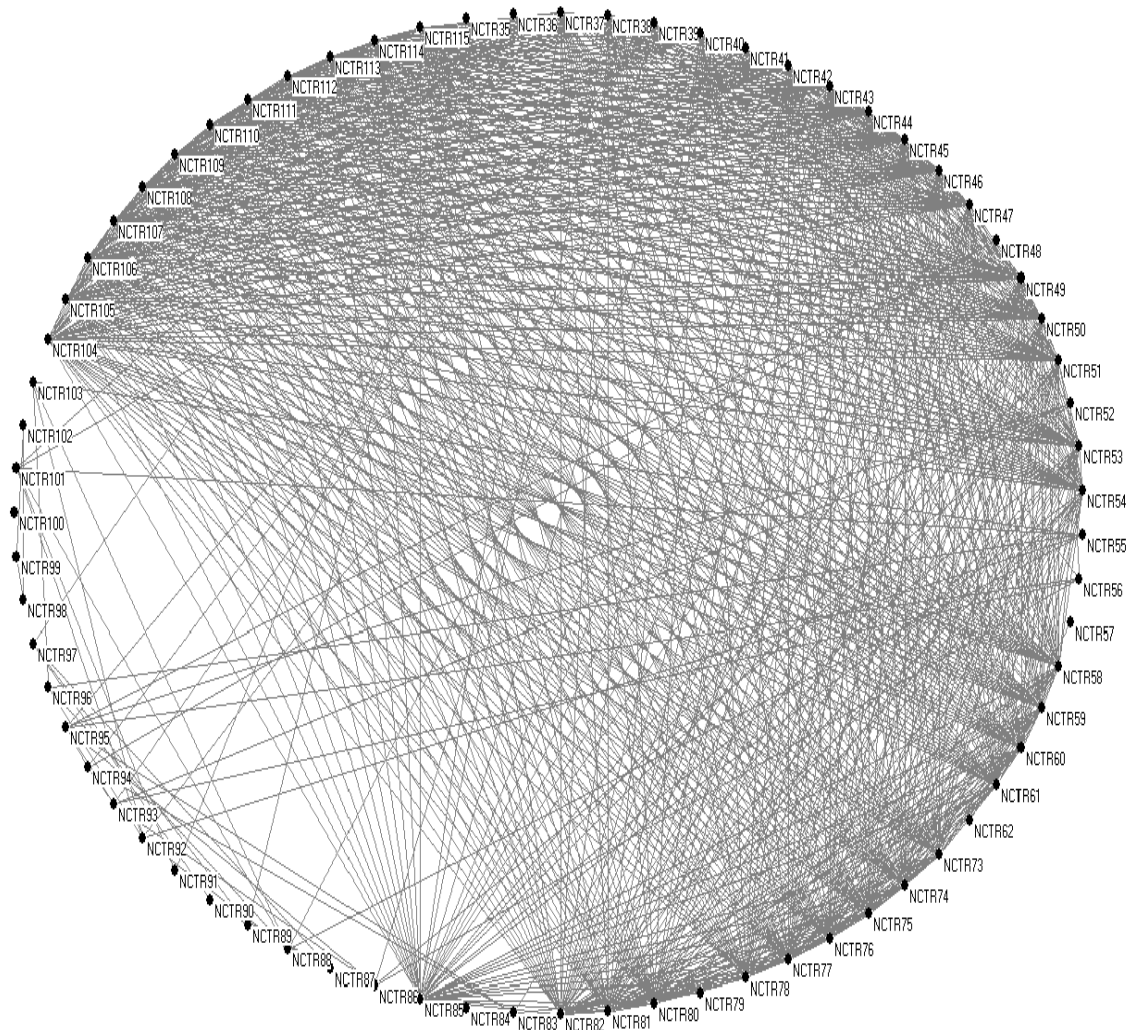
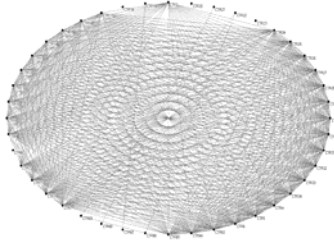
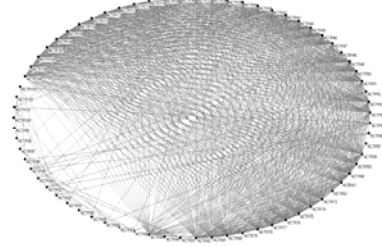
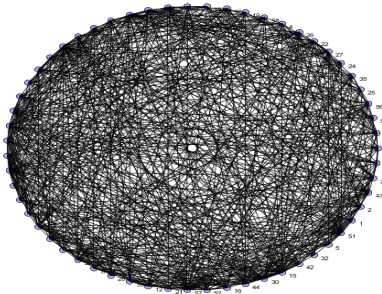
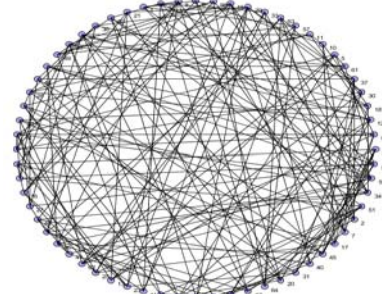
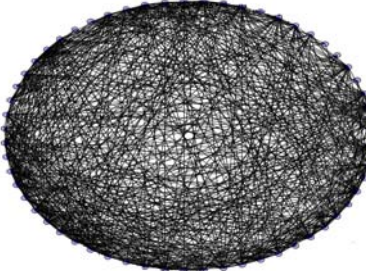
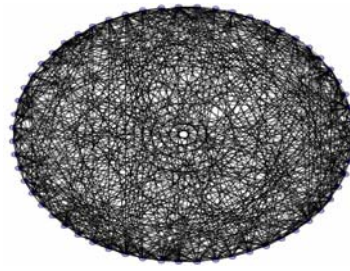


Figure 10. Complex Pro-EDICToRs sub-network for NCTR samples

We also compared these networks with well known models of random networks such as: Barabasi-Albert, Kleinberg Small World, Erdos-Renyi, and Epsstein Power Law, finding similarities in some cases.¹⁸⁷ The **Table 3** illustrates the pictures of the networks derived with the software CentiBin. This table also summarizes the basic TIs for all the networks. The ideal random networks derived have different remarkable properties. For instance the Kleinberg network is by definition a small-world network having a relative low average topological distance $Dist = 2.6$ despite its large diameter ($D = 4$) and very low average node degree ($\delta = 6$). For instance, the rest of the ideal random networks present notably higher node degrees ($\delta > 20$) in all cases despite having lower diameter ($D = 2$) in all cases.

Table 3 Comparison with some ideal random network models

CTR sub-Network	Value	TIs	Value	Non-CTR sub-Network
	44	n	71	
	533	m	837	
	1056	W	1560	
	1.0	D	1.0	
	24.2	δ	23.6	
	1.0	Dist	1.0	
Barabasi Albert Random Network	Value	TIs	Value	Kleinberg Small World Network
	59	n	64	
	631	m	192	
	5582	W	10438	
	2	D	4	
	21	δ	6	
	1.6	Dist	2.6	
Erdos Renyi Random Network	Value	TIs	Value	Epsstein Power Law Network
	59	n	59	
	640	m	648	
	5554	W	5548	
	2	D	2	
	22	δ	22	
	1.6	Dist	1.6	

^a The TIs used are: number of nodes (n), number of edges (m), Wiener index (W), diameter (D), and the network average values for node degree (δ), topological distance (Dist).

As often happen in science, the non-ideal or real-world networks resemble some models of ideal networks in some aspects and others models in other features.^{188, 189} For instance, both CTR and NCTR sub-networks have very low topological distances (Dist = 1) like in small-world networks^{190, 191} but have a notably low diameter (D = 1) and high node degrees ($\delta_{CTR} = 24.2$ and $\delta_{NCTR} = 23.6$). Notably the Wiener index, which encodes network graph branching, is comparatively low for our non-ideal network ($W_{CTR} = 1056$ and $W_{NCTR} = 1056$). These values are low not only compared with respect to the small-world network (W = 10 438) but with respect to the remnant networks too. The **Table 4** summarizes the relative differences of CTR and NCTR sub-networks with other networks $D\%(TI)_{CTR,i}$ and $D\%(TI)_{NCTR,i}$ as well as the average values of each differences across all networks families or TIs types.

Table 4. Summary of the comparative study of the actual vs. ideal networks

Network TIs	TI average	Barabasi-Albert	Kleinberg	Erdos-Renyi	Epsstein
$D\%(TI)_{CTR,i}$					
n	26.9	25.4	31.3	25.4	25.4
m	56.9	15.5	177.6	16.7	17.7
W	83.2	81.1	89.9	81.0	81.0
D	56.3	50.0	75.0	50.0	50.0
δ	84.6	15.2	303.3	10.0	10.0
Dist	44.2	38.7	61.2	38.7	38.3
$D\%(TI)_{CTR,i}$	Network average	37.7	123.1	37.0	37.1
$D\%(TI)_{non-CTR,i}$					
n	18.0	20.3	10.9	20.3	20.3
m	107.1	32.6	335.9	30.8	29.2
W	75.2	72.1	85.1	71.9	71.9
D	56.3	50.0	75.0	50.0	50.0
δ	80.1	12.4	293.3	7.3	7.3
Dist	44.2	38.7	61.2	38.7	38.3
$D\%(TI)_{non-CTR,i}$	Network average	37.7	132.5	36.7	36.6

^a The TIs used are: number of nodes (n), number of edges (m), Wiener index (W), diameter (D), and the network average values for node degree (δ), and topological distance (Dist); $D\%(TI)_{n,i} = (TI_n - TI_i) \cdot 100 / TI_i$, where TI refers to an specific topological index, n points to the predicted sub-network (CTR or non-CTR) and i to the corresponding random network (Barabasi-Albert, Kleinberg, Erdos-Renyi).

Partial Order of ProEDICToRs. Ordering of samples may be very useful for the MS-based classification and comparison of blood proteome samples for the study of drug-induced cardiac toxicity. In principle, we may propose different 1D alternative orders for all the 115 samples (x_i). These orders may be constructed based on different sample features. In this study is easy to realize that some of the parameters calculated above can play the role of ranking attributes by themselves to order the samples. For instance, a total 1D order may base on least three different inputs: QPTR model LDA probabilities, PCA F_1 -scores and/or Complex network node degree. In any case, a total order based on one single parameter is less rich in information content and may easily fail in capturing all the biologically remarkable sample characteristics due to the high complexity of the blood proteome samples. Consequently, is more reasonable to construct a 2D order of samples based on more than one feature at time. In PO theory one may use different combinations of sample features to construct the PO scheme. The general principle is to order or rank different elements or samples (x_i) using multiple ranking attributes or sample features. Consequently, the study of the best set of attributes or sample features used to build the PO becomes of the major importance. ^{100-105, 109, 111, 112, 163, 164, 192-198}

The **Table 5** depicts general statistics of different POs based on two or the three parameters for all the samples or both groups of samples (CTR vs NCTR) separately as well as.

Table 5. General statistics (All samples, CTR, and NCTR) for different PO schemes.

Parameter ^a	Comparing posets by Tanimoto's analysis								
	PCA vs. CN			LDA vs. CN			LDA vs. PCA		
	CTR	NCTR	All	CTR	NCTR	All	CTR	NCTR	All
T	0.75	0.78	0.77	0.67	0.40	0.47	0.55	0.82	0.70
$ CTR \cap NCTR $	411	909	1320	411	909	1320	411	909	1320
$CTR \Delta NCTR$	137	247	384	200	1344	1544	332	199	531

^a T is the Tanimoto's coefficient and $CTR \Delta NCTR = |CTR| + |NCTR| - 2|CTR \cap NCTR|$

In our case, is desirable a PO scheme maximizing the separation of CTR from NCTR samples and the concordance between the ordering ranks based on different features. In this sense, we used Tanimoto and Chi-square analysis to compare the different POs proposed. The similarity measure denoted the Tanimoto index, $T(g_1, g_2)$ may substitutes the Spearman's rank correlation in this cases when more than one attribute is used simultaneously. PO of samples forbids the use of the Spearman's coefficient, which is applicable only to orders using the same single variable.¹⁶⁴ In **Table 5** one can note high PO overall similarity between the orders of PCA vs. CN as well as LDA vs. PCA with total Tanimoto's coefficients of 0.77 and 0.70 respectively. In any case, the PCA-CN based PO is more coherent in ordering Pro-EDICToRs with coefficients equal to 0.75 and 0.78 for CTR and NCTR samples respectively. Conversely, LDA-PCA based PO is not very equilibrated presenting high similarity in NCTR samples ordering (0.82) but low similarity in CTR samples ordering (0.55). A more extensive characterization of these alternative POs is reported in **Table 6**.

Table 6. Other important statistics for different PO schemes.

Parameters ^a	Cases for Tanimoto's analysis		
	PCA vs. CN	LDA vs. CN	LDA vs. PCA
T(All)	0.77	0.47	0.70
χ^2	2.81	-0.11	3.70
P(IB)	0.53	0.21	0.47
d(N)	0.1	0.07	0.09
t(N)	0.39	0.53	0.32
NL	23	31	23
NEL	6	4	7
V(N)	3269	5339	3618
U(N)	6908	2802	6116
K(N)	336	370	242
NEC	59	59	73
C	Yes	No	Yes

^a T(all): Overall Tanimoto's coefficient, χ^2 : Chi-square statistic, P(IB): Stability of ranking, d(N): Diversity, t(N): Selectivity, NL: Number of Levels, NEL: Number of Elements in the Largest Level, V(N): Comparability, U(N): Contradictions, K(N): Level of degeneracy, NEC: Number of equivalent classes, and C: Complexity.

Inspection of **Table 6** not only confirms that PCA-CN and LDA-PCA based POs orderings are the more coherent with the higher Chi-square coefficients (2.81 and 3.7) but shows similar trends for both POs in many other statistic parameters. In our opinion, we recommend to use as the first alternative LDA-PCA based PO considering also that LDA probabilities give more direct prediction of CTR vs. NCTR classification. The **Figure 11** illustrates the interface of the WHASSE software depicting a HASSE diagram with the PO of blood proteome samples.

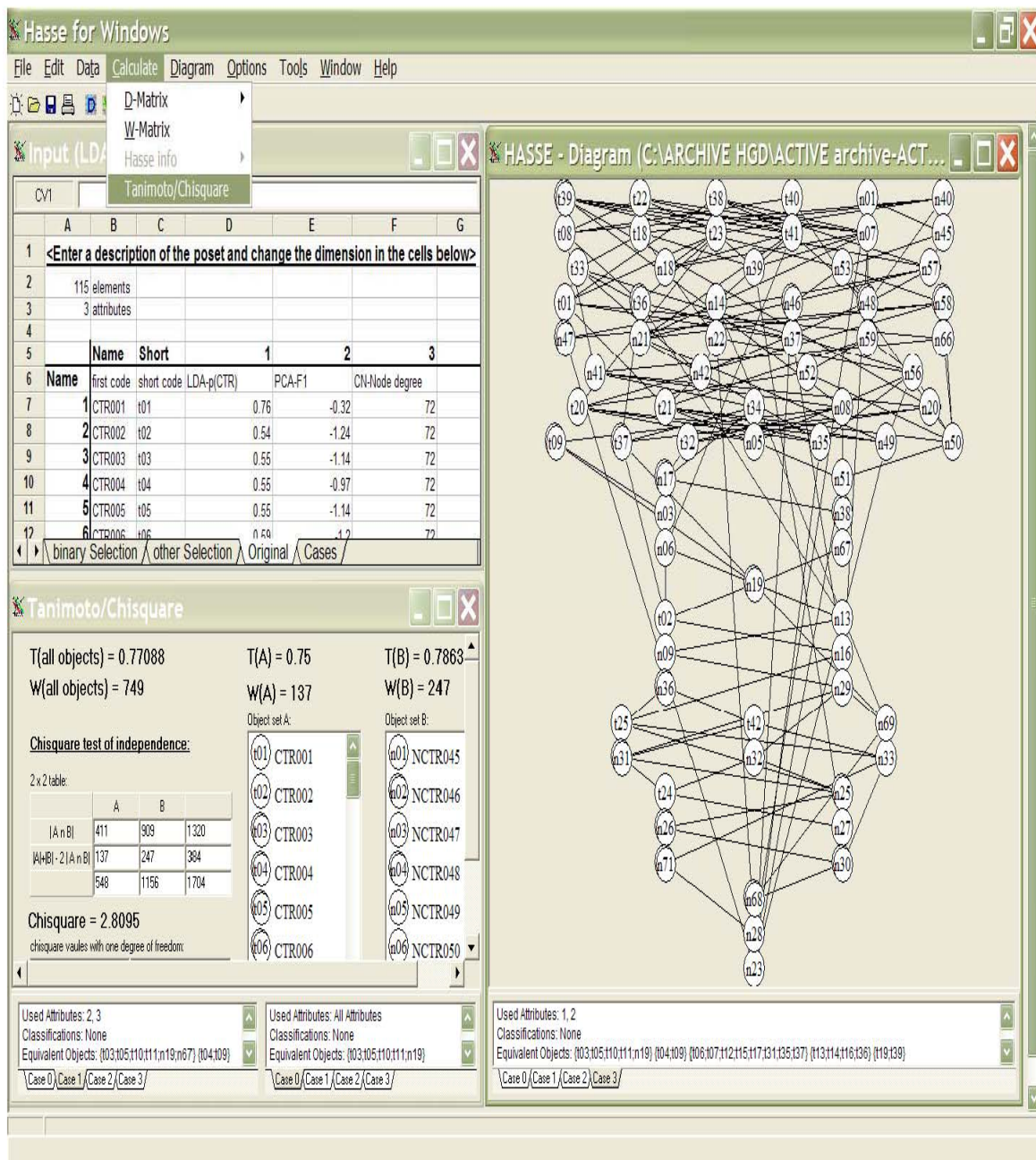


Figure 11. View of the PO analysis performed with WHASSE software

4. Conclusions

The node-overlapping parameters (nop_k) derived from a Cartesian spectrum graph for the SP-MS are numerical indices very useful to derive QPTR models with LDA technique. These QPTR models can be employed for the proteome based early detection of drug-induced cardiac toxicities, which we called Pro-EDICToRs. The nop_k values can be also used as inputs on other studies including PCA data reduction and Complex Network construction for Pro-EDICToRs data. The outputs of these studies (LDA, PCA, and Complex networks) can be used as order ranking attributes in PO analysis. It allows ordering and comparison of blood proteome data. The present result opens a new door to the application of SP-MS graph parameters to toxicoproteomics in the near future. After this work one may conclude that in general graph-based QPTRs will help to unravel proteome-disease relationships hidden in the SP-MS, which cannot be explained by a single biomarker.

Acknowledgments. The authors thank projects funded by the Xunta de Galicia (PXIB20304PR and BTF20302PR) and the Ministerio de Sanidad y Consumo (PI061457). González-Díaz H. acknowledges tenure track research position funded by the Program Isidro Parga Pondal, Xunta de Galicia. González-Díaz H. thanks Prof. Bruggemann R. by kindly release Hasse for Windows (WHASSE). Cruz-Montegudo M. thanks undergraduated doctoral scholarship from FCT as well as, support from Department of Organic Chemistry, Faculty of Pharmacy, University of Porto, Portugal.

Supporting Information Available: The values of the five variables in the model for the 115 samples used, their observed and predicted classifications, as well as the predicted CTR/NCTR probabilities are depicted in a supplementary material related to this work. This material is available upon authors request.

References

1. van Dalen, E. C.; van den Brug, M.; Caron, H. N.; Kremer, L. C., Anthracycline-induced cardiotoxicity: comparison of recommendations for monitoring cardiac function during therapy in paediatric oncology trials. *Eur J Cancer* **2006**, 42, (18), 3199-205.
2. Jones, R. L.; Ewer, M. S., Cardiac and cardiovascular toxicity of nonanthracycline anticancer drugs. *Expert Rev Anticancer Ther* **2006**, 6, (9), 1249-69.
3. Urbanova, D.; Urban, L.; Carter, A.; Maasova, D.; Mladosevicova, B., Cardiac troponins--biochemical markers of cardiac toxicity after cytostatic therapy. *Neoplasma* **2006**, 53, (3), 183-90.
4. Ward, J. B., Jr.; Henderson, R. E., Identification of needs in biomarker research. *Environ Health Perspect* **1996**, 104 Suppl 5, 895-900.
5. Anderson, N. L.; Anderson, N. G., The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **2002**, 1, (11), 845-67.
6. Liotta, L. A.; Ferrari, M.; Petricoin, E., Clinical proteomics: written in blood. *Nature* **2003**, 425, (6961), 905.
7. Mehta, A. I.; Ross, S.; Lowenthal, M. S.; Fusaro, V.; Fishman, D. A.; Petricoin, E. F., 3rd; Liotta, L. A., Biomarker amplification by serum carrier protein binding. *Dis Markers* **2003**, 19, (1), 1-10.
8. Hu, S.; Loo, J. A.; Wong, D. T., Human body fluid proteome analysis. *Proteomics* **2006**, 6, (23), 6326-53.
9. Kantor, A. B., Comprehensive phenotyping and biological marker discovery. *Dis Markers* **2002**, 18, (2), 91-7.
10. McDonald, W. H.; Yates, J. R., 3rd, Shotgun proteomics and biomarker discovery. *Dis Markers* **2002**, 18, (2), 99-105.
11. 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.; Sistiare, F. D., Toxicoproteomics: serum proteomic pattern diagnostics for early detection of drug induced cardiac toxicities and cardioprotection. *Toxicol Pathol* **2004**, 32 Suppl 1, 122-30.
12. 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., Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **2002**, 359, (9306), 572-7.
13. Petricoin, E. F., 3rd; Ornstein, D. K.; Paweletz, C. P.; Ardekani, A.; Hackett, P. S.; Hitt, B. A.; Velasco, 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., Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* **2002**, 94, (20), 1576-8.
14. Bartels, C., Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spectrom.* **1990**, 19, 363–368.
 15. Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V., A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *Comput Appl Biosci* **1995**, 11, (4), 427-34.
 16. Taylor, J. A.; Johnson, R. S., Sequence Database Searches via de Novo Peptide Sequencing by Tandem Mass Spectrometry. *Rapid Communications In Mass Spectrometry* **1997**, 11, 1067–1075.
 17. Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A., De Novo Peptide Sequencing via Tandem Mass Spectrometry. *J Comput Biol* **1999**, 6, (3/4), 327–342.
 18. Frank, A.; Pevzner, P., P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **2005**, 77, 964-973.
 19. Estrada, E.; Uriarte, E., Recent advances on the role of topological indices in drug discovery research. *Curr Med Chem* **2001**, 8, 1573-1588.
 20. González-Díaz, H.; Vilar, S.; Santana, L.; Uriarte, E., Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. *Current Topics in Medicinal Chemistry* **2007**, 7, (10), 1025-39.
 21. Randić, M.; Zupan, J.; Vikić-Topić, D., On representation of proteins by star-like graphs. *J Mol Graph Model* **2007**, 290-305.
 22. Zupan, J.; Randić, M., Algorithm for coding DNA sequences into "spectrum-like" and "zigzag" representations. *J Chem Inf Model* **2005**, 45, (2), 309-13.
 23. Randić, M.; Lers, N.; Plavšić, D.; Basak, S.; Balaban, A. T., Four-color map representation of DNA or RNA sequences and their numerical characterization. *Chemical Physics Letters* **2005**, 407, 205-208.
 24. Randić, M.; Balaban, A. T., On a four-dimensional representation of DNA primary sequences. *J Chem Inf Comput Sci* **2003**, 43, (2), 532-9.
 25. Randić, M., A Graph Theoretical Characterization of Proteomics Maps. *Int J Quant Chem* **2002** 90, 848–858.
 26. Bajzer, Z.; Randić, M.; Plavšić, D.; Basak, S. C., Novel map descriptors for characterization of toxic effects in proteomics maps. *J Mol Graph Model* **2003**, 22 1–9.
 27. Liao, B.; Luo, J.; Li, R.; Zhu, W., RNA Secondary structure 2D graphical representation without degeneracy. *International Journal of Quantum Chemistry* **2006** 106 (8), 1749-1755.
 28. Liao, B.; Xiang, X.; Zhu, W., Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *J Comput Chem* **2006**, 27, (11), 1196-1202.
 29. Zhu, W.; Liao, B.; Ding, K., A condensed 3D Graphical representation of RNA secondary structures. *Journal of Molecular Structure: THEOCHEM* **2005** 757 193-198.
 30. Yu-Hua, Y.; Liao, B.; Tian-Ming, W., A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it. *Journal of molecular structure:THEOCHEM* **2005**, 755, 131-136.
 31. Liao, B.; Wang, T.; Ding, K., On A Seven-Dimensional Representation of RNA Secondary Structures. *Molecular Simulation* **2005**, 31, (14), 1063-1071.
 32. Liao, B.; Ding, K.; Wang, T., On A Six-Dimensional Representation of RNA Secondary Structures. *J Biomol Struc Dynamics* **2005**, 22, 455-464.
 33. Randić, M., On graphical and numerical characterization of proteomics maps. *J Chem Inf Comput Sci* **2001**, 41, (5), 1330-8.
 34. Randić, M.; Zupan, J.; Nović, M., On 3-D graphical representation of proteomics maps and their numerical characterization. *J Chem Inf Comput Sci* **2001**, 41, (5), 1339-44.
 35. Randić, M.; Zupan, J.; Nović, M.; Gute, B. D.; Basak, S. C., Novel matrix invariants for characterization of changes of proteomics maps. *SAR QSAR Environ Res* **2002**, 13, (7-8), 689-703.
 36. Randić, M.; Basak, S. C., A comparative study of proteomics maps using graph theoretical biodescriptors. *J Chem Inf Comput Sci* **2002**, 42, (5), 983-92.
 37. Randić, M., Quantitative characterizations of proteome: dependence on the number of proteins considered. *J Proteome Res* **2006**, 5, (7), 1575-9.

38. Randic, M.; Witzmann, F. A.; Kodali, V.; Basak, S. C., On the dependence of a characterization of proteomics maps on the number of protein spots considered. *J Chem Inf Model* **2006**, *46*, (1), 116-22.
39. Randic, M.; Novic, M.; Vracko, M., Novel characterization of proteomics maps by sequential neighborhoods of protein spots. *J Chem Inf Model* **2005**, *45*, (5), 1205-13.
40. Bajzer, Z.; Randic, M.; Plavsic, D.; Basak, S. C., Novel map descriptors for characterization of toxic effects in proteomics maps. *J Mol Graph Model* **2003**, *22*, (1), 1-9.
41. Randic, M.; Novic, M.; Vracko, M., On characterization of dose variations of 2-D proteomics maps by matrix invariants. *J Proteome Res* **2002**, *1*, (3), 217-26.
42. Nandy, A., Novel Method for Discrimination of Conserved Genes through Numerical Characterization of DNA Sequences. *Int E J Mol Design* **2003**, *2*, 000-000.
43. Roy, A.; Raychaudhur, C.; Nandy, A., Novel techniques of graphical representation and analysis of DNA sequences-A review. *J. Biosci.* **1998**, *23*, (1), 55-71.
44. Nandy, A., Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *CABIOS (Comput-Appl-Biosci.)* **1996**, *12*, (1), 55-62.
45. Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y., Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett* **2006**, *580*, (3), 723-30.
46. Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y., Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **2006**, *6*, 4023-4037.
47. Benigni, R.; Andreoli, C.; Giuliani, A., QSAR models for both mutagenic potency and activity: application to nitroarenes and aromatic amines. *Environ Mol Mutagen* **1994**, *24*, (3), 208-19.
48. Caballero, J.; Fernandez, M., Linear and nonlinear modeling of antifungal activity of some heterocyclic ring derivatives using multiple linear regression and Bayesian-regularized neural networks. *J Mol Model (Online)* **2005**, 1-14.
49. Caballero, J.; Garriga, M.; Fernandez, M., Genetic neural network modeling of the selective inhibition of the intermediate-conductance Ca²⁺-activated K⁺ channel by some triarylmethanes using topological charge indexes descriptors. *J Comput Aided Mol Des* **2005**, *19*, (11), 771-89.
50. Cai, C. Z.; Han, L. Y.; Chen, X.; Cao, Z. W.; Chen, Y. Z., Prediction of functional class of the SARS coronavirus proteins by a statistical learning method. *J Proteome Res* **2005**, *4*, (5), 1855-62.
51. Chou, K. C.; Cai, Y. D., Predicting protein quaternary structure by pseudo amino acid composition. *Proteins* **2003**, *53*, (2), 282-9.
52. Chou, K. C.; Cai, Y. D., Prediction of protein subcellular locations by GO-FunD-PseAA predictor. *Biochem Biophys Res Commun* **2004**, *320*, (4), 1236-9.
53. Chou, K. C.; Shen, H. B., Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. *Biochem Biophys Res Commun* **2006**.
54. Fernandez, M.; Caballero, J., Modeling of activity of cyclic urea HIV-1 protease inhibitors using regularized-artificial neural networks. *Bioorg Med Chem* **2006**, *14*, (1), 280-94.
55. Fernandez, M.; Caballero, J.; Tundidor-Camba, A., Linear and nonlinear QSAR study of N-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors. *Bioorg Med Chem* **2006**, *14*, (12), 4137-50.
56. Huang, Y.; Li, Y., Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics* **2004**, *20*, (1), 21-8.
57. Shen, H. B.; Chou, K. C., Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. *Biochem Biophys Res Commun* **2005**, *337*, (3), 752-6.
58. Shen, H.; Chou, K. C., Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. *Biochem Biophys Res Commun* **2005**, *334*, (1), 288-92.
59. Smith, F. M.; Gallagher, W. M.; Fox, E.; Stephens, R. B.; Rexhepaj, E.; Petricoin, E. F., 3rd; Liotta, L.; Kennedy, M. J.; Reynolds, J. V., Combination of SELDI-TOF-MS and data mining provides early-stage response prediction for rectal tumors undergoing multimodal neoadjuvant therapy. *Ann Surg* **2007**, *245*, (2), 259-66.

60. Gruvberger-Saal, S. K.; Eden, P.; Ringner, M.; Baldetorp, B.; Chebil, G.; Borg, A.; Ferno, M.; Peterson, C.; Meltzer, P. S., Predicting continuous values of prognostic markers in breast cancer from microarray gene expression profiles. *Mol Cancer Ther* **2004**, 3, (2), 161-8.
61. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E., A Model for the Recognition of Protein Kinases Based on the Entropy of 3D van der Waals Interactions. *J Proteome Res* **2007**, 6, (2), 904-908.
62. Heijne, W. H.; Stierum, R. H.; Slijper, M.; van Bladeren, P. J.; van Ommen, B., Toxicogenomics of bromobenzene hepatotoxicity: a combined transcriptomics and proteomics approach. *Biochem Pharmacol* **2003**, 65, (5), 857-75.
63. Marrero-Ponce, Y., Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J Chem Inf Comput Sci* **2004**, 44, (6), 2010-26.
64. Komura, D.; Nakamura, H.; Tsutsumi, S.; Aburatani, H.; Ihara, S., Multidimensional support vector machines for visualization of gene expression data. *Bioinformatics* **2005**, 21, (4), 439-44.
65. Lavine, B. K.; Davidson, C. E.; Rayens, W. S., Machine learning based pattern recognition applied to microarray data. *Comb Chem High Throughput Screen* **2004**, 7, (2), 115-31.
66. Bathen, T. F.; Krane, J.; Engan, T.; Bjerve, K. S.; Axelson, D., Quantification of plasma lipids and apolipoproteins by use of proton NMR spectroscopy, multivariate and neural network analysis. *NMR Biomed* **2000**, 13, (5), 271-88.
67. Zhao, C. Y.; Zhang, H. X.; Zhang, X. Y.; Liu, M. C.; Hu, Z. D.; Fan, B. T., Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* **2005**.
68. Li, H.; Ung, C. Y.; Yap, C. W.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Chen, Y. Z., Prediction of genotoxicity of chemical compounds by statistical learning methods. *Chem Res Toxicol* **2005**, 18, (6), 1071-80.
69. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T., QSAR models for the prediction of binding affinities to human serum albumin using the heuristic method and a support vector machine. *J Chem Inf Comput Sci* **2004**, 44, (5), 1693-700.
70. Cui, J.; Han, L. Y.; Cai, C. Z.; Zheng, C. J.; Ji, Z. L.; Chen, Y. Z., Prediction of functional class of novel bacterial proteins without the use of sequence similarity by a statistical learning method. *J Mol Microbiol Biotechnol* **2005**, 9, (2), 86-100.
71. Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z., SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* **2003**, 31, (13), 3692-7.
72. Cai, Y. D.; Lin, S. L., Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim Biophys Acta* **2003**, 1648, (1-2), 127-33.
73. Chou, K. C.; Cai, Y. D., Using functional domain composition and support vector machines for prediction of protein subcellular location. *J Biol Chem* **2002**, 277, (48), 45765-9.
74. Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C., Support vector machines for predicting the specificity of GalNAc-transferase. *Peptides* **2002**, 23, (1), 205-8.
75. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U., Complex networks: Structure and dynamics. *Physics Reports* **2006**, 424, 175-308.
76. Barabasi, A. L., Sociology. Network theory-the emergence of the creative enterprise. *Science* **2005**, 308, (5722), 639-41.
77. Barabasi, A. L.; Oltvai, Z. N., Network biology: understanding the cell's functional organization. *Nat Rev Genet* **2004**, 5, (2), 101-13.
78. Blais, A.; Dynlacht, B. D., Constructing transcriptional regulatory networks. *Genes Dev* **2005**, 19, (13), 1499-511.
79. Brazhnik, P.; de la Fuente, A.; Mendes, P., Gene networks: how to put the function in genomics. *Trends Biotechnol* **2002**, 20, (11), 467-72.
80. De, P.; Singh, A. E.; Wong, T.; Yacoub, W.; Jolly, A. M., Sexual network analysis of a gonorrhoea outbreak. *Sex Transm Infect* **2004**, 80, (4), 280-5.
81. Dezsó, Z.; Barabasi, A. L., Halting viruses in scale-free networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **2002**, 65, (5 Pt 2), 055103.

82. Dobrin, R.; Beg, Q. K.; Barabasi, A. L.; Oltvai, Z. N., Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics* **2004**, 5, 10.
83. Estrada, E., Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* **2006**, 6, (1), 35-40.
84. Estrada, E., Protein bipartivity and essentiality in the yeast protein-protein interaction network. *J Proteome Res* **2006**, 5, (9), 2177-84.
85. Estrada, E.; Rodriguez-Velazquez, J. A., Subgraph centrality in complex networks. *Phys Rev E Stat Nonlin Soft Matter Phys* **2005**, 71, (5 Pt 2), 056103.
86. Gomez, S. M.; Lo, S. H.; Rzhetsky, A., Probabilistic prediction of unknown metabolic and signal-transduction networks. *Genetics* **2001**, 159, (3), 1291-8.
87. Guimera, R.; Mossa, S.; Turtschi, A.; Amaral, L. A., The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc Natl Acad Sci U S A* **2005**, 102, (22), 7794-9.
88. Honey, C. J.; Kotter, R.; Breakspear, M.; Sporns, O., Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc Natl Acad Sci U S A* **2007**, 104, (24), 10240-5.
89. Huynen, M. A.; Snel, B.; von Mering, C.; Bork, P., Function prediction and protein networks. *Curr Opin Cell Biol* **2003**, 15, (2), 191-8.
90. Jansen, R.; Gerstein, M., Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol* **2004**, 7, (5), 535-45.
91. Kel, A.; Reymann, S.; Matys, V.; Nettessheim, P.; Wingender, E.; Borlak, J., A novel computational approach for the prediction of networked transcription factors of aryl hydrocarbon-receptor-regulated genes. *Mol Pharmacol* **2004**, 66, (6), 1557-72.
92. Klovdahl, A. S.; Graviss, E. A.; Yaganehdoost, A.; Ross, M. W.; Wanger, A.; Adams, G. J.; Musser, J. M., Networks and tuberculosis: an undetected community outbreak involving public places. *Soc Sci Med* **2001**, 52, (5), 681-94.
93. LaCount, D. J.; Vignali, M.; Chettier, R.; Phansalkar, A.; Bell, R.; Hesselberth, J. R.; Schoenfeld, L. W.; Ota, I.; Sahasrabudhe, S.; Kurschner, C.; Fields, S.; Hughes, R. E., A protein interaction network of the malaria parasite Plasmodium falciparum. *Nature* **2005**, 438, (3), 103-107.
94. Mason, O.; Verwoerd, M., Graph theory and networks in Biology. *IET Syst Biol* **2007**, 1, (2), 89-119.
95. Zhu, T.; Phalakornkule, C.; Ghosh, S.; Grossmann, I. E.; Koepsel, R. R.; Ataai, M. M.; Domach, M. M., A metabolic network analysis & NMR experiment design tool with user interface-driven model construction for depth-first search analysis. *Metab Eng* **2003**, 5, (2), 74-85.
96. Zhou, X. J.; Kao, M. C.; Huang, H.; Wong, A.; Nunez-Iglesias, J.; Primig, M.; Aparicio, O. M.; Finch, C. E.; Morgan, T. E.; Wong, W. H., Functional annotation and network reconstruction through cross-platform integration of microarray data. *Nat Biotechnol* **2005**, 23, (2), 238-43.
97. Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H., Predicting protein-protein interactions based only on sequences information. *PNAS* **2007**, 104, (11), 4337-4341.
98. Stam, C. J.; Jones, B. F.; Nolte, G.; Breakspear, M.; Scheltens, P., Small-world networks and functional connectivity in Alzheimer's disease. *Cereb Cortex* **2007**, 17, (1), 92-9.
99. Williams, R. J.; Berlow, E. L.; Dunne, J. A.; Barabasi, A. L.; Martinez, N. D., Two degrees of separation in complex food webs. *Proc Natl Acad Sci U S A* **2002**, 99, (20), 12913-6.
100. Todeschini, R.; Consonni, V.; Mauri, A.; Ballabio, D., Characterization of DNA Primary Sequences by a New Similarity/Diversity Measure Based on the Partial Ordering *J Chem Inf Model* **2006**, 46, (5), 1905-1911.
101. Ye, Y.; Godzik, A., Multiple flexible structure alignment using partial order graphs. *Bioinformatics* **2005**, 21, (10), 2362-9.
102. Grasso, C.; Modrek, B.; Xing, Y.; Lee, C., Genome-wide detection of alternative splicing in expressed sequences using partial order multiple sequence alignment graphs. *Pac Symp Biocomput* **2004**, 29-41.
103. Lee, C., Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics* **2003**, 19, (8), 999-1008.
104. Grasso, C.; Quist, M.; Ke, K.; Lee, C., POAVIZ: a Partial order multiple sequence alignment visualizer. *Bioinformatics* **2003**, 19, (11), 1446-8.

105. Lee, C.; Grasso, C.; Sharlow, M. F., Multiple sequence alignment using partial order graphs. *Bioinformatics* **2002**, 18, (3), 452-64.
106. Randic, M.; Lers, N.; Vukicevic, D.; Plavsic, D.; Gute, B. D.; Basak, S. C., Canonical labeling of proteome maps. *J Proteome Res* **2005**, 4, (4), 1347-52.
107. Bruggemann, R.; Pudenz, S.; Carlsen, L.; Sorensen, P. B.; Thomsen, M.; Mishra, R. K., The use of Hasse diagrams as a potential approach for inverse QSAR. *SAR QSAR Environ Res* **2001**, 11, (5-6), 473-87.
108. Bruggemann, R.; Halfon, E.; Welzl, G.; Voigt, K.; Steinberg, C. E., Applying the concept of partially ordered sets on the ranking of near-shore sediments by a battery of tests. *J Chem Inf Comput Sci* **2001**, 41, (4), 918-25.
109. Lerche, D.; Bruggemann, R.; Sorensen, P.; Carlsen, L.; Nielsen, O. J., A comparison of partial order technique with three methods of multi-criteria analysis for ranking of chemical substances. *J Chem Inf Comput Sci* **2002**, 42, (5), 1086-98.
110. Hollert, H.; Heise, S.; Pudenz, S.; Bruggemann, R.; Ahlf, W.; Braunbeck, T., Application of a sediment quality triad and different statistical approaches (Hasse diagrams and fuzzy logic) for the comparative evaluation of small streams. *Ecotoxicology* **2002**, 11, (5), 311-21.
111. Lerche, D.; Sorensen, P. B.; Bruggemann, R., Improved estimation of the ranking probabilities in partial orders using random linear extensions by approximation of the mutual ranking probability. *J Chem Inf Comput Sci* **2003**, 43, (5), 1471-80.
112. Jensen, T. S.; Lerche, D. B.; Sorensen, P. B., Ranking contaminated sites using a partial ordering method. *Environ Toxicol Chem* **2003**, 22, (4), 776-83.
113. Bruggemann, R.; Welzl, G.; Voigt, K., Order theoretical tools for the evaluation of complex regional pollution patterns. *J Chem Inf Comput Sci* **2003**, 43, (6), 1771-9.
114. Voigt, K.; Bruggemann, R.; Pudenz, S., Chemical databases evaluated by order theoretical tools. *Anal Bioanal Chem* **2004**, 380, (3), 467-74.
115. Simon, U.; Bruggemann, R.; Pudenz, S., Aspects of decision support in water management--example Berlin and Potsdam (Germany) I--spatially differentiated evaluation. *Water Res* **2004**, 38, (7), 1809-16.
116. Simon, U.; Bruggemann, R.; Pudenz, S., Aspects of decision support in water management--example Berlin and Potsdam (Germany) II--improvement of management strategies. *Water Res* **2004**, 38, (19), 4085-92.
117. Sanchez, R.; Grau, R.; Morgado, E., *MATCH Commun Math Comput Chem* **2004**, 52, 29-46.
118. Lambertenghi-Deliliers, G.; Zanon, P. L.; Pozzoli, E. F.; Bellini, O., Myocardial injury induced by a single dose of adriamycin: an electron microscopic study. *Tumori* **1976**, 62, (5), 517-28.
119. Zhang, J.; Herman, E. H.; Ferrans, V. J., Dendritic cells in the hearts of spontaneously hypertensive rats treated with doxorubicin with or without ICRF-187. *Am J Pathol* **1993**, 142, (6), 1916-26.
120. Herman, E. H.; Zhang, J.; Rifai, N.; Lipshultz, S. E.; Hasinoff, B. B.; Chadwick, D. P.; Knapton, A.; Chai, J.; Ferrans, V. J., The use of serum levels of cardiac troponin T to compare the protective activity of dexrazoxane against doxorubicin- and mitoxantrone-induced cardiotoxicity. *Cancer Chemother Pharmacol* **2001**, 48, (4), 297-304.
121. Zhang, J.; Herman, E. H.; Knapton, A.; Chadwick, D. P.; Whitehurst, V. E.; Koerner, J. E.; Papoian, T.; Ferrans, V. J.; Sistare, F. D., SK&F 95654-induced acute cardiovascular toxicity in Sprague-Dawley rats--histopathologic, electron microscopic, and immunohistochemical studies. *Toxicol Pathol* **2002**, 30, (1), 28-40.
122. Randic, M., Graphical representations of DNA as 2-D map. *Chemical Physics Letters* **2004**, 386, (4), 468-471.
123. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Gonzalez-Diaz, Y.; Sanchez-Gonzalez, A., Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *J Comput Chem* **2007**, 28, (6), 1042-1048.
124. González-Díaz, H.; Pérez-Castillo, Y.; Podda, G.; Uriarte, E., Computational Chemistry Comparison of Stable/Nonstable Protein Mutants Classification Models Based on 3D and Topological Indices. *J Comput Chem* **2007**, 28, 1990-1995.

125. González-Díaz, H.; Pérez-Bello, A.; Cruz-Montegudo, M.; González-Díaz, Y.; Santana, L.; Uriarte, E., Chemometrics for QSAR with low sequence homology: Mycobacterial promoter sequences recognition with 2D-RNA entropies. *Chemom Intell Lab Syst* **2007**, 85, 20-26.
126. González-Díaz, H.; Agüero-Chapin, G.; Varona, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Gianni, P., 2D-RNA-Coupling Numbers: A New Computational Chemistry Approach to Link Secondary Structure Topology with Biological Function. *J Comput Chem* **2007**, 28, 1049–1056.
127. Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E., A QSAR model for in silico screening of MAO-A inhibitors. Prediction, synthesis, and biological assay of novel coumarins. *J Med Chem* **2006**, 49, (3), 1149-56.
128. González-Díaz, H.; Vina, D.; Santana, L.; de Clercq, E.; Uriarte, E., Stochastic entropy QSAR for the in silico discovery of anticancer compounds: prediction, synthesis, and in vitro assay of new purine carbanucleosides. *Bioorg Med Chem* **2006**, 14, (4), 1095-107.
129. González-Díaz, H.; Pérez-Bello, A.; Uriarte, E.; Gonzalez-Diaz, Y., QSAR study for mycobacterial promoters with low sequence homology. *Bioorg Med Chem Lett* **2006**, 16, (3), 547-53.
130. González-Díaz, H.; Saiz-Urra, L.; Molina, R.; Uriarte, E., Stochastic molecular descriptors for polymers. 2. Spherical truncation of electrostatic interactions on entropy based polymers 3D-QSAR. *Polymer* **2005**, 46, 2791–2798.
131. González-Díaz, H.; Pérez-Bello, A.; Uriarte, E., Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. *Polymer* **2005**, 46 6461–6473.
132. Maruyama, T., A Markov process of gene frequency change in a geographically structured population. *Genetics* **1974**, 76, (2), 367-77.
133. González-Díaz, H.; Molina-Ruiz, R.; Hernandez, I. *MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design)*, 2.0; 2005.
134. González-Díaz, H. H., I., BIOMARKS version 1.0, contact information: gonzalezdiazh@yahoo.es or qohumbe@usc.es **2005**.
135. González-Díaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Diaz, Y., 3D-QSAR study for DNA cleavage proteins with a potential anti-tumor ATCUN-like motif. *J Inorg Biochem* **2006**, 100, (7), 1290-7.
136. Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y., Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett* **2006**, 580 723-730.
137. Saiz-Urra, L.; González-Díaz, H.; Uriarte, E., Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials. *Bioorg Med Chem* **2005**, 13, (11), 3641-7.
138. González-Díaz, H.; Uriarte, E.; Ramos de Armas, R., Predicting stability of Arc repressor mutants with protein stochastic moments. *Bioorg Med Chem* **2005**, 13, (2), 323-31.
139. González-Díaz, H.; Uriarte, E., Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. *Biopolymers* **2005**, 77, (5), 296-303.
140. González-Díaz, H.; Uriarte, E., Proteins QSAR with Markov average electrostatic potentials. *Bioorg Med Chem Lett* **2005**, 15, (22), 5088-94.
141. Gonzalez-Diaz, H.; Molina, R.; Uriarte, E., Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. *FEBS Lett* **2005**, 579, (20), 4297-301.
142. Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E., Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. *Proteins* **2004**, 56, (4), 715-23.
143. González-Díaz, H.; Molina, R.; Uriarte, E., Markov entropy backbone electrostatic descriptors for predicting proteins biological activity. *Bioorg Med Chem Lett* **2004**, 14, (18), 4691-5.
144. Van Waterbeemd, H., Discriminant Analysis for Activity Prediction. In *Chemometric methods in molecular design*, Van Waterbeemd, H., Ed. Wiley-VCH: New York, 1995; Vol. 2, pp 265-282.
145. *STATISTICA*, 6.0 for Windows; Statsoft Inc.: 2001.
146. Kowalski, R. B.; Wold, S., Pattern recognition in chemistry. In *Handbook of statistics*, Krishnaiah, P. R.; Kanal, L. N., Eds. North Holland Publishing Company: Amsterdam, 1982; pp 673-697.
147. Van Waterbeemd, H., *Chemometric methods in molecular design*. Wiley-VCH: New York, 1995; Vol. 2.

148. Cruz-Monteagudo, M.; González-Díaz, H.; Agüero-Chapin, G.; Santana, L.; Borges, F.; Domínguez, R. E.; Podda, G.; Uriarte, E., Computational Chemistry Development of a Unified Free Energy Markov Model for the Distribution of 1300 Chemicals to 38 Different Environmental or Biological Systems. *J Comput Chem* **2007**, *28*, 1909-1922.
149. Stewart, J.; Gill, L., *Econometrics*. 2nd edition ed.; Prentice Hall: London, 1998.
150. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W., Standardized Multiple Regression Model. In *Applied Linear Statistical Models*, Fifth ed.; McGraw Hill: New York, 2005; pp 271-277.
151. Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E., Jr., A general treatment of solubility. 3. Principal component analysis (PCA) of the solubilities of diverse solutes in diverse solvents. *J Chem Inf Model* **2005**, *45*, (4), 913-23.
152. van de Waterbeemd, H.; el Tayar, N.; Carrupt, P. A.; Testa, B., Pattern recognition study of QSAR substituent descriptors. *J Comput Aided Mol Des* **1989**, *3*, (2), 111-32.
153. StatSoft.Inc. *STATISTICA (data analysis software system), version 6.0*, www.statsoft.com.Statsoft, Inc., 6.0; 2002.
154. Microsoft.Corp. *Microsoft Excel* 2002.
155. Zhang, W., Computer inference of network of ecological interactions from sampling data. *Environ Monit Assess* **2007**, *124*, (1-3), 253-61.
156. Koschützki, D. *CentiBiN Version 1.4.2*, 2006.
157. Junker, B. H.; Koschutzki, D.; Schreiber, F., Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **2006**, *7*, 219.
158. Batagelj, V.; Mrvar, A., Pajek 1.15. **2006**.
159. Ludemann, A.; Weicht, D.; Selbig, J.; Kopka, J., PaVESy: Pathway Visualization and Editing System. *Bioinformatics* **2004**, *20*, (16), 2841-4.
160. Pudenz, S.; Bruggemann, R.; Bartel, H. G., QSAR of ecotoxicological data on the basis of data-driven if-then-rules. *Ecotoxicology* **2002**, *11*, (5), 337-42.
161. Carlsen, L.; Sorensen, P. B.; Thomsen, M.; Bruggemann, R., QSAR's based on partial order ranking. *SAR QSAR Environ Res* **2002**, *13*, (1), 153-65.
162. Geyer, H. J.; Scheunert, I.; Bruggemann, R.; Steinberg, C.; Korte, F.; Kettrup, A., QSAR for organic chemical bioconcentration in Daphnia, algae, and mussels. *Sci Total Environ* **1991**, 109-110, 387-94.
163. Bruggemann, R.; Sorensen, P. B.; Lerche, D.; Carlsen, L., Estimation of averaged ranks by a local partial order model. *J Chem Inf Comput Sci* **2004**, *44*, (2), 618-25.
164. Sorensen, P. B.; Bruggemann, R.; Carlsen, L.; Mogensen, B. B.; Kreuger, J.; Pudenz, S., Analysis of monitoring data of pesticide residues in surface waters using partial order ranking theory. *Environ Toxicol Chem* **2003**, *22*, (3), 661-70.
165. Marengo, E.; Leardi, R.; Robotti, E.; Righetti, P. G.; Antonucci, F.; Cecconi, D., Application of Three-Way Principal Component Analysis to the Evaluation of Two-Dimensional Maps in Proteomics. *J Proteome Res* **2003**, *2*, 351-360.
166. Fallico, V.; McSweeney, P. L.; Siebert, K. J.; Horne, J.; Carpino, S.; Licitra, G., Chemometric analysis of proteolysis during ripening of Ragusano cheese. *J Dairy Sci* **2004**, *87*, (10), 3138-52.
167. Bender, A.; van Dooren, G. G.; Ralph, S. A.; McFadden, G. I.; Schneider, G., Properties and prediction of mitochondrial transit peptides from Plasmodium falciparum. *Mol Biochem Parasitol* **2003**, *132*, (2), 59-66.
168. Lilien, R. H.; Farid, H.; Donald, B. R., Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J Comput Biol* **2003**, *10*, (6), 925-46.
169. Nandy, A., Two-dimensional graphical representation of DNA sequences and intron-exon discrimination in intron-rich sequences. *Comput Appl Biosci* **1996**, *12*, (1), 55-62.
170. Nandy, A.; Basak, S. C., Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J Chem Inf Comput Sci* **2000**, *40*, (4), 915-9.
171. de Julián-Ortiz, J. V.; de Gregorio Alapont, C.; Ríos-Santamarina, I.; García-Doménech, R.; Gálvez, J., Prediction of properties of chiral compounds by molecular topology *J Mol Graph Model* **1998**, *16*, (1), 14-18

172. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borrás, R., New agents active against Mycobacterium avium complex selected by molecular topology: a virtual screening method. *J Antimicrob Chemother* **2004**, 53, (1), 65-73.
173. Zweig, M. H., Apolipoproteins and lipids in coronary artery disease. Analysis of diagnostic accuracy using receiver operating characteristic plots and areas. *Arch Pathol Lab Med* **1994**, 118, (2), 141-4.
174. Zweig, M. H.; Broste, S. K.; Reinhart, R. A., ROC curve analysis: an example showing the relationships among serum lipid and apolipoprotein concentrations in identifying patients with coronary artery disease. *Clin Chem* **1992**, 38, (8 Pt 1), 1425-8.
175. Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W., Multicollinearity and its effects. In *Applied Linear Statistical Models*, Fifth ed.; McGraw Hill: New York, 2005; pp 278-289.
176. Dunn, W. J., 3rd; Koehler, M. G.; Grigoras, S., The role of solvent-accessible surface area in determining partition coefficients. *J Med Chem* **1987**, 30, (7), 1121-6.
177. Hellberg, S.; Sjoström, M.; Skagerberg, B.; Wold, S., Peptide quantitative structure-activity relationships, a multivariate approach. *J Med Chem* **1987**, 30, (7), 1126-35.
178. Mei, H.; Liao, Z. H.; Zhou, Y.; Li, S. Z., A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers* **2005**, 80, (6), 775-86.
179. Mikula, S.; Niebur, E., A novel method for visualizing functional connectivity using principal component analysis. *Int J Neurosci* **2006**, 116, (4), 419-29.
180. Benigni, R.; Cotta-Ramusino, M.; Gallo, G.; Giorgi, F.; Giuliani, A.; Vari, M. R., Deriving a quantitative chirality measure from molecular similarity indices. *J Med Chem* **2000**, 43, (20), 3699-703.
181. Lejon, T.; Strom, M. B.; Svendsen, J. S., Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. *J Pept Sci* **2001**, 7, (2), 74-81.
182. Yu, H.; Kim, P. M.; Sprecher, E.; Trifonov, V.; Gerstein, M., The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* **2007**, 3, (4), e59.
183. Sun, F. T.; Miller, L. M.; Rao, A. A.; D'Esposito, M., Functional connectivity of cortical networks involved in bimanual motor sequence learning. *Cereb Cortex* **2007**, 17, (5), 1227-34.
184. Nemunaitis, J.; Senzer, N.; Khalil, I.; Shen, Y.; Kumar, P.; Tong, A.; Kuhn, J.; Lamont, J.; Nemunaitis, M.; Rao, D.; Zhang, Y. A.; Zhou, Y.; Vorhies, J.; Maples, P.; Hill, C.; Shanahan, D., Proof concept for clinical justification of network mapping for personalized cancer therapeutics. *Cancer Gene Ther* **2007**.
185. McDonald, D. B., Predicting fate from early connectivity in a social network. *Proc Natl Acad Sci U S A* **2007**, 104, (26), 10910-4.
186. Ma, H. W.; Zeng, A. P., The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **2003**, 19, (11), 1423-30.
187. Junker, B. H.; Koschuetzki, D.; Schreiber, F., Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **2006**, 7, (1), 219.
188. Farkas, I. J.; Derenyi, I.; Barabasi, A. L.; Vicsek, T., Spectra of "real-world" graphs: beyond the semicircle law. *Phys Rev E Stat Nonlin Soft Matter Phys* **2001**, 64, (2 Pt 2), 026704.
189. Zhou, H., Distance, dissimilarity index, and network community structure. *Phys Rev E Stat Nonlin Soft Matter Phys* **2003**, 67, (6 Pt 1), 061901.
190. Vazquez, A., Spreading dynamics on small-world networks with connectivity fluctuations and correlations. *Phys Rev E Stat Nonlin Soft Matter Phys* **2006**, 74, (5 Pt 2), 056101.
191. Louzoun, Y.; Muchnik, L.; Solomon, S., Copying nodes versus editing links: the source of the difference between genetic regulatory networks and the WWW. *Bioinformatics* **2006**, 22, (5), 581-8.
192. Grasso, C.; Lee, C., Combining partial order alignment and progressive multiple sequence alignment increases alignment speed and scalability to very large alignment problems. *Bioinformatics* **2004**, 20, (10), 1546-56.
193. Pfleiderer, C.; Reznik, D.; Pintschovius, L.; Lohneisen, H. V.; Garst, M.; Rosch, A., Partial order in the non-Fermi-liquid phase of MnSi. *Nature* **2004**, 427, (6971), 227-31.
194. Carlsen, L., A combined QSAR and partial order ranking approach to risk assessment. *SAR QSAR Environ Res* **2006**, 17, (2), 133-46.

195. Lerche, D.; Matsuzaki, S. Y.; Sorensen, P. B.; Carlsen, L.; Nielsen, O. J., Ranking of chemical substances based on the Japanese Pollutant Release and Transfer Register using partial order theory and random linear extensions. *Chemosphere* **2004**, 55, (7), 1005-25.
196. Sandford, D.; Fendorf, M.; Stacy, A. M.; Holstein, W. L.; Crawford, M. K., Observation of Hendricks-Teller partial order in a tetragonal cuprate superconductor: La_{1.68}Nd_{0.14}Na_{0.10}K_{0.082}CuO₄. *Phys Rev B Condens Matter* **1994**, 50, (13), 9419-9425.
197. Sorensen, P. B.; Mogensen, B. B.; Carlsen, L.; Thomsen, M., The influence on partial order ranking from input parameter uncertainty. Definition of a robustness parameter. *Chemosphere* **2000**, 41, (4), 595-601.
198. Lerche, D.; Sorensen, P. B.; Larsen, H. S.; Carlsen, L.; Nielsen, O. J., Comparison of the combined monitoring-based and modelling-based priority setting scheme with partial order theory and random linear extensions for ranking of chemical substances. *Chemosphere* **2002**, 49, (6), 637-49.