Currents Bioinformatics. 2015; 10 (5): 606-631

Graph-based processing of macromolecular information

Cristian R. Munteanu [1,2], Vanessa Aguiar-Pulido[1], Ana Freire[3], Marcos Martínez-Romero [4], Ana B. Porto-Pazos [1], Javier Pereira[4] and Julián Dorado[1]

[1]*Artificial Neural Networks and Adaptive Systems Lab, Computer Science Faculty, University of A Coruna, Spain*
[2] *Department of Bioinformatics - BiGCaT, Maastricht University, P.o. Box* 616, *UNS50 Box* 19, *NL-6200 MD, Maastricht, The Netherlands*
[3] *Web Research Group, DTIC, Universitat Pompeu Fabra, Barcelona, Spain*
[4] *IMEDIR Center, University of A Coruña, Spain*

Abstract
The complex information encoded into the element connectivity of a system gives rise to the possibility of graphical processing of divisible systems by using the Graph theory. An application in this sense is the quantitative characterization of molecule topologies of drugs, proteins and nucleic acids, in order to build mathematical models as Quantitative Structure - Activity Relationships between the molecules and a specific biological activity. These types of models can predict new drugs, molecular targets and molecular properties of new molecular structures with an important impact on the Drug Discovery, Medicinal Chemistry, Molecular Diagnosis, and Treatment. The current review is focused on the mathematical methods to encode the connectivity information in three types of graphs such as star graphs, spiral graphs and contact networks and three in-house scientific applications dedicated to the calculation of molecular graph topological indices such as S2SNet, CULSPIN and MInD-Prot. In addition, some examples are presented, such as results of this methodology on drugs, proteins and nucleic acids, including the Web implementation of the best molecular prediction models based on graphs.

Keywords: Molecular information, QSAR, Markov descriptors, Graphs, Complex networks, Protein topological indices.

1. INTRODUCTION

A molecule represents the basic item in the complex biological processes and their properties have biochemical consequences, translated in physiological or pathological cellular activities. Due to continuous population growth and aging, elevated costs and time requirements for bringing a new drug on the market, and the increased resistance and adaptability of the pathogens, there is an acute need for innovative medicine based on detailed and accurate molecular information. From over 20,000 non-redundant human proteins there are only few hundred targets for approved drugs. With all important discoveries in omics and screening methods, the number of the new drugs approved by USA Food and Drug Administration (FDA) is reduced, around 25 compounds [1].

The main problems regarding the discovery of effective and precise drugs are the adaptability of the molecular targets and the complex interactions between one drug and multiple targets and between the consequent interactions of the produced metabolites. In order to numerically characterize a molecule, the scientists are using the known physical, chemical, biological properties (experimental or theoretical/calculated). If it is known a series of numbers that characterizes a specific drug or its molecular target and the relation between the drug, target and/or a physiological or pathological condition, it is possible to build mathematical models that can predict drug biological properties or future molecular targets for treatments of specific medical conditions. If using these numbers we cannot obtain a model that predicts the molecular property of interest, there is a need for other methods which can extract information from molecules, generate a new series of invariant numbers for a molecule, and encode this hidden molecular information. The current review presents one of these methods based on connectivity-derived information.

The focus of the review will be on the molecular c1assification problem that allows predicting new specific biological function of molecules (Fig. 1). There are two ways to process the molecular information:

l. Direct processing: the numerical attributes that characterize a molecule (e.g. physical properties) are available and they can be used to obtain a c1assification model that can discriminate between a group of molecules that has a specific property and another group without this property; e.g. the chemical groups from a molecule are used to discriminate between active and non-active drugs;

2. Indirect processing: the available numerical attributes are not available or they are not enough to obtain c1assification models; e.g. the molecular property cannot be directly linked to specific chemical groups in a molecule. In general, this processing is used for macromolecules with complex chemical structures. In this case, there is a need for a method that generates numerical invariants specific for a specific molecule. A possibility is using the Complex Network / Graph theory [2] that considers the molecule as a system with a specific structure by dividing the molecule in elements, connected by specific relationships. This method represents the graphical processing of information from the internal structure or topology of any system. Thus, drugs, proteins or nuc1eic acids can be considered systems. A drug can be divided into atoms connected by chemical bonds. A protein can be divided into amino acids that are connected in a sequence. The elements that define the nuc1eic acid are the nucleotides or the corresponding nuc1eobases (A, T, G, C), connected by phosphate or hydrogen bonds.

Therefore, the indirect method based on graphical representation of the information represents a solution for c1assifications problems that cannot be directly solved. This is the case of the macromolecules with a complex 3D structure, where it is impossible to know *a priori* which specific topology and physical-chemical property of each amino acid (as node) can be used to c1assify the macromolecules for a specific property. An example in this sense is the case of protein c1assification as cancer-related or not to a complex disease such as cancer: it is impossible to say a *priori* which parts / topologies of the proteins are involved in cancer development and how these parts can numerically contribute to this specific property. In conc1usion, these topological indices offer the possibility to encode indirect molecular information using specific formulas.

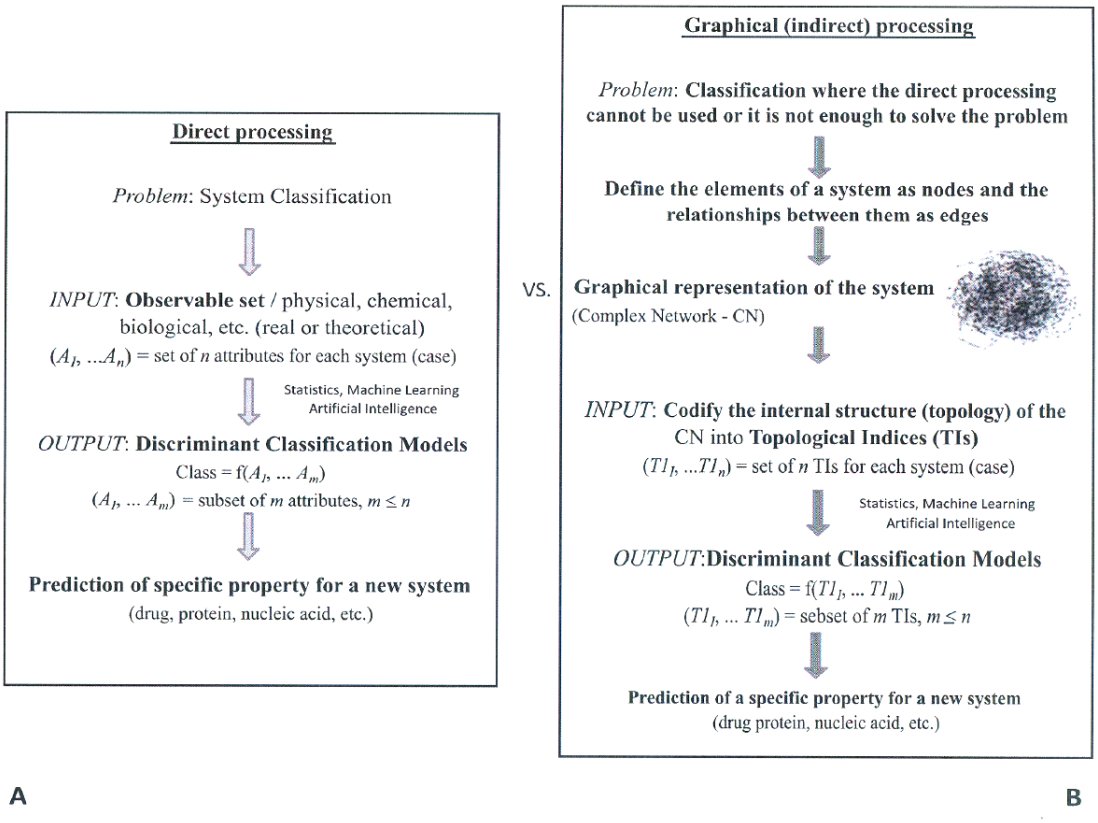


Fig. (1). Direct (A) and (8) graphical (indirect) processing ofinformation.

The molecular information used for the graphical processing with graphs depends on the type of molecule (Fig.2):

– For small/medium molecules (drugs): chemical formula - atom types, chemical bonds, atom physical- chemical properties;

– For macromolecules:

  - Proteins: primary amino acid sequence - amino acid type, frequency and location in the sequence, protein secondary structure -hydrogen bonds, 3D structure - amino acid type and spatial contact between the amino acids;

  - Nucleic acids: primary structures - types of nuc1eobases, frequency and positions in the sequence, secondary structures -hydrogen bonds between nucleobases and 3D structures - type and spatial position of nucleobases.

The chemical structures are represented as a SMILES formula (Simplified Molecular-Input Line-Entry System) [3] for inputs in the graphical processing. The open-standards version of the SMILES language for chemistry is maintained by the OpenSMILES community (http://www.opensmil es.org/), as part of the Blue Obelisk project [4]. The primary sequence information can be found in the FAST format for proteins and nucleic acids. The secondary structure and the 3D coordinate information are included in the PDB files for proteins and nucleic acids.

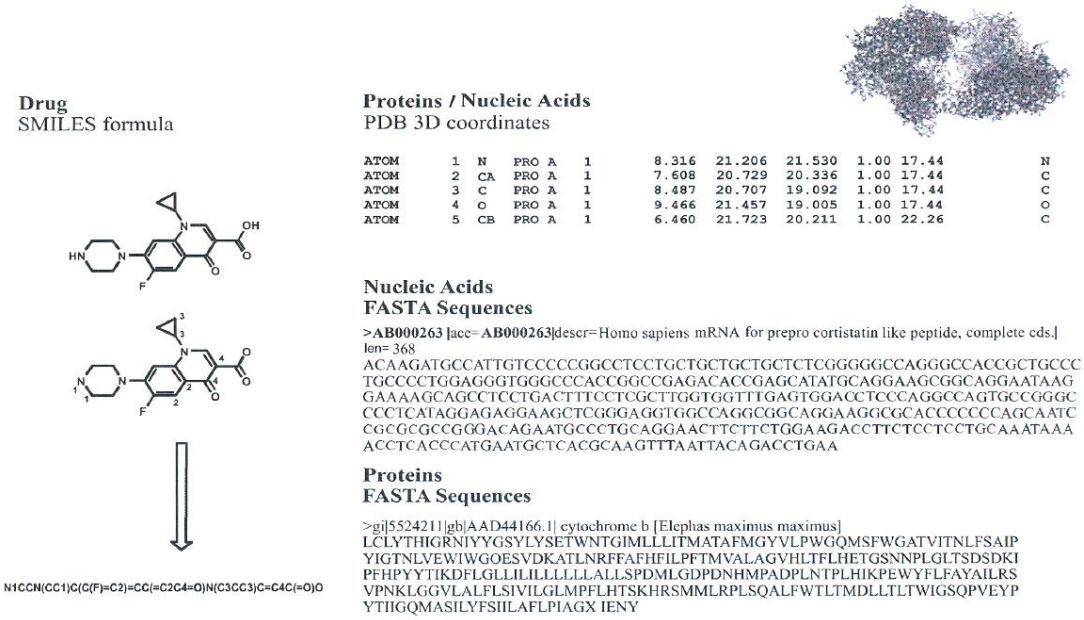


Fig. (2). Molecular information used for graphical processing (SMILES formula from https://es.wikipedia.org/wiki/SMILES).

2. GRAPHS AND TOPOLOGICAL INDICES

2.1. Graph connectivity

Connectivity represents the basic concept of the Graph or Complex Network theory [2], a branch of applied mathematics, which is used in almost all scientific fields, especially in Chemistry [5, 6], Biology [7], or Sociology [8, 9]. Therefore, the real or conceptual connected networks can be graphically represented as a graph: a collection of items or nodes (vertices) with the corresponding connections between them or links (edges, ares). A graph (G) is represented as an ordered pair of *V* as a set of vertices (nodes) and *E* as a set of edges (connections), which are 2-element subsets. Graphically, a graph is plotted as a diagram with dots for vertices and lines for edges. The edges can be undirected, edge (i,j) being identical to edge (j,i) or directed (digraph) (see Fig. 3). In practice, the terms graph and network are used as synonyms, but networks generally refer to real complex systems.

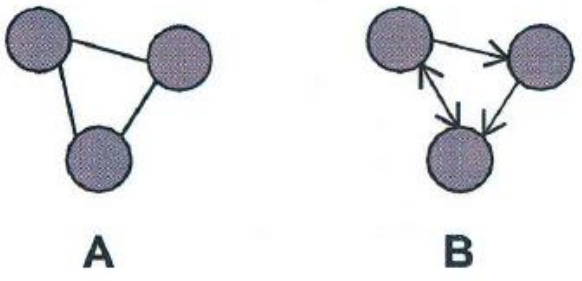


Fig. (3). Types of graphs: (A) undirected and (B) directed

The nodes of a graph can be represented as any type of real or theoretical items such as particles [10], atoms [11], amino acids [12], drugs, proteins [13], nucleic acids [14], parasites [15], diseases [16], amplitude regions of blood mass spectra [17] or other graphs. The corresponding edges can represent any relation between the nodes, from particle interactions or chemical bonds, to complex social interactions.

The number of nodes and edges is defining the complexity of the graphs/networks. For example: with 5 nodes it is possible to build 34 possible graphs with different topology (structure), la nodes can be linked as 12,005,168 different graphs, 20 no des can generate a 39-digit number of graphs and 25 nodes can be drawn in a 67-digit number of graphs [18]. The number of connections varies from hundreds such as drug-targets interactions to 100 trillion in the case of brain neurons.

The graphs have different shapes and the current review is using a few of them such as star graphs, spiral graphs, lattice graph and contact networks (Fig. 4).

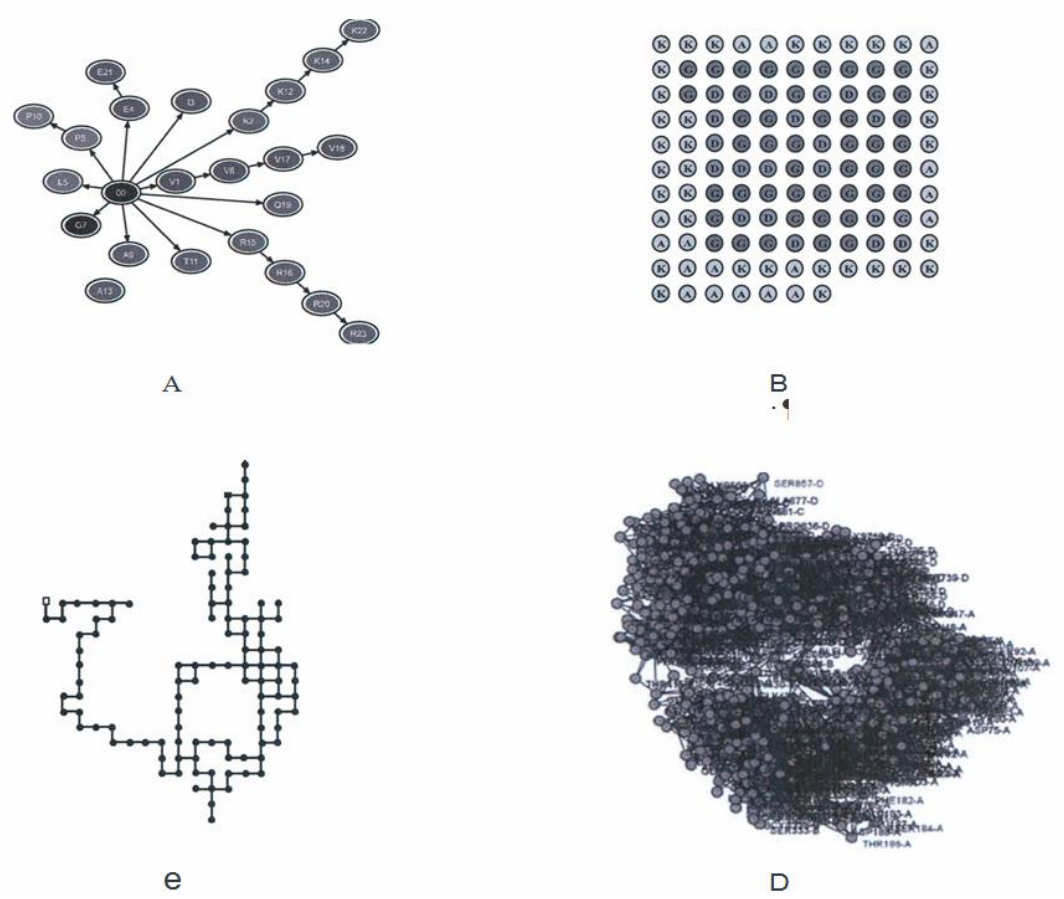


Fig. (4). Shapes of graphs: (A) star graph, (B) spiral graph, (C) lattice graph and (D) contact network.

The current review will focus on the drug, protein and nucleic acid molecules as complex systems, where the nodes are represented by atoms, amino acids and nucleobases and the edges are chemical bonds, spatial contact distance and hydrogen bonds.

2.2. Matrices and Vectors for Graphs

In order to characterize graphs quantitatively, specific invariants or graph descriptors are calculated. This series of graph descriptors are entitled molecular descriptors when the studied system is a molecule or macromolecule. Due to the fact that the topology (connectivity) information is determining this ca1culation, these invariants are called topological indices of the molecular graphs. For the ca1culation of these indices, mathematical elements such as adjacency (connectivity) matrix, node degree, weight vectors, etc. (Fig. 5) are used. All dimensions of these elements depend on the number of nodes *(n),* and the number of edges *(e).* The presentation of the mathematical elements for macromolecules (proteins or nucleic acids) will be more difficult to be presented. Thus, Fig. (5) presents a simple molecule such as toluene with only *n=7* (nodes = atoms) and *e=7* (links/edges = chemical bonds). The resulting elements are presented as follows:

- Connectivity (adjacency) matrix (**M**) has dimension n x n = 7 x 7, with boolean values of "1" if there is a graph connection and "0" for the opposite case; Transaction probability matrix (**P**) with p_{ij} elements obtained by division of m_{ij} elements of the connectivity matrix by the sum(m_{ij}) for each row *i;*

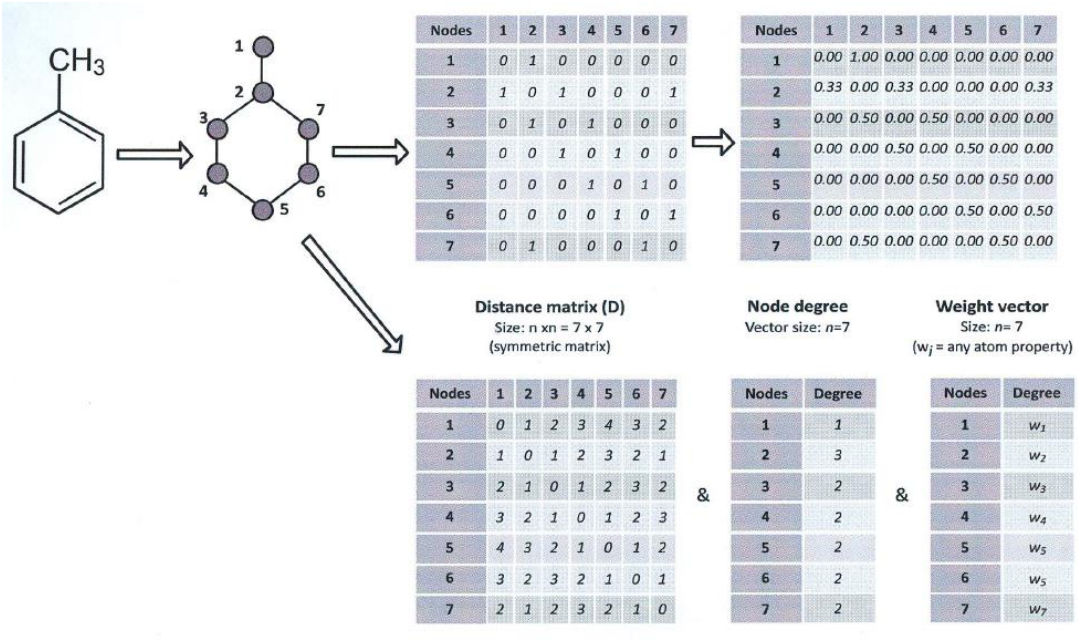
- Distance matrix (**D**) with d_{ij} elements as the length of a shortest path that can connect vertices *i* and *j* in a specific graph. Therefore, the distance matrix contains more information compared with the adjacency matrix, by giving the exact distance between the nodes instead of telling only whether or not two vertices are connected.

- The node degree vector (**deg**) has n = 7 elements (deg_j) that correspond to the number of connections for each node and it can be obtained by surnming the connectivity elements by rows;

- The weight vector (**w**) has n elements (w_j) of weights such as any physical-chemical property of atoms (ex.: electronegativity).

The resulted matrices and vectors are similar for any other type of molecule or graph (star graph, spiral graph, contact networks, etc.). In the next step, these mathematical elements are used to generate topological indices as invariants for a molecular structure. Therefore, using a database of molecules with a specific biological activity and with these calculated topological indices (TIs), it is possible to build a prediction model for new molecular structures such as a Quantitative Structure - Activity Relationship (QSAR) [19].

The Markov Chain theory [20] is used to include the probability interactions between nodes at *k* distance. Therefore, matrix **P** is powered by *k* (0-5) and the calculated TIs will be for a specific *k.* The matrices and vectors for macromolecule graphs are similar to the ones presented above, the differences being the increased number of nodes and edges, and the type of links/nodes defined by the type of graphical representation such as star, spiral, lattice and contact graphs. In the case of the new information-based descriptors, the mathematical elements are linked with the Shannon entropy and event-based matrices [21]. In the next sections, couples of graph types are presented: star, spiral, lattice graphs and contact networks.



Nodes	1	2	3	4	5	6	7
1	0	1	0	0	0	0	0
2	1	0	1	0	0	0	1
3	0	1	0	1	0	0	0
4	0	0	1	0	1	0	0
5	0	0	0	1	0	1	0
6	0	0	0	0	1	0	1
7	0	1	0	0	0	1	0

Nodes	1	2	3	4	5	6	7
1	0.00	1.00	0.00	0.00	0.00	0.00	0.00
2	0.33	0.00	0.33	0.00	0.00	0.00	0.33
3	0.00	0.50	0.00	0.50	0.00	0.00	0.00
4	0.00	0.00	0.50	0.00	0.50	0.00	0.00
5	0.00	0.00	0.00	0.50	0.00	0.50	0.00
6	0.00	0.00	0.00	0.00	0.50	0.00	0.50
7	0.00	0.50	0.00	0.00	0.00	0.50	0.00

Nodes	1	2	3	4	5	6	7
1	0	1	2	3	4	3	2
2	1	0	1	2	3	2	1
3	2	1	0	1	2	3	2
4	3	2	1	0	1	2	3
5	4	3	2	1	0	1	2
6	3	2	3	2	1	0	1
7	2	1	2	3	2	1	0

Nodes	Degree
1	1
2	3
3	2
4	2
5	2
6	2
7	2

Nodes	Degree
1	w_1
2	w_2
3	w_3
4	w_4
5	w_5
6	w_5
7	w_7

Fig. (5). Mathematical elements for the molecular graph of toluene.

2.3. Star Graphs

The SG representation can be obtained for any type of character sequence, including the genetic or protein primary sequence, a string of l-letters corresponding to nucleobases or amino acids (Fig. 6). The star graph (SG) containing *n* nodes (Fig. 4A) has one node with *n*-1 degrees of freedom (center of the star) and the other nodes *(n*-1*)* with only one degree of freedom [2]. The center of the star graph is a dummy node that does not correspond to any items of the represented system such as proteins with amino acids as nodes or nucleic acids with nucleobases as nodes (vertices).

The no des are grouped into branches ("rays") using classes of nodes. In the case of the proteins, it is possible to use 20 branches that correspond to the 23 types of amino acids: the standard and the non-standard ones (Fig. 6A). The groups (SG branches) for the peptide sequences are as follows: A, R, N,D, C, E, Q, G, H, 1, L, K, M, F, P, S, T, W, Y, V, U, 0, and the set of B, Z, J, X. The amino acids can be grouped by any other amino acid property such as hydrophilic properties, electronegativity, etc. In the case of nucleic acids, there are 4 branches for each type of nucleobases (Fig. 6B): A, T, G and C.

The sequence of amino acids/nucleobases can be transformed in a SG by reading the sequence from the left to the right and placing each item as a node in a specific branch defined as groups. This type of graph is a non-embedded one. If the initial connectivity of the amino acids / nucleobases from the sequence is added to the graph (each element is connected with the neighborhood elements in the sequence), the resulted one is an embedded graph (see Fig. 6). The standard types of files with these sequences are the FASTA formats. Moreover, the SG can be applied to numeric sets such as proteome mass spectra and electroencephalogram (EEG). Numerical intervals can be defined as a character (a group) and, therefore, a set of numbers can be converted in a string of characters. The obtained string will be transformed in a SG, similar with the protein or nucleic acid sequences.

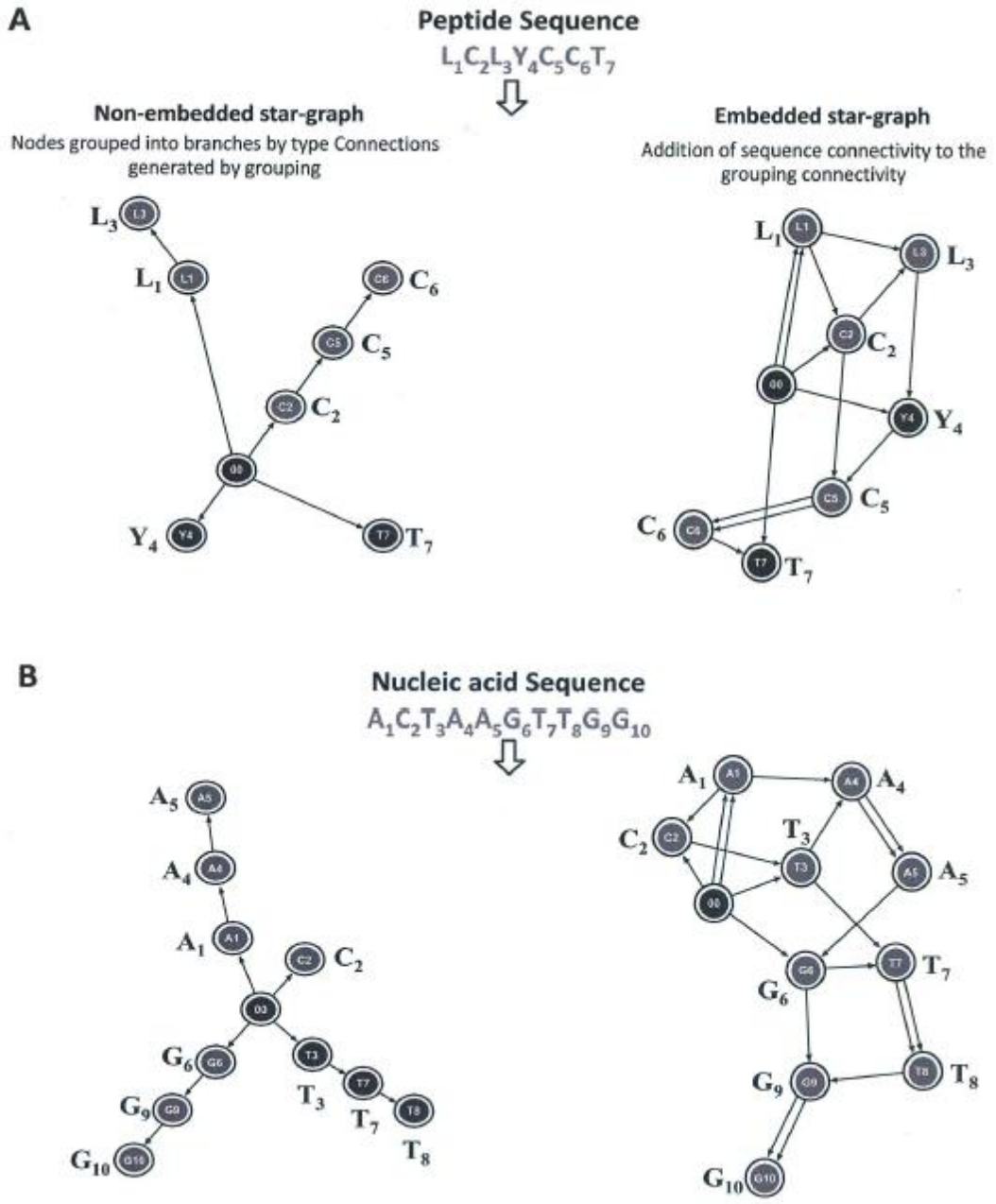


Fig. (6). Star graph generation using peptide (**A**) and nucleic acid (**B**) sequences.

In order to numerically compare graphs, the corresponding matrices and vectors are used. These elements are encoding information about a specific graph topology. TIs are calculated using normalized matrices. The following TIs [22] of molecular SGs can be calculated:

– Trace of the *k* connectivity matrices *(Tr_n)*:

$$Tr_k = \sum_i (M^k)_{ii,} \quad (1)$$

where *k* has values from 0 to the power limit], *M* is the connectivity matrix of the graph with *n*n* dimension, *ii* represents the i[th] diagonal element and *i* has values between 1 and *n.*

– Harari number *(H)*:

$$H = \sum_{i<j} \frac{m_{ij}}{d_{ij,}} \tag{2}$$

where d_{ij} are distance matrix elements and m_{ij} are the M elements;

– Wiener index *(W):*

$$W = \sum_{i<j} d_{ij}, \tag{3}$$

– Gutman topological index *(S_6):*

$$S_6 = \sum_{ij} deg_i \, ^{*} deg_j / d_{ij}, \tag{4}$$

where deg_i are the degree matrix elements;

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

$$S = \sum_{i<j} \left(deg_i + deg_j\right) \, ^{*} d_{ij}, \tag{5}$$

– Balaban distance connectivity index *(J):*

$$J = (edges - nodes + 2) \, ^{*} \sum_{ij} m_{ij} \, ^{*} \mathrm{sqrt}\left(\sum_k d_{ik} \, ^{*} \sum_k d_{kj}\right), \tag{6}$$

where *nodes*+1 = Number of AA / Number of nodes of the Star Graph + origin, $\sum_k d_{jk} \sum_k d_{ik}$ is the node distance degree;

– Kier-Hall connectivity indices *(^nX):*

$$^{0}X = \sum_i 1 \,/\, \mathrm{sqrt}(deg_i), \tag{7}$$

$$^{2}\mathrm{X} = \sum_{\mathrm{i<j<k}} \mathrm{m_{ij}} * \mathrm{m_{jk}} / \mathrm{sqrt}\left(\mathrm{deg_i} \, ^{*} \mathrm{deg_j} \, ^{*} \mathrm{deg_k}\right) \tag{8}$$

$$^{3}\mathrm{X} = \sum_{\mathrm{i<j<k<m}} \mathrm{m_{ij}} \, ^{*} \mathrm{m_{jk}} \, ^{*} \mathrm{m_{km}} / \mathrm{sqrt}\left(\mathrm{deg_i} \, ^{*} \mathrm{deg_j} \, ^{*} \mathrm{deg_k} \, ^{*} \mathrm{deg_m}\right), \tag{9}$$

$$^{4}\mathrm{X} = \sum_{\mathrm{i<j<k<m<o}} \mathrm{m_{ij}} \, ^{*} \mathrm{m_{jk}} \, ^{*} \mathrm{m_{km}} \, ^{*} \mathrm{m_o} / \mathrm{sqrt}\left(\mathrm{deg_i} \, ^{*} \mathrm{deg_j} \, ^{*} \mathrm{deg_k} \, ^{*} \mathrm{deg_m} \, ^{*} \mathrm{deg_o}\right) \tag{10}$$

$$^{5}\mathrm{X} = \sum_{\mathrm{i<j<k<m<o<q}} \mathrm{m_{ij}} \, ^{*} \mathrm{m_{jk}} \, ^{*} \mathrm{m_{km}} \, ^{*} \mathrm{m_o} \, ^{*} \mathrm{m_q} / \mathrm{sqrt}\left(\mathrm{deg_i} \, ^{*} \mathrm{deg_j} \, ^{*} \mathrm{deg_k} \, ^{*} \mathrm{deg_m} \, ^{*} \mathrm{deg_o} \, ^{*} \mathrm{deg_q}\right) \tag{11}$$

– Randic connectivity index *(^1X):*

$$^{1}X = \sum_{ij} m_{ij} / \mathrm{sqrt}\left(deg_i \, ^{*} deg_j\right), \tag{12}$$

2.4. Spiral Graphs

The Ulam spiral graph is based on the geometrical shape that adopts the prime numbers when placing the natural numbers into a spiral [23]. In 1963 the mathematician Stanislaw M. Ulam has discovered this topology. The steps for constructing a spiral graph are as follows: write down a regular grid of numbers, starting with one at the center, and spiraling out the rest of integer numbers (see Fig. 7 A). The numerical series and sequences can be used to plot numbers that reveals hidden patterns. In molecular sciences, the spiral graph was used to represent DNA nucleobases sequences defined by a sequence of only four classes: A, T, G, and C.

The Ulam spiral is divided into different regions entitled gnomons (see Fig. 7B). The gnome is defined by remembering the oblong numbers. This numbers can be represented by the product $n(n+1)$ with *n:* 2, 6, 12, 20, 30, 42, 56, 72, ... These numbers divide into natural numbers in different intervals growing in size *(2n).* Thus, a gnomon is defined by a serial couple of oblong numbers and it creating growing size rectangles. It can be observed that each element of the spiral belongs to only a gnomon. In consequence, each element has Ulam coordinate U_n into one gnomon. Fig. (7C) shows the correspondent spiral graph by using letters.

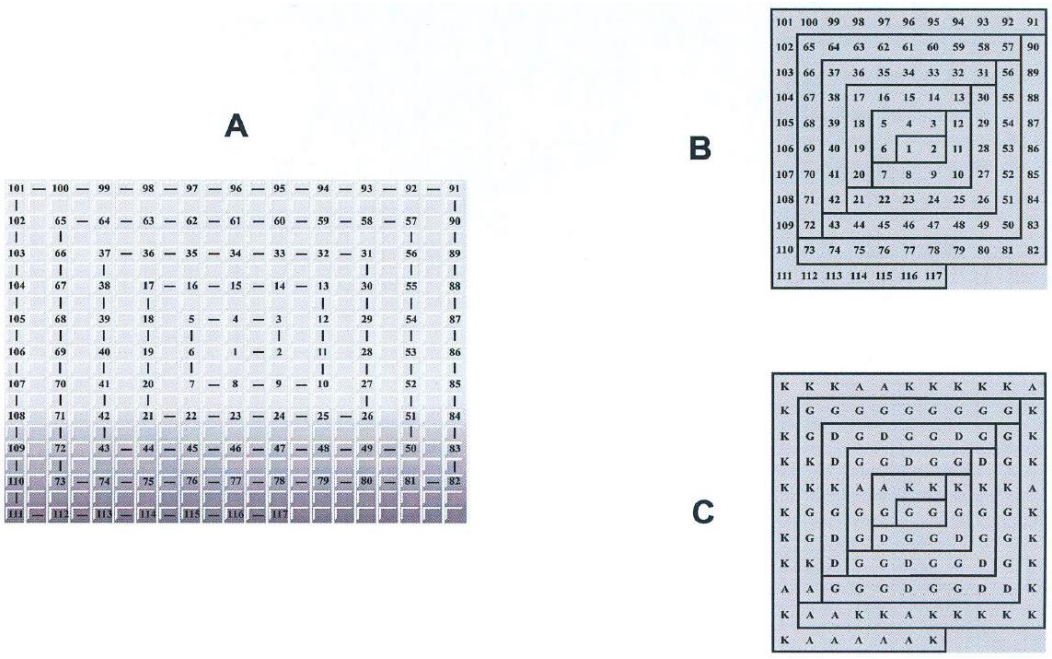


Fig. (7). Spiral composed by numbers (**A**), gnomons division using numbers (**B**) and gnomons division using letters (**C**).

Fig. (8) presents the peptide and nucleic acid sequences and the corresponding spiral graphs. Typically the input sequences can be downloaded as FASTA files. For a spiral graph, there are two types of TIs: frequencies (Fr) and Shannon Entropies (Sh). The indices of the spiral graph can be ca1culated as following:

- *By classes in gnomons:* the two TI types are ca1culated for each class in each gnomon. If a class is not present in a certain gnomon, its Frequency and Shannon Entropy in this gnomon take zero values. This type of ca1culation is used for the case when the sequences contains only few classes and the sequences are not very large. Otherwise, a high number of indices would be obtained and it could make more difficult the further statistical process.

- *By classes in global graph:* TIs are calculated for each class in the whole graph as the sum of their values in all the gnomons. Thus, the number of Tis is reduced in the case of large sequences.

- *By gnomons:* TIs are calculated at gnomon level, independently of the classes as the sum of the TIs of all the classes in a gnomon. This methods is great for sequences with a lot of classes but with a moderate length.

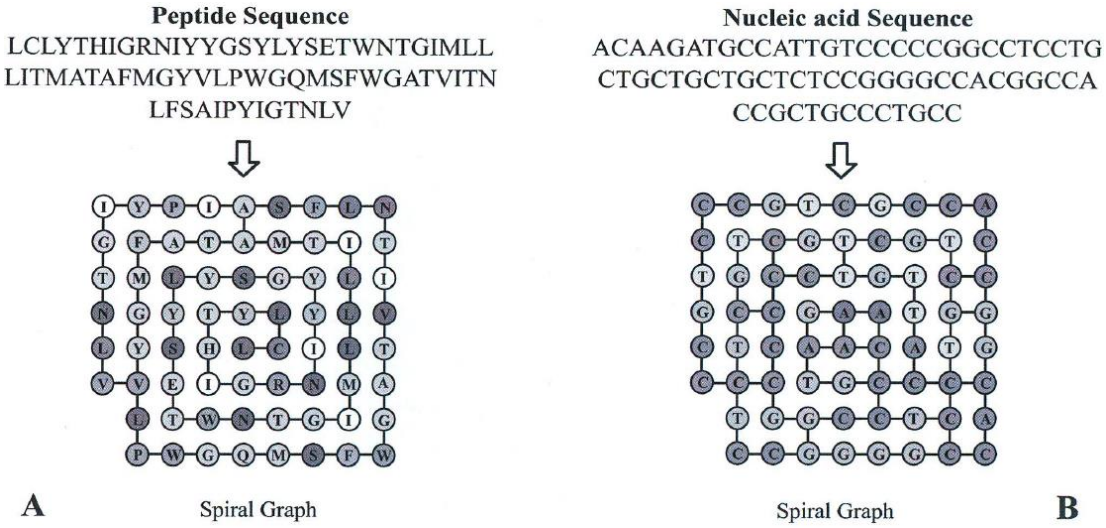


Fig. (8). Spiral graph for a peptide (**A**) and nucleic acid (**B**) primary sequence

In the spiral graph, each node has a class and the nodes are connected to the following sequence letter, and to the nodes from the same class (with the same letter). By definition, the node degree is the number of connections with the other nodes, and the graph total degree represents the sum of all node degrees. In addition, it can be defined the gnomon degrees as the sum of the degrees of the nodes present into a specific gnomon.

Thus, the topological indices of a spiral graph ca1culated by classes in the global graph can be:

- Frequencies:

$$Fr(c) = \sum_n \left[deg\big(n(c)\big) / \sum_i deg(i) \right] \tag{13}$$

where *c* represents the class, n_c is a no de from class e into the spiral graph G_u, *deg* represents the node degree;

- Shannon Entropies:

$$Sh(c) = -Fr(c) * \log(Fr(c)) \tag{14}$$

2.5. Lattice Graphs

The visualization and numerical characterization of the biological/chemical information is extensively presented in Ref. [24]. Several scientists have designed different types of graph-based representations. One of this shape is the lattice- like patterns [25,26].

Let us consider a general example such as a set of elements *(n)*, identified by s_j *(j* = 1...*n)*, and arranged into a sequence / numerical series. An example in this sense is the one-letter code for DNA / protein sequences, chromosome genes, microarray data or amplitude signals of the proteome mass spectrum The processing of the information with the lattice graph (LG) representation has the following steps: in the first one, all these elements s_j are arranged as a vector $s = [s_1\ s_2,\ s_3,\ ...\ s_j,\ ...\ s_n]$; in the next step, for each element s_j is assigned one or more (up to *m)* properties / weights (^kw_j) memorized as vectors:

$$^1w = \left[{}^1w_1,\ {}^1w_2,\ {}^1w_3, ...\ {}^1w_j, ...\ {}^1w_n \right]$$

$$^2w = \left[{}^2w_1,\ {}^2w_2,\ {}^2w_3, ...\ {}^2w_j, ...\ {}^2w_n \right]$$

...

$$^kw = \left[{}^kw_1,\ {}^kw_2,\ {}^kw_3, ...\ {}^kw_j, ...\ {}^kw_n \right]$$

...

$$^mw = \left[{}^mw_1,\ {}^mw_2,\ {}^mw_3, ...\ {}^mw_j, ...\ {}^mw_n \right]$$

An example is the mass spectra data where the elements s_j can be considered the *n* signals in the MS and at least two weights can be assigned to each signal s_j the mass/charge ratio (m/z), of s_j and the intensity I_j of s_j. Thus, the following vectors are generated: $^1\mathbf{w} = [(m/z)_1\ (m/z)_2,\ (m/z)_3\ ...\ (m/z)_j,\ ...\ (m/z)_n]$ and $^2\mathbf{w} = [I_1,\ I_2,\ I_3,\ ...\ I_j,\ ...\ I_n]$. In addition, all the elements (s_j) can be grouped into one or several classes (q) using the sets of conditions $\mathbf{C}_q$ (simple or composed). For all the elements of the same class one letter symbol is assigned. An example is the nucleic acid sequence, where labels A, T, G, or C can be used for each nucleotide for Adenine, Thymine, Guanine, or Cytosine. Another example are the signal series where the labels such as H or L can be used for each signal s_j in an MS if the intensity value I_j is Higher (H) or Lower (L) than the MS average intensity.

In the next step, the information is graphically processed by assigning each element or signal of the sequence a node graph with the Cartesian coordinates $\mathbf{r}_2 = (x, y)$ (2D Euclidean space). The first node (sometimes not a data point) is placed into the center of the system at coordinate $\mathbf{r}_2 = (0,0)$. The coordinates into the lattice graph for the next nodes are calculate using the same methods for DNAs [27], using multiple weight ^kw_j of the elements s_j.

a) Increases in +1 the y axe if ^kw_j follow the set of conditions $\mathbf{C}_1$ (upwards-step) or:

b) Increases in + 1 the x axe if ^kw_j follow the set of conditions $\mathbf{C}_2$ but not $\mathbf{C}_1$ (leftwards-step) or:

c) Decreases in -1 the y axe if ^kw_j follow the set of conditions $\mathbf{C}_3$ but not $\mathbf{C}_1$ nor $\mathbf{C}_2$ (rightwards-step) or:

d) Decreases in -1 the x axe otherwise (downwards- step).

This representation allows displaying large sequences into a simple 2D picture such as the lattice graph and it can be considered as the 2D overlapping / alignment maps. The polymer folding by optimizing the lattice structure and resembling the real folding has been described using the pseudo-folding lattice based on hydrophobicity and polarity (HP) [28]. Therefore, the 2D

graph representations for DNNRNA and protein sequences have been introduced by several scientist [29][30] and it was entitled as polymer sequence pseudo-folding lattice networks due to the lattice- like shape where the sequences fold in a non-natural way.

Therefore, similar with Nandy's DNA representation, a new 2D-lattice for protein sequence was introduced [31-33]. For each of the four amino acid groups there is a single axis direction, according to the amino acid physicochemical praperties [34]. Using polarity and acid character, four classes of amino acids can be defined as polar, non-polar, acid and basic. The positive or negative charge of the amino acids prevails over the polar/non-polar property (the classes do not overlap). Thus, the vector $\mathbf{s} = [s_0, s_1, s_2, ... sj, ... s_n]$ is used to record the labels of *n* amino acids s_j, (from the protein sequence). Two weight vectors have been used to numerically characterize s_j vector ${}^1\mathbf{w} = [q_0, q_l, q_2, ... q_j, ... q_n]$ that contains the electrostatic charges and vector ${}^2\mathbf{w} = [\mu_1, \mu_2, \mu_3, ... \mu_i, ... \mu_n]$ with the dipolar moments of all the amino acid.

The sets of conditions consist of logical order operations. The node of the initial amino acid s_0 is placed at the coordinates (0,0) in a Cartesian 2D space. The coordinates of the successive amino acids are calculated with the same method for DNA lattice graphs:

$\mathbf{C}_1$: Increases in + 1 the y axis if $q_j > 0$ (upwards-step) or:
$\mathbf{C}_2$: Increases in + 1 the x axis if $q_j = 0$ and $\mu_j \neq 0$ (rightwards- step) or:
$\mathbf{C}_3$: Decreases in -1 the y axis if $q_j < 0$ (downwards-step) or:
$\mathbf{C}_4$: Decreases in -1 the x axis otherwise (leftwards-step).

Compared with the DNA lattice graph, the difference is that the new representation for protein sequences contains 20 amino acids, not 4 base types. Therefore, these amino acids are grouped into only 4 groups.

The first Markov Model (MM) entitled MARCH- INSIDE has been used to codify the information of 135 mycobacterial promoter sequences (Mps) [35] and 511 random control group sequences (Cgs). In this methodology any atom, nuc1eotide or amino acid is a states of the Markov Chain (MC). MM calculates the probabilities $({}^k p_{ij})$, where the charge distribution of nucleotide moves from any nucleotide in the vicinity *i* at time t_0 to another nucleotide *j* along the protein backbone as specified by the Markov chain theory [20], in discrete time periods until a stationary state is achieved [36]. Based on previous prediction of RNA from sequences [34], the lattice graphs can encode information about the Mps, similarly to the ones for DNA, by using four types of nucleotides. Table 1 and Fig. (9) present the 2D lattice graph for Mps of the gene Alpha in *Mycobacterum bovis* (BCG). The graph construction rules for the next nucleotide are described below:

a) Increase by + 1 the abscissa axis coordinate for thymine (rightwards-step) or:

b) Decrease by -1 the abscissa axis coordinate for cytosine (leftwards-step) or:

c) Increase by + 1 the ordinate axis coordinate for adenine (upwards-step) or:

d) Decrease by -1 the ordinate axis coordinate for guanine (downwards-step).

DNA Lattice Network			
	$c_1g_2a_3c_4t_5t_6t_7c_8g_9c_{10}c_{11}c_{12}g_{13}a_{14}a_{15}t_{16}c_{17}g_{18}a_{19}c_{20}$ $a_{21}t_{22}t_{23}t_{24}g_{25}g_{26}c_{27}c_{28}t_{29}c_{30}c_{31a32}c_{33}a_{34}c_{35}a_{36}c_{37}g_{38}g_{39}t_{40}$ $a_{41}t_{42}g_{43}t_{44}t_{45}c_{46}t_{47}g_{48}g_{49}c_{50}c_{51}c_{52}g_{53}a_{54}g_{55}c_{56}a_{57}c_{58}a_{59}c_{60}$ $g_{61}a_{62}c_{63}g_{64}a_{65}$		
n	**Nucleotide**	**x**	**y**
1	$c_1a_3t_5g_{25}$	0	0
2	$g_2c_{10}g_{26}$	0	-1
3	c_4t_{16}	-1	0
4	t_6c_8	1	0
5	t_7	2	0
6	g_9	1	-1
7	$c_{11}c_{27}t_{29}$	-1	-1
8	$c_{12}a_{14}g_{18}c_{28}c_{30}g_{48}$	-2	-1
9	$g_{13}g_{49}$	-2	-2
10	$a_{15}c_{17}a_{19}t_{45}t_{47}$	-2	0
11	$c_{20}a_{32}t_{44}c_{46}$	-3	0
12	a_{21}	-3	1
13	t_{22}	-2	1
14	t_{23}	-1	1
15	t_{24}	0	1
16	c_{31}	-3	-1
17	$c_{33}g_{43}$	-4	0
18	$a_{34}t_{42}$	-4	1
19	$c_{35}a_{41}$	-5	1
20	a_{36}	-5	2
21	c_{37}	-6	2
22	g_{38}	-6	1
23	g_{39}	-6	0
24	t_{40}	-5	0
25	c_{50}	-3	-2
26	c_{51}	-4	-2
27	$c_{52}a_{54}$	-5	-2
28	$g_{53}g_{55}$	-5	-3
29	c_{56}	-6	-3
30	a_{57}	-6	-2
31	c_{58}	-7	-2
32	a_{59}	-7	-1
33	$c_{60}a_{62}$	-8	-1
34	g_{61}	-8	-2
35	$c_{63}a_{65}$	-9	-1
36	g_{64}	-9	-2

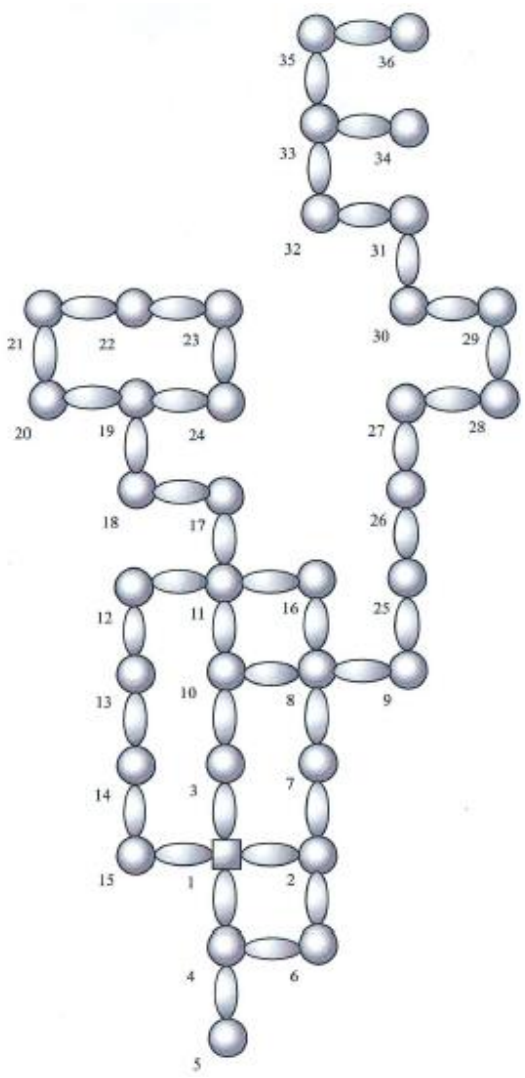


Fig. (9). Lattice Graph for the Mps of gene Alpha in *Mycobacterum bovis* (BCG).

The basis for the TI ca1culation is represented by a stochastic matrix $^{1}\mathbf{\Pi}$ for each lattice graph with the elements as probabilities $^{1}P_{ij}$ of reaching node n_i with the charge Q_i moving through a node path of length $k = 1$ from another node n_j with charge Q_j [34]:

$$P_{ij} = \frac{\frac{Q_j}{d_{j0}}}{\sum_{m=1}^{n} \alpha_{il} \cdot \frac{Q_j}{d}_{l0}} = \frac{\phi_j}{\sum_{m=1}^{n} \alpha_{il} \cdot \phi_1} \quad (15)$$

$$P_j = \frac{\frac{Q_j}{d_{j0}}}{\sum_{m=1}^{n} \frac{Q_j}{d}_{l0}} = \frac{\phi_j}{\sum_{m=1}^{n} \phi_1} \quad (16)$$

α_{ij} is 1 if nodes n_i and n_j are adjacent or 0 otherwise. Q_j represents the sum of the electrostatic charges of all nucleotide placed at this node. The number of no des *(n)* of the graph is equal to the number of rows or columns in $^{1}\mathbf{\Pi}$ and, sometimes, it can be smaller than the number of DNA bases in the sequence. Three families of invariant lattice graph TIs can be ca1culated for the DNA sequence:

$$^{LG}\pi_k = \sum_{i=j}^{n} {}^{k}p_{ij} \quad (17)$$

$$^{LG}\xi_k = \sum_{i=j}^{n} {}^{k}p_j \cdot \phi_j \quad (18)$$

$$^{LG}\theta_k = -\sum_{i=j}^{n} {}^{k}p_j \cdot \log\left({}^{k}p_j\right) \quad (19)$$

$^{LN}\pi_k$ are the Markov sgectral moments as the sum of the matrix main diagonal: $^{LG}\pi_k$= Tr($^{k}\mathbf{\Pi}$) = Tr[($^{1}\mathbf{\Pi}_k$)]= (Tr = trace). $^{LG}\xi_k$ are the mean values of the electrostatic potentials and $^{LG}\theta_k$ represent the Markov entropies.

2.6. Contact Networks

The contact networks (CN) define the edges using the spatial distance (3D information) to determine if two networks elements are in contact. The main applications for CN are proteins. Therefore, the system of a protein molecule, in which the nodes are the amino acids, is transformed in a CN using the Cartesian 3D coordinates (x, y, z) ofthe alpha- carbons from the PDB files (protein Data Bank) [37]. The nodes are connected if the distance between two amino acid alpha carbon atoms is les s than a *cutoff* value (r_{off}, default values = 7Å) (see Fig. l0A). Depending on the distance from the geometrical center of the protein, each amino acid can be virtually localized in spherical regions *(R)* such as core *(c),* inner *(i),* middle *(m)* and surface *(s)* (Fig. 10B). The sum of these regions represents the total space of the protein as the *total* region *(t).* The diameters of these regions are the percentage of the longest distance r_{max} with respect to the protein chain geometrical center: e between 0% and 25%, *i* between 25% and 50%, *m* between 50% and 75%, and s between 75% and 100%. Additionally, there is a total region *(t)* with diameter

between 0% and 100%. The Markov Chain theory is used to ca1culate the probabilities to interact of any two amino acids placed at a topological distance *k* with integer values between 0 and 5. For each region *R*, the obtained values are averaged using all values of *k*.

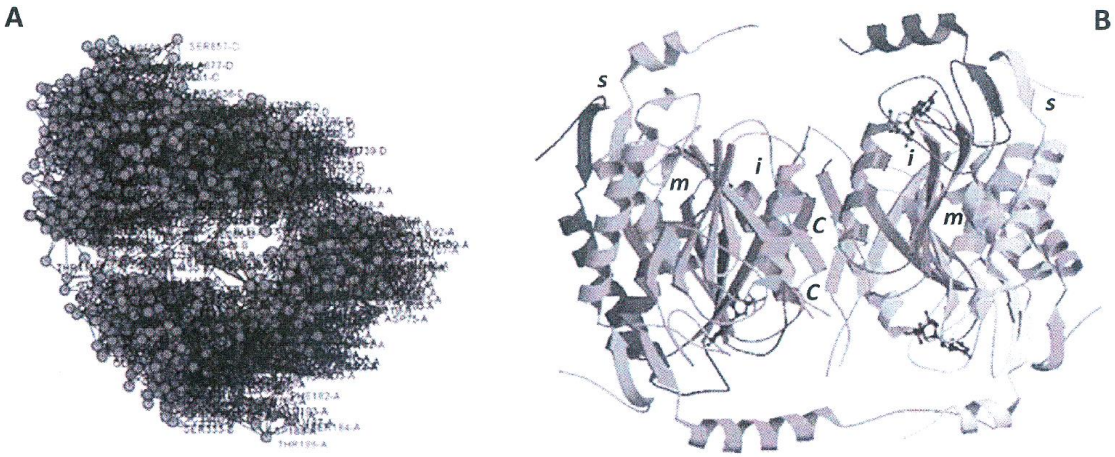


Fig. (10). Contact networks for a protein (A) and the protein regions (B).

3. SOFTWARE FOR MOLECULAR GRAPH INDICES

Because almost anything can be divided in elements related by properties, the graph theory can be used in different studies such as search of molecular targets, interactions between macromolecules, drug discovery, metabolic pathways, diseases analysis, etc. This work is focused on molecular systems such as drugs, proteins and nuc1eic acids.

The molecular descriptors are the main role into the QSAR model searching. This section is focused on couple applications that are used for the calculation of molecular descriptors (TIs and other types) [22]: Chemistry Development Kit (CDK), DRAGON, MoDesLab, ToMoCoMD, MARCH-INSIDE, E-Ca1c and CODESSA PRO. These applications can calculate over 10,000 molecular descriptors in order to be used in Bioinformatics, Chemoinformatics, Drug Design, or Medicinal Chemistry. These tools are generally dedicated to small and medium molecules but some of them, such as MARCH-INSIDE, can ca1culate descriptors for macromolecules. In another subsection, three in-house desktop tools for calculating topological indices for macromolecules (proteins/nucleic acids) are presented, such as S2SNet [38] for star graphs, CULSPIN [39] for spiral graphs and MInD-Prot [40] for contact networks.

3.1. Molecular Descriptor Tools

The Chemistry Development Kit (CDK) is an open- source lava library for Chemoinformatics and Bioinformatics created by Christoph Steinbeck, Egon Willighagen and Dan Gezelter [41]. CDK is a library, instead of a user program and, therefore, it has been integrated into various environments such as R (programming language) [42], Bioc1ipse [43], KNIME [44] and Excel (called LICSS, excel- cdk) [45].

DRAGON tool (http://www.talete.mi.it/products/dragon_ description.htm) is a well known tool to calculate an important number of molecular descriptors (inc1uding the most known TIs). DRAGON was released in 1994 by the Milano Chemometrics Group using the name of WHIM/3D QSAR [46]. DRAGON ver. 6.0 can calculate 4855 molecular descriptors that are divided into 29 types. E-DRAGON (v. 1.0, http://www.vcc1ab.org/lab/edragon/) is the free online version of DRAGON (v. 5.4) and it allows the calculation of more than 1600 molecular descriptors divided into 20 logical blocks [47]. Some examples in the literature on the use of this software are presented in Refs. [48-50].

MoDesLab (http://www.modeslab.com/) provides different tools in order to perform QSAR studies: from the input of a large number of molecules, for the calculation of molecular descriptors such as Kier and Hall, Kappa, Balaban indices, and Abraham descriptors sub-structural descriptors. In addition, it permits the definition of the atom, bonds and fragments properties, and the use of SMILES formulas [51- 53].

Y. Marrero-Ponce and Romero V. developed a new tool entitled ToMoCoMD. It is composed of four subprograms that allow editing structures (draw mode) and the ca1culation of molecular descriptors 2D/3D (ca1culation mode). The software calculates various types of TIs from algebraic forms such as quadratic, linear and the bi-linear [54]. A recent review has presented many ToMoCoMD applications to QSPR / QSAR studies for anti-parasitic drugs [55].

The new version is entitled as ToMoCoMD-CARDD (http://tomocomd.com/) and it represents an interactive and user-friendly free multi-platform framework with two suites with parallel functionalities. The first suite contains a set of modules derived from algebraic considerations (QuBiLS suite, Fig. 11) such as QuBiLs-MAS (Quadratic, Bilinear and Linear Maps based on Graph-Theoretic Electronic-Density Matrices and Atomic weightings), QuBiLs-MIDAS (Quadratic, Bilinear and N-Linear Maps based on N-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings), and QuBiLs-POMAS (Quadratic, Bilinear and Linear Maps based on Molecular Surface-based Potential Matrices and Atomic Weightings).

The second suite of ToMoCoMD-CARDD consists in a collection of molecular descriptor calculating modules that is using relation frequency matrices, molecular fingerprints and a pool of the most relevant indices reported in the literature such as DIVATI (DIscrete DeriVAtive Type Indices), GT- STAF (Graph Theoretical Thermodynamic STAte Functions), FREMESSA (FREquency-type Matrices Extended claSSical Algorithms), FREMXALF (FREquency-type MatriX-based ALgebraic Forms), MOLFIP (MOLecular FIngerPrints), and DESPOOL (DEScriptor POOLs). For the handling of the chemical structures and the calculation of the atomic properties, Chemical Development Kit (CDK) library has been used. The framework contains a useful API library, that permits an easily integration with other software for chemo- informatics applications. ToMoCoMD demonstrated the capacity to offer solution for large spectra of problems. Some examples for small molecule are the prediction of tyrosinase inhibitors using the atom linear indices [56], prediction of aquatic toxicity [57], and predicting Caco-2 cell permeability [58]. In the case of macromolecules, the tool has been used to predict the protein stability effects of a complete set of alanine substitutions in the Are repressor [59, 60], and to build the nucleic acid QSAR models [61].

MARCH-INSIDE is a simple calculation method, but it is very effective for QSAR studies in Medicinal Chemistry. It was developed by González-Díaz's team (Fig. 12) and it uses the Markov Chains theory in order to generate numerical parameters describing the chemical structure of the drugs and their molecular targets. In recent review papers, there have been examples of using this program for predicting anti-microbial and anti-parasitic agents, and their molecular targets [55]. Thus, MARCH-INSIDE can be used for macromolecules using lattice graphs and contact networks.

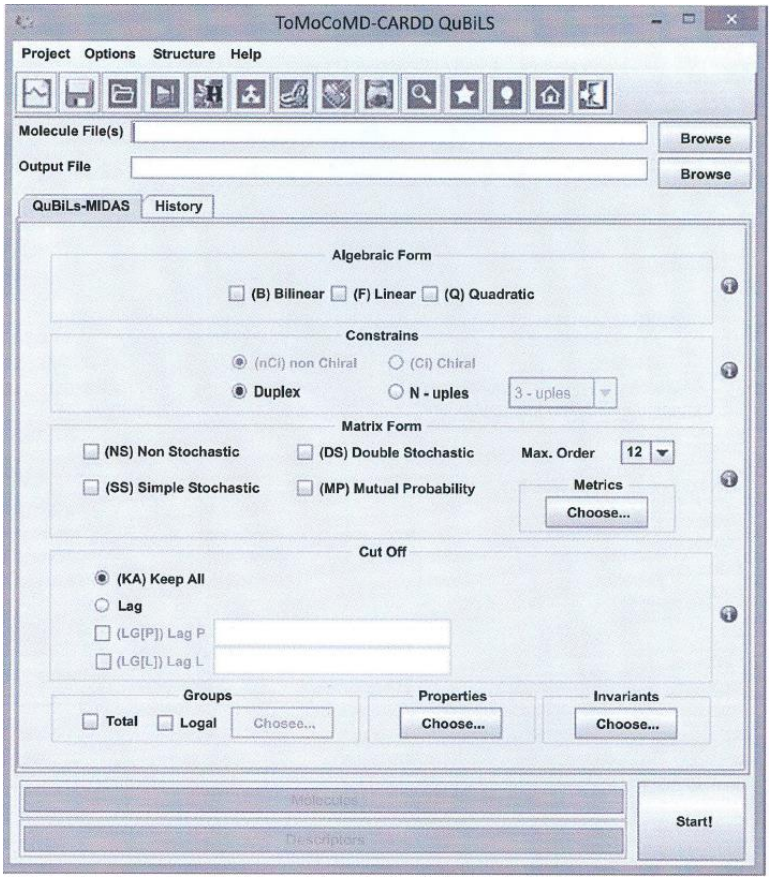


Fig. (11). Graphical interface ofToMoCoMD-CARDD QuBiLS.

E-Ca1c (ver.l.l/1999) is a tool that ca1culates Electro- topological state indices (E-values) of molecules, including the Electro-topological State (E-State) and hydrogen E-State (HE-State) values of the individual atoms and the atom ratios. Parts from Mo1conn SciQSAR-Z and 2D [62] have been used for the TIs calculation.

CODESSA PRO, Comprehensive Descriptors for Structural and Statistical Analysis (http://www.codessa- pro.com) is a tool designed by Alan R. Katritzky, Karelson Mati and Ruslan Petrukhin. It is designed to build QSAR/QSPR models by integrating all necessary mathematical measures and computational tools to (i) ca1culate a range of molecular descriptors using the 3D geometric structure and lor the quantum-mechanical wave function of chemical compounds, (ii) develop QSPR models for linear and non-linear chemical and physical properties or biological activity of chemical compounds, (iii) conduct a cluster analysis of experimental data and molecular descriptors, (iv) interpret the developed models, and (v) predict the property values of any chemical with a known molecular structure. CODESSA PRO includes 116 molecular descriptors divided into 8 groups: constitutional, topological and geometric, electrostatic CPSA, quantum chemical, related to molecular orbitals and thermodynamics. Some examples of the use of this tool are described in Refs. [63- 66].

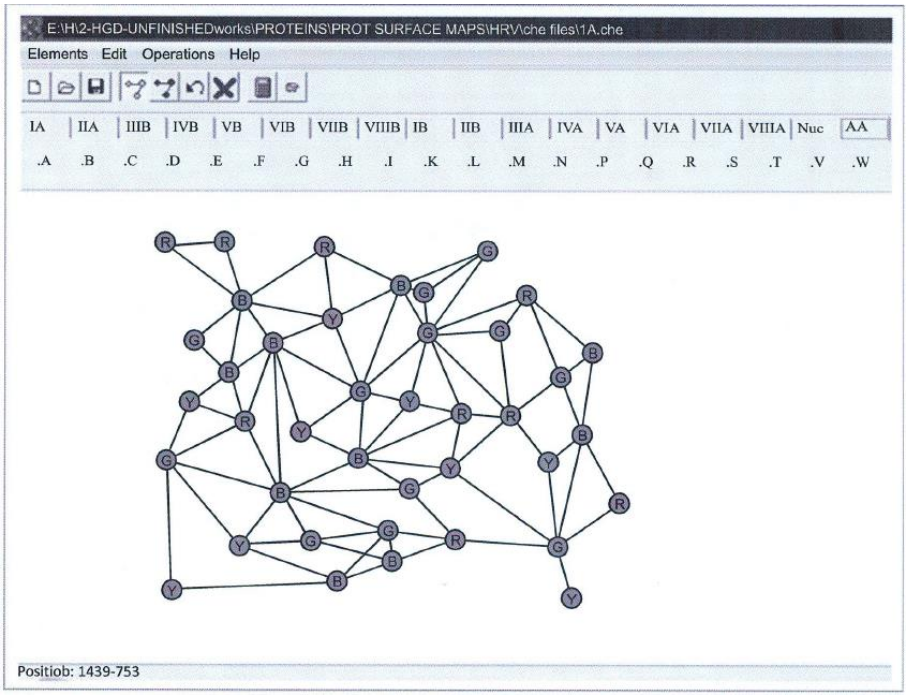


Fig. (12). MARCH-INSIDE interface

3.2. Tools for Markov Topological Indices of Macromolecules

The topological indices are based on the molecular topology and they combine this information with other physical-chemical properties such as the electronegativity for amino acids. The Markov topological indices use the Markov Chain theory in order to include the node transition (interaction) placed at a distance *k* or after *k* steps of a Markov chain. Munteanu's team developed several graph- based scientific applications to calculate the topological indices for an entire system as a graph or for each node from a system. The dedicated tools for the ca1culation of the Markov TIs are S2SNet, CULSPIN and MInD-Prot.

3.2.1. S2SNet

S2SNet, Sequence to Star Network [38] is a free Python tool with the user interface programmed in wxPython [67] and with Graphviz to plot the resulted graphs (Fig. 13). S2SNet is able to encode any sequence of characters into Star Network (SN) topological indices (TIs) and it can plot the resulted graphs. There are several examples of sequences: peptide amino acid chains, DNA/RNA strands and mass spectra data. The current desktop version was compiled under Microsoft Windows XP/Vista/7. The application is available for free upon request for academic use only. S2SNet can carry out the following tasks: it can transform l-character sequences into topological Star Network indices, it can transform numeric data into 1- character sequences, and it can transform N -character sequences into l-character sequence using a different codification. The same tool can edit/view the user's input and output text files, create DOT language files, and plot and display networks as PNG images.

MARCH-INSIDE [68] can generate Spiral Graphs from a list of sequences and Star Graphs only from the graphical interface, one by one. In addition, there are software such as Centibin [69] and Pajek [70] that can calculate some S2SNet topological indices and it can process file by file (mat/net). Therefore, S2SNet has several advantages: it is dedicated to SG calculations, it can transform a list of sequences into Star Graph TIs (the sequence can be anything, not only proteins/nuc1eic acids). S2SNet can generate embedded graphs, it calculates an extended list of Star Graphs topological indices, and it can plot the resulted graphs into different Graphviz formats. If the initial data are not a string of characters such as a peptide or nucleic acid sequence, S2SNet has some conversion tools as follows:

– If the input is a numerical set such as an electroencephalography or a mass spectra record, S2SNet can transform the set of values into a string of characters by creating ranges of values as groups (automatically from the input or manually by the user);

– If the input contains 3-letter codes for amino acids or the 3-nucleotide codons, the tool can transform the input into the corresponding string of 1-letter amino acids; ex.: ALA or GCU/GCC/GCA/GCG will be replaced by A.

The calculated TIs include the ones described in Section 2.3: Shannon Entropy *(Sh_n)* , Traces *(Tr_n),* Harary number *(H),* Wiener index (*W*), Gutrnan topological index (S_6), Schultz topological index (non-trivial part) (*S*), Balaban distance connectivity index *(J),* Kier-Hall connectivity indices *(^nX)* and Randic connectivity index *(^1X).* An example of calculation of TIs (non-embedded graph) with S2SNet for 70DC (chain A) protein from Protein Data Bank (http://www.rcsb.org) is presented in Fig. (14).

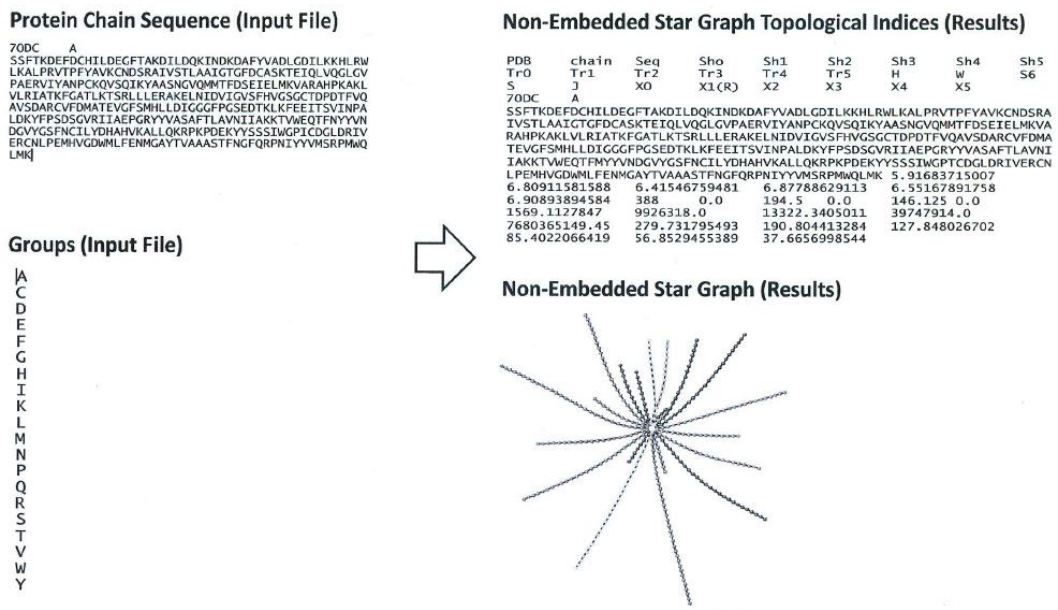


Fig. (14). Example of TI calculation for 70DC (chain A) using non-embedded SG.

S2SNet can be used to numerically characterize any type of system that can be represented as a character string, where a letters represents the elements of the system.

3.2.2. CULSPIN

CULSPIN is an interactive application created with Python/wxPython with a notebook format (Fig. 15) that can transforms any character sequence into a spiral of Ulam by connecting the nodes from the same class (with the same letter). CULSPIN calculates two types of spiral graph TIs, calculated at several levels: for each one of the classes in each Ulam gnomon, for each one of the classes in the whole graph and for each gnomon regardless of the class type. In addition, the 2D graph (U-graphs) generated by the application can be visualized and exported. The calculated TIs could be the input of the statistical analysis in order to find QSAR models. Examples of input sequences: protein amino acids chains, nucleic acids and protein mass spectra.

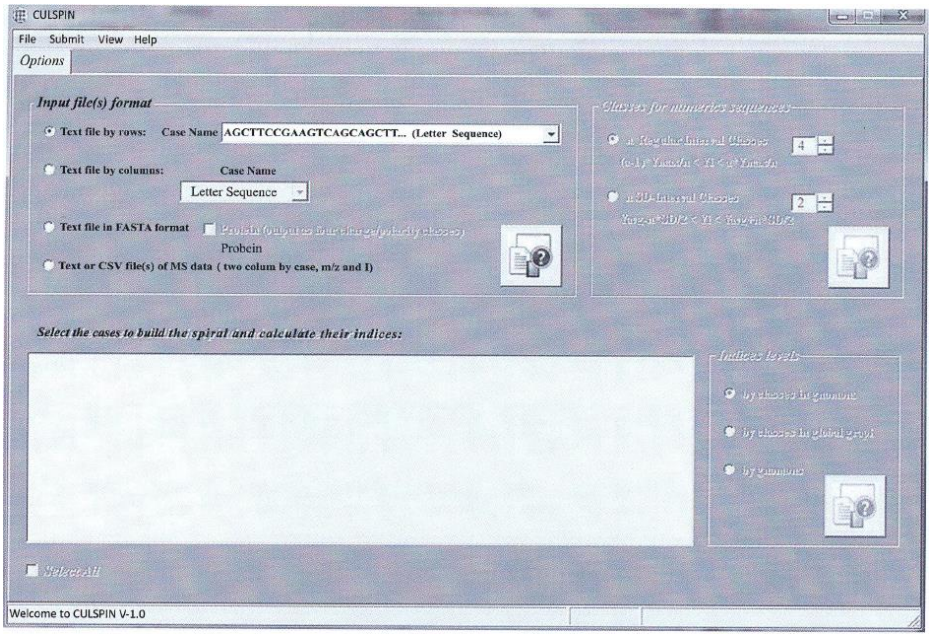


Fig. (15). CULSPIN interface

With CULSPIN, the user can process sequence files, FASTA sequences, numeric/series, mass spectra, convert any sequence into their corresponding U-graph connecting the no des that belong to the same class (they have the same letter), compute two families of Tis, plot the U-graph of the selected sequence, and export the connectivity information of each one of the U-graphs.

The input data can be introduced in different formats as text and numbers, by rows and column, FASTA or CSV. In the case of proteins, if the option Protein is selected, each amino acid present in the sequences is codified in one of the four different amino acids classes determined by their side chain properties: non-polar and neutral; polar and neutral; acidic and polar; and basic and polar.

The details about TIs calculated by CULSPIN (frequencies and Shannon entropies) are presented in Section 2.4. If the numeric input format is used, the tool offers two different heuristics to transform this input into letter sequences:

– *n* Regular Interval Classes: in this option numeric data are divided into *n* intervals or classes (2 :s *n* :s 10) and a letter is assigned to each one of them. Thus, the elements or signs of the numeric sequence are encoded with the letter from the class to which it belongs.

– *n* σ-Interval Classes: in this option the numeric data are divided into *2n+2* intervals ($2 \leq n \leq 4$) that are function of its standard deviation. Each class is assigned a letter and the element or signal of the numeric sequence are encoded with the letter from the class to which it belongs.

In the cases of MS data, this program version transforms the original data into numeric sequences obtained by means of the *m1z* and intensity values multiplication and, subsequently, it makes the transformation in the letter sequences using the heuristic selected by the user.

The spiral graph, similar to the SG, can be used to solve Bio or non-Bio problems using molecular inputs such as proteins/nucleic acids or non-molecular ones such as the mass spectra.

3.2.3. MInD-Prot - Markov Indices for Drugs and Proteins

MInD-Prot tool [40] represents a Python/wxPython application for the ca1culation of the Mean properties Markov indices for drugs and proteins. lt uses the PDB/F ASTA files for proteins (3D coordinates or peptide sequence) as inputs and the SMILES codes for drugs. The user-friendly graphical interface is presented in Fig. (16).

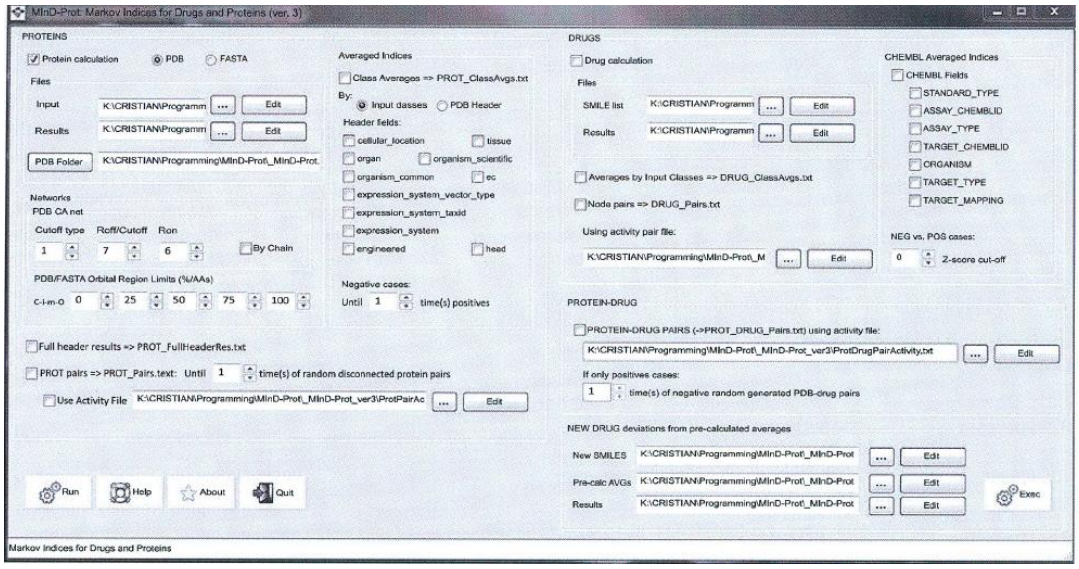


Fig. (16). MInD-Prot graphical interface.

MInD-Prot calculates Markov Mean Properties *(MP)* using different molecule physicochemical properties for the characterization of chemical structures. It is based on the called MARCH-INSIDE (MI), introduced by González-Díaz *et al.* [55, 71, 72]. It uses essentially the same algorithm for all c1asses of molecular structures but performs different approximations for low-weight molecules (drugs, ligands, metabolites) with respect to large bio-polymers (proteins in this case). In the case of proteins, the values of the amino acid physicochemical properties are ca1culated as a sum of all atomic properties from each type of amino acid. Four types of

physicochemical properties have been used: Electronegativity Mulliken (EM), Polarizability Kang-Jhon (PKJ), Van der Waals area (vdWA) [22] and Atom Contribution to P (AC2P) [73].

The representation of a protein chain is considered as a static model where the amino acids are spatially distributed with the corresponding 3D coordinates (x_i, y_i, z_i) for the Cα atoms. These coordinates are used to obtain the amino acid contact network for a protein chain by using a cutoff distance (r_{off}) of 7Å. The amino acids at a distance less than r_{off} are connected (a_{ij} = 1 element in the connectivity matrix **A**). The 3D protein structure is divided into spherical spatial regions *(R):* core (*c*), inner *(i)*, middle *(m)* and surface *(s).* The probabilities to interact any two amino acids placed at a topological distance *k* (0-5) are ca1culated using Markov Chain theory. The obtained values are averaged by all *k* values for each region *R.* Consequently, we can ca1culate different k-averaged parameters (MP_R) for the amino acids from a region *(R = c, i, m,* s, *t)* [74-78] and a specific physicochemical property.

The indices for each physicochemical property are obtained after the following steps:

- Calculation of a squared connectivity matrix of Cα atoms (M) by using the 3D coordinates from a PDB protein file; *n* x *n* matrix where *n* is the number of the amino acids in the protein chain and m_{ij} elements have values of 1 for connected amino acids and 0 for the non-connected ones;

- Calculation of the weighted matrix (**W**) by adding the physicochemical property values for each type of connected amino acid (w_j elements from vector **w** = amino acid weight vector);

- Calculation of the interaction probability matrices ($^k\mathbf{\Pi}$) by the normalization of **W**;

- Calculation of similar interaction probability matrices ($^k\mathbf{\Pi}$) for other *k* steps of interactions *(k =* 0-5), for a specific molecular property;

- The matrices $^k\mathbf{\Pi}$ are used to ca1culate the 3D Markov mean properties corresponding to the entire protein chain, ^kMP_t, for a specific *k* (see Eq. 15); the central matrix $^k\mathbf{\Pi}$ is multiplied from the left by the probability vector $^0\mathbf{p}$ for all amino acids without considering the network connectivity; the result is multiplied from the right by the vector of the amino acid weights (**w**); the values correspond to elements from 1 to *n* (the total number of the amino acids in the protein chain);

- The other *MPs* corresponding to the other protein regions *(c, i, m,* s) are obtained from the same formula by multiplying only the values that correspond to the amino acids in a specific 3D region;

- Finally, the ^kMP_R values are averaged for all *k* values as the Markov Mean Properties MP_R (see Eq. 16).

MInD-Prot [40] calculates a total of 20 descriptors MP_R for each peptide sequence that correspond to 4 types of physicochemical properties, and that are averaged for all the *k* values into 5 regions *R: EMR, PKJ*$_R$, *vdWA*$_R$ and *AC2P*$_R$.

$$^{k}MP_1 = [\ ^{0}p(w_1)^0\ p(w_2)\ \ldots\ \ ^{0}p(w_n)]$$

$$\begin{bmatrix} {}^{1}p_{1,2} & {}^{1}p_{1,2} & {}^{1}p_{1,3} & \cdot & {}^{1}p_{1,n} \\ {}^{1}p_{2,1} & {}^{1}p_{2,2} & {}^{1}p_{2,3} & \cdot & {}^{1}p_{2,n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ {}^{1}p_{n,1} & \cdot & \cdot & \cdot & {}^{1}p_{n\prime n} \end{bmatrix}^{k} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_n \end{bmatrix} \tag{15}$$

$$= \sum_{j=i}^{n} {}^{k}p(w_j) \cdot w_j$$

$$MP_R = \sum_{k=0}^{5} {}^{k}MP_R \tag{16}$$

The indices that are encoding the 3D structure and physicochemical property information of the protein chains can be used as input for the Machine Learning methods from Weka [79]. This way, QSAR c1assification model could be developed to predict specific protein properties.

The main GUI is divided into several parts: *Proteins, Drugs, Protein - Drug Pairs* and *NEW DRUG deviations from pre-calculated CHEMBL averages.* Protein calculations include the following parameters:

– *File parameters:* input PDBchain list file as PDB/PDBchain, protein simple result file, the PDB folder local database; if the PDB is missing, it will be downloaded from the online PDB databank;

– *Alpha Carbon Network parameters: Cutoff, Roff, Ron* parameters for deciding whether two amino acid alpha carbons are linked, protein Orbital Region Limits in percentage as core, inner, middle, outer; *By Chain* calculation: if the user enables this control, the tool will consider the entire protein and will ca1culate all the chains even if you have PDBchains in the input; if this is disabled, the calculation will consider each PDBchain in the list (if the chain info is missing, the entire protein network will be constructed).

– *Averaged Indices* (PROT_ClassAvgs.txt): *by PDB header information* such as head, expression _ system, expression _ system _ taxid, name, chain, organism_ scientific, molecule, express ion _ system _ vector _ type, ec, organism _ common, expression _ system _plasmid, engineered, expression _ system _ strain, cell line, cellular_location, gene, organism taxid or *by input class* from the PDBchain list file (as PDBchain[tab]Class).

– *Full header information output* (PROT_FullHeader Res.txt): it gets the full information from PDBs and adds it to the simplest result (one column for each header field).

– *Protein PAIR:* by using the similarity PDB of the chains as positive pairs and generating the negative ones until X times the positive pairs. As alternative, the activity file can be used (default: ProtPair Activity.txt, with the PDB 1 [tab]PDB2[tab]Class).

Section *Drugs* inc1udes several parameters:

- *File parameters:* SMILES list file (DrugName[tab]SMILES formula), drug simple result file, averaged results by input classes (DRUG_ ClassAvg.txt).

- *Averaged result* (DRUG_ClassAvgs.txt): by using the input c1ass from the SMILES list file (Drug Name[tab]SMILES formula[tab]Class).

Section *Drug PAIRS* always uses the input activity file as DrugName1[tab]DrugName2 [tab]Class; the identification of the drugs in the results is made by drug name, not by SMILES formula. Section *CHEMBL Averaged Indices* uses standard CHEMBL files [80, 81] to calculate the average TIs and the deviations.

The *output* contains the following header information:

- Z-score = classical Z-score using the standard deviation and average for STANDARD_VALUES for each type of STANDARD_TYPE & ST ANDARD_UNIT;

- Activity Class = values of 1 or 0 depending on the Z- score values and the *cutoff* from the GUI; only the records with Activity class of 1 are used to calculate the averages;

- P curate = depends on field CURATED_BY; inc1uded in deviation by multiplication with the difference between TI and Average: if CURATED BY = "Autocuration": P curate = 0.50, if CURATED BY = "Intermediate": P curate = 0.75, if CURATED BY = "Expert": P curate = 1.0.

- All CHEMBL fields from the input, drug Tis, average s (for each TI and CHEMBL field type), deviations between the TIs and Averages (for each TI and CHEMBL field type).

Parameter for CHEMBL calculation:

- CHEMBL input files to calculate the TIS from the control named **SMILES list**;

- The result will be printed in the file from the control named **Results**;

- You can chose the **CHEMBL files** to be used to average the TI values and the corresponding deviation values: STANDARD_TYPE, ASSAY_CHEMBLID, ASSAY_TYPE, TARGET_ CHEMBLID, ORGANISM, TARGET_TYPE, TARGET_MAPPING

- *Cutoffvalues* to calculate the Activity Class using the Z-score values.

- Section PROTEIN-DRUG PAIRS always uses an input activity file as PDBChain[tab] DrugName[tab] Activity. If there is only one type of c1ass (positive cases), it can be generated random protein-drug pairs until X times the positive cases. This option will work only if both PROTEIN and DRUG calculations are enabled.

In section *NEW DRUG deviations from pre-calculated CHEMBL averages*, the previous CHEMBL result file can be used to calculate the deviation of a user list of SMILES using the pre-calculated values of the averages for the corresponding CHEMBL fields. It has the following parameters: New SMILES - file with the user's drug list (DrugName[TAB]SMILES format), Pre-calc AVGs - file resulted from a previous calculation using the CHEMBL fields that contain the averages for TI and field types, and Results - the final output file. The user can create/edit and browse all the files directly from the interface by using the native NotePad from Windows. All the options have default values in order to perform a minimum number of calculations. SMILES code

of drugs is transformed into MOL format using free BABEL software [82] (http://www.eyesopen.com/docs/babel/current/htm1/index.html).

MInD-Prot is dedicated to the ca1culation of drug or protein TIs and the mixed ones. It can average at different levels of information the calculated TIs and it can generate pairs of drug-protein, drug-drug, protein-protein networks that can be used for interaction studies in order to find new drugs, new indications for a known drug or new molecular targets for specific diseases.

3.2.4. lnformation Theory-Based Descriptors

Barigye *et al.* [83] presented an extended description of the information theory-based chemical structure codification in molecular descriptors. The review is presenting the link between the Shannon entropy or the entropy of information and these descriptors that represent "graph invariants that view the molecular graph as a source of different probability distributions to which information theory definitions can be applied" [84]. The sources of information could be the chemical formula (0D molecular structure representation), the chemical graphs, and the matrix representations. In addition, the channel-coding theorem applied to chemical system has been presented. Several information theory-based indices (IFIs) have been discussed: information index on chemical composition, topological information content index, vertex orbital information indices, edge orbital information indices, chromatic information content, molecular symmetry index, connection orbital information content or Bertz index, centric information indices, information bond index, vertex complexity index, information distance index, and electronic delocalization entropy.

Recently, novel IFIs have been presented using versatile event-based approaches [85]. The idea is based on the definition of an event in this context as the criterion followed in the "discovery" of molecular substructures, which consequently are the basis for the generalized incidence and relations frequency matrices construction. Therefore, Shannon's, mutual, conditional and joint entropy-based IFIs could be computed. In previous studies, it was introduced an event as the "connected subgraphs". The new study introduced other types of events such as terminal paths, vertex path incidence, quantum subgraphs, walks of length k, Sach's subgraphs, MACCs, E-state and substructure fingerprints and, Ghose and Crippen atom-types for hydrophobicity and refractivity. In addition, the authors defined magnitude-based IFIs, by introducing the use of the magnitude criterion in the definition of mutual, conditional and joint entropy-based IFIs. The new descriptors have been tested by comparison with the other similar molecular descriptors applied to 34 derivatives of 2-furylethylenes. Details about Shannon's, mutual, conditional and joint entropy information indices could be find in Ref. [21].

Another class of novel graph-theoretical invariant for generating new 2/3D molecular descriptors have been introduced by Marrero-Ponce *et al.* [86] by describing the chemical structures of organic molecules at atomic- molecular level. Therefore, it is the first tine time when it was proposed the concept of the derivative of a molecular graph with respect to a given event, in order to obtain a new family of molecular descriptors. A new matrix representation of the molecular graph has been introduced by generalization of graph's theory's traditional incidence matrix. The new matrix is entitled as generalized incidence matrix and it arises from the Boolean representation of molecular sub- graphs that participate in the formation of the graph molecular skeleton. A very recent study of Martinez- Santiago *et al.* is presenting details about discrete derivatives for atom-pairs as a novel graph-theoretical invariant for generating new molecular descriptors [87].

4. CLASSIFICATION MODELS FOR MOLECULES

The previous section presented the calculation of different topological indices for molecular star graph, spiral graph and contact networks with specialized graph software such as S2SNet, CULSPIN and MInD-Prot. Therefore, it is possible to quantitatively characterize a molecule such as a drug, protein, and nucleic acid by a set of invariants based on the molecular topology (TIs) and physical and chemical properties of the molecular components such as atoms, amino acids and nucleobases. The biological properties of well characterized molecules from experiments and the TIs give the possibility to search for mathematical models that can predict specific biological properties for new molecules. Thus, these QSAR models are mathematical relationships between the TIs (molecular topology) and a molecular property.

In conclusion, Fig. (1) shows that the TIs can be used as indirect codification and quantification of biological properties, when it is impossible to know what molecular property can be used for a molecular classification. In addition, it is impossible to know *a priori* which type of graph will be better for a specific type of molecule classification or for a specific type of biological property. The published studies about these types of classifications demonstrated that some specific TIs of specific type of graphs generate better molecular classifications, without the comparison of all the possible types of graphs.

Because of the fact that the TIs codify hidden topological patterns mixed with physical-chemical amino acid weights, it is impossible to have a direct explanation of the reason of using specific TIs. In contrast, it is possible to observe that for a specific classification, a type of physical-chemical property such as Van der Waals interactions or polarity participates to a specific classification for a specific biological function.

There are a few criteria for choosing a specific type of graph:

1. The available input data - they depend on whether a sequence or a 3D structure of the macromolecule can be obtained. An example in this sense is the mass spectra experiments where the results are presented as sequences of peptides without any known function. Thus, the star, spiral and lattice types of graphs can be used.

2. The encoding information - the star, spiral and lattice types of graphs use the sequence information that includes node type, frequency of a node type and the neighborhood nodes in the sequence. In addition, the contact graphs use the 3D spatial information.

3. The graph structure - each type of graph defines the node connectivity using different criteria. The star, spiral and lattice graphs use the sequence of the node and additional rules such as node groups (star graph), neighborhood similar type of node (spiral graph) and a specific set of conditions $\mathbf{C}_q$ (lattice graph). The contact graphs use only the distance condition. Another detail: star graph can be structured in a maximum of 23 groups (amino acid types) for proteins and only 4 groups (nucleic base types) for nucleic acids; the lattice graph is constructed using only 4 directions and the spiral graph evolves in a spiral direction using only 4 classes of amino acids (proteins). Thus, star/spiral/lattice graphs are 2D representations of the information and the contact networks are 3D ones.

4. The abstract degree of the graph - the spiral and lattice graphs are the most abstract because they are linked to mathematical structures and pseudo structures. The star graph is less abstract because for proteins it represents an amino acid separation by type in each branch. The simplest type is the contact graph where we have nodes linked if they are close in space.

5. The physical-chemical property used as weight - star graphs do not use weights as default but they can use any type of no de property (see S2SNet); spiral graphs for proteins use 4 classes of amino acids that correspond to the side chain properties (see CULSPIN); lattice graphs use 4

classes of amino acids, similarly to the spiral graph; and the contact graphs can use any type of amino acid property. In MInD-Prot, the contact graphs use 4 amino acid properties including electronegativity, polarizability and Van der Waals area.

The type of TI to use for a specific type of graph can result from a variable selection method for a specific classification. One of the TIs that is linked with the information theory is the Shannon entropy and one of the most simple TIs are the spectral moments / traces because they are based only on the diagonal elements of the connectivity matrix.

4.1. QSAR Models

An important tool for the drug design and development is represented by the QSAR classification models based on molecular descriptors [88] that can help scientists to discriminate between drugs, proteins, nucleic acids or other type of molecules. The relations between the molecular structure and the molecular activity have been intensively used for diverse scientific problems. Some of the newest interests in QSAR applications are linked with anti- colorectal cancer agents [89, 90], anti-breast cancer agents [91], agents used in nervous system disorders [92], A(3) adenosine receptor antagonists [93], agrochemical fungicides [94], anti-viral drugs [95], tyrosine kinases inhibitors [96], and T-type calcium channel blocker [97]. All QSAR models have been generated using mathematical approaches such as statistics, Machine Learning or Artificial Intelligence. These tools can be simple such as a linear correlation between the molecular invariants and their activity or complex such as artificial neural networks.

4.2. Machine Learning for Classification

In order to respect the statistical independence condition, different classification techniques need to be tested using a 10-fold cross-validation to split data [98]. Dataset is randomly partitioned into 10 equal-sized bins: 9 bins were picked 10 times to train the models and the remaining bin is used to test them, each time leaving out a different bin. This tries to minimize influence of the configuration of training and validation sets.

The prediction model performance in the case of a two- class problem can be evaluated using the confusion matrix. There are several numbers of well-known accuracy measures for a two-class classifier in the literature such as classification rate, precision, sensitivity, specificity, F- measure and Area Under the Receiver Operating Characteristic Curve (AUROC) [99]. The higher the precision, the less effort wasted on testing and inspection; and the higher the recall, the fewer defective modules go undetected. However, there is a trade-off between precision and recall and therefore a combination of both is needed in a single efficiency measure, known as F-measure, which considers both precision and recall equally important [100].

The ROC is the Relative Operating Characteristic curve and it represents the comparison of two operating characteristics as the criterion changes: true positive rate and false positive rate [101]. ROC curve plot illustrates the performance of a binary classifier system as its discrimination threshold is varied. It plots the fraction of true positives out of the positives (TP Rate = true positive rate) vs the fraction of false positives out of the negatives (FP Rate = false positive rate), for different thresholds. TP Rate represents the sensitivity of the model, and FP Rate is (1 - specificity) and it is entitled true negative rate. Therefore, ROC represents a cost/benefit analysis.

The dataset of a QSAR model can be divided randomly into two parts (training and validation) extracting a total of 20% of the training data. Furthermore, a preprocessing of the data was performed to check the non-correlation between variables by means of the *findCorrelation* function in statistical software R [102] that searches through the correlation matrix columns to

remove the pair-wise correlations. A selection of variables can be done with several types of methods (genetic algorithm, ANNs, etc.).

There are a great number of applications that can build complex models for a specific data set. Some of them are Weka [79], STATISTICA [103], Matlab [104] and R [102]. Weka represents a collection of Machine Leaming algorithms for solving different data mining problems: AdaBoost (AB) [105], MultiLayer Perceptron (MLP) [106, 107], Naïve Bayes (NB) [108], Random Forest (RF) [109], LibLinear (LL) [110], J48 [100], and SVM [111].

STATISTICA has fewer methods compared to Weka because it is dedicated to statistics rather than to Machine Learning techniques, but it has some advantages: it dynamically combines the variables in order to search the best model and it can export the models as C files in order to be implemented in other user applications. In the next section some of the applications of QSAR models for specific problems are presented using topological indices of molecular graphs.

4.3. Applications of Graphs

The graph TIs have been used to build different models that can predict molecular properties such as the relation of proteins to cancers [112-115], human breast and colon cancer [116], prostate cancer [17], enzymatic activity [117], natural proteins [118], and drug toxicity [17].

4.3.1. Star Graph-Based Models

The star graphs have been used to solve several types of problems such as molecular classification for a specific property (enzymatic activity, anti-cancer agents, etc.) and the personalized diagnostics using blood proteome mas s spectra in cancers/toxicity and electroencephalography (EEG) in neurologic diseases.

A molecular property of proteins predicted with SO models is the *enzymatic activity* [117]. The important number of new proteins without any enzymatic characterization gives rise to a demand of protein QSAR theoretical models. This study presented a series of mixed protein parameters to obtain an enzyme/non-enzyme classification such as composition, sequence and connectivity, also called topological indices (TIs) and the computationally expensive 3D descriptors. The model was based on a set of 966 proteins (enzymes and non-enzymes) as PDB/DSSP files, Python/Biopython scripts, STATISTICA and Weka tools. Some of the indices are pure composition indices (residue fractions), DSSP secondary structure protein composition and 3D indices (surface and access), mixed indices such as composition-sequence indices (Chou's pseudoamino acid compositions or coupling numbers), 3D-composition (surface fractions) and DSSP secondary structure amino acid composition/propensities and classical TIs for the Randic's protein sequence Star graphs using S2SNet tool. The QSAR model can be developed using General Discriminant Analysis models (GDA), ANNs and other machine learning (ML) techniques. The results have been presented using complexity, average of Shannon's entropy (Sh) and data/method type. These results show that there is no direct relation between the complexity and the accuracy of the model for enzymes.

The character of random or natural protein have been predicted with a model based on SG TIs [118]. The study encode the information from the protein primary structure into Tis, the input for the natural/random protein classification mode1. The model was based on a set of 1,046 chains of natural proteins selected from the pre-compiled CulledPDB list from PISCES Dunbrack's Web Lab [119]. The protein homology was 20%, the structure resolution 1.6Å and the R-factor lower than 25%. The set of random amino acid chains contained 1,046 sequences generated wiht Python similar to the natural protein set. The model was found using the General Discriminant Analysis method [120] from STATISTICA. The best model was obtained with the forward stepwise model

with the accuracies of 90.77%. This model used for the first time the SG TIs to predict if a peptide is natural or random by using only the amino acid sequence information.

The relation of the proteins to the human breast and colon cancer can be predicted with a model built in Ref. [116]. The diagnostic of cancer is very complex because the specific markers can interfere and they can produce negative results. This is the reason there is a need of simple and fast theoretical models that can help the cancer diagnosis. This study converts the protein primary structure data in specific Randic's SG TIs using S2SNet application. The database of the QSAR model contained a set of 1,054 proteins related or not to two types of cancer, human breast cancer (HBC) and human colon cancer (HCC). The best input-coded multi- target classification model was obtained with the Discriminant Analysis Method with the accuracies of 90.0% (forward stepwise model type). This study demonstrate the useful of S2SNet in clinical proteomics.

The drug induces toxicity which has been predicted using the SG topological indices applied to numeric series (not direct to molecular structures) [17]. Similarly to QSAR, this study used a Quantitative Proteome-Property Relationship (QPPR) based on a SG theory to predict properties of polymeric complex systems. Thus, the Mass Spectrometry (MS) analysis of blood proteome (BP) is a very useful information source for the early detection of diseases and drug-induced toxicities. The spiral and star graph representation of the blood proteome MS was transformed into specific TIs such as spectral moments of the stochastic matrix associated with the spiral graph. They have been used to describe non-linear relationships between the different regions of the MS characteristic of BP. MARCH-INSIDE approach has been used to ca1culate the spectral moments for the SG for different BP samples and S2SNet to determine several SG TIs. The QPPR model has been obtained with Linear Discriminant Analysis (LDA) and it has been used to detect drug induced cardiac toxicities from BP samples. J48 decision tree classifier has the best performance among the other Machine Learning classification algorithms. The results show the ability of this approach to be applied into the polymer sciences. A similar approach has been used to detect prostate cancer using SG of the blood mass spectra [17].

4.3.2. Spiral Graph-Based Models

Spiral graphs have been less applied to classifications compared to SGs. An important application is the classification of proteins related to the human colon cancer [121]. This study proposed a new a Quantitative Structure- Disease Relationship (QSDR) classification model to predict proteins linked to human colon cancer using spiral graph TIs of protein amino acid sequences. The model uses eleven Shannon entropy indices, it was obtained with the Naive Bayes method and has an excellent predictive ability (90.92%) with AUROC of 0.91.

Other application of spiral graphs is the personalized diagnosis. This way, the information of mixtures of macromolecules such as the mass spectra but not the direct molecular graphs is used. Ref. [122] evaluated the drug- induced cardiotoxicity using blood proteome mass spectra. TIs have been used to build Quantitative Structure-Activity, Property or Toxicity Relationship (QSAR, QSPR and QSTR) models. The serum proteome Mass Spectra (MS) represents a potential information source for the early detection of biomarkers for diseases and/or drug-induced toxicities. In this work, the authors have been introduced for the first time a new graph representation for the blood proteome MS samples with the correspondent TIs: Spiral Markov Connectivity *(SMC_k)* of the MS Spiral graph ca1culated with the MARCH-INSIDE [123]. The SMC_k values have been used to find Quantitative Proteome-Property Relationship (QPPRs) models. TIs have been ca1culated for 62 blood samples and they have been used to find the best QPPR model that can discriminate between proteome MS for patients susceptible to suffer drug-induced cardiotoxicity from the control samples. The QPPR model is characterized by good Accuracy, Sensitivity, and Specificity (73.08% - 87.5%). The model demonstrated its power for clinical proteomics.

The drug toxicity predicted with SGs has been evaluated before with spiral graphs too [124]. Low range mas s spectra (MS) characterization of serum proteome offers the best chance of discovering proteome-(early drug-induced cardiac toxicity) relationships, called here Pro-EDICToRs. There are thousands of proteins involved and there is a difficult task to find a single disease-related protein. Therefore, the search for a model based on general MS patterns becomes a more realistic choice. Similarly to QSAR, the model was a Quantitative Proteome- Toxicity Relationship (QPTR) that links MS 3D-Markovian electronic delocalization entropies (3D-MEDNEs) [125] to drug-induced toxicological properties from BP information. 62 serum proteome samples have been transformed in SG and LG. Each sample has more than 370,100 intensity (I i) signals with m/z bandwidth above 700-12,000 each. The TIs have been calculated with MARCH-INSIDE too1. The best proposed QPTR has accuracy between 83.8% and 87.1 % and leave-one-out (LOO) predictive ability of 77.4-85.5%. This work demonstrated that the idea behind classic drug QSAR models may be extended to construct QPTRs with proteome MS data.

4.3.3. Lattice Graph-Based Models

Lattice graphs have been used to predict nucleic acid property such as mycobacterial DNA promoters [126]. The promoter sequences are important for the regulation of important mycobacterial pathogens. This study proposed two DNA promoter QSAR models based on pseudo-folding lattice network (LN) and SG TIs. The best model is based on two LN stochastic electrostatic potentials and it has Accuracy of 90.87%, Selectivity of 82.96% and Specificity of92.95%.

The QSAR model for alignment-free prediction of human breast cancer biomarkers was constructed similarly to the SG study and it was based on electrostatic potentials of protein pseudofolding HP-lattice networks [127]. This QSAR model was based on 122 proteins that related to human breast cancer (HBC) from experiments [128] from over 10,000 human proteins. The control group was made up of 200 proteins that are not related to HBC (non-HBCp). The ca1culated TIs are electrostatic potential parameters and the statistical method was the Linear Discriminant Analysis. The best model was validated with an external prediction series with good classification of 80%. The best QSAR model could predict genes and/or proteins linked to the HBC.

4.3.4. Contact Network-Based Models

The contact networks are used more frequently because they include complex information of the amino acids such as the 3D relative position combined with electrostatic properties. The next rows will present few applications of the contact networks.

The enzymatic activity of the proteins has been the target for a new classification model based on molecular graph/network [129]. This study presented a new and fast Markov chain model (MCM) able to predict the enzyme classification (EC) number. A comparison between linear and non-linear classifiers has been done using linear discriminant analysis (LDA) and/or artificial neural networks (ANN). The LDA model was based on three variables and predicted the first EC number with an overall accuracy of 79%. The data set contained 4,755 proteins (859 enzymes and 3,896 non-enzymes) divided into both training and external validation series. The ANN model a very good overall accuracy of 98.85%. The model was implemented at portal Bio-AIMS (http://bio-aims.udc.es/EnzClassPred.php) as a free on-line tool using PHP/HTML/Python and MARCH-INSIDE routines. This tool could be used to predict peptides of prokaryote and eukaryote parasites and their hosts as well as other superior organisms, for drug development.

The prediction of drug-protein interactions represents the main step in the drug design process in the Pharmaceutical companies. A classification model that uses the drug and protein target molecular topology and that can predict these interactions is presented in Ref. [130]. In this work, the authors selected drug-target pairs (DTPs/nDTPs) of drugs with high affinity/non-affinity for different targets. Generally, the QSAR models predict activity against only one protein target. In addition, there is no model implemented as a free Web server. Therefore, this study presented a multitarget QSAR (mt-QSAR) classifier using MARCH-INSIDE to calculate the molecular TIs of drusg and targets, and Linear Discriminant Analysis (LDA) method to find the best classification model. The LDA model had an accuracy of 94.4% for training and 94.9% for the external validation series and it was implemented into an online server entitled MARCH-INSIDE Nested Drug-Bank Exploration & Screening Tool (MIND-BEST, http://bio-aims.udc.esIMIND-BEST.php). In addition, two experiments have been done to verify the model. A similar model was created using non-linear methods for the same prediction of the PPIs in parasites [131].

The unique targets in trypanosome proteome have been predicted in Ref. [132] using protein-protein interactions (PPIs). *Trypanosoma brucei* causes important diseases such as African trypanosomiasis in humans (HA T or African sleeping sickness) and Nagana in cattle. In addition, *Trypanosoma cruzi* generates Chagas disease in South America, an acute illness in young children. The researchers have been tried to study the protein-protein interactions (PPIs) in pathogen Trypanosome species and they showed the low sequence identities between some parasite proteins and their human host, making the PPIs possible drug targets. Still there is no general models to predict Unique PPIs in Trypanosome (TPPIs). This work introduced new protein- protein complex invariants based on the Markov average electrostatic potential ξ_k(Ri) for amino acids located in different regions (Ri) of i^h protein and placed at a distance k one from each other. Over 30 different types of parameters have been ca1culated for 7,866 pairs of proteins: 1,023 TPPIs and 6,823 non- TPPIs, from more than 20 organisms, including parasites and human or cattle hosts. The best model is represented by a linear formula with a prediction above 90% of TPPIs and non-TPPIs using only two molecular descriptors as $^d\xi_k(s) = | \xi_k(sl) - \xi_k(s2)|$, the absolute difference between the ξ_k(si) values on the surface of the two proteins of the pairs. The nonlinear ANN models showed poorer results. The linear QSAR model was implemented as a free web server named TrypanoPPI (http://bio-aims.udc.es/TrypanoPPLphp) and it represents the first model that predicts how unique a protein-protein complex in Trypanosome proteome is with respect to other parasites and hosts.

Other similar protein interactions have been predicted using a model described in Ref. [133]. This work proposed a theoretical models to predict biologically relevant Parasite Self Proteins (PSP), which are expressed differentially in a given parasite and are dissimilar to proteins expressed in other parasites and have a high probability to become new vaccines (unique sequence) or drug targets (unique 3D structure). The model is working for PSPs in eight different HPs *(Ascaris, Entamoeba, Fasciola, Giardia, Leishmania, Plasmodium, Trypanosoma,* and *Toxoplasma)* with an accuracy of 90%. The model has inputs the protein residue graphs, and it was obtained with Artificial Neural Networks (ANN). The TIs are Markov spectral moments ca1culated with MARCH-INSIDE. The model was implemented as MISS-Prot (MARCH-INSIDE Scores for Self-Proteins) available at http://bio-aims.udc.es/MISSProt-HP.php. The new tool uses 3D structures deposited at PDB (mode 1) or Peptide Mass Fingerprinting (PMFs) and MS/MS for query proteins with unknown 3D structures (mode 2). In addition, MISS-Prot allows the prediction of PSP proteins in 16 additional species including parasite hosts, fungi pathogens, disease transmission vectors, and biotechnologically relevant organisms.

CONCLUSION

The graphical processing of the molecular information demonstrated a great capability to encode complex and hidden information. The diversity of graph representations and the flexibility to define any system as parts (nodes) and relations (edges) give the possibility to quantify complex information from macromolecules. The current review is presenting several basic concepts about graphical processing such as mathematical elements, types of graphs, different molecular descriptors for small molecules and macromolecules, the most common tools to calculate these molecular descriptors, and the most important applications on prediction of macromolecule properties or interactions.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work is supported by "Collaborative Project on Medical Informatics (CIMED)" PI13/00280 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 20l3- 2016 and the European Regional Development Funds (FEDER) and by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. GRC2014/049), the Galician Network for Colorectal Cancer Research (REGICC) (Ref. R2014/039) and by the Galician Network of Drugs R+D REGID (R2014/025) and the European Fund for Regional Development (FEDER) in the European Union.

REFERENCES

[1] Cepeda MS, Lobanov V, Berlin JA. From ClinicalTrials.gov trial registry to an analysis-ready database of clinical trial results. Clin Trials 2013; 10(2): 347·8.

[2] Harary F. Grapb Theory. Westview Press: MA 1969.

[3] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Model 1988; 28(1): 31-6.

[4] Guba R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL. The Blue Obelisk- interoperability in chemical informatics. J Chem Inf Model 2006; 46(3): 991-8.

[5] Balaban AT. Graph Theory and Topology in Chemistry. Elsevier: Amsterdam 1987.

[6] Bonchev D, Rouvray DH. Chemical Graph Theory. Gordon & Breach: New York 1991.

[7] Mason O, Verwoerd M. Graph theory and networks in Biology. IET Syst Biol 2007; 1(2): 89·119.

[8] Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. Social science. Computational social science. Science 2009; 323(5915): 721-3.

[9] Jones JJ, Settle JE, Bond RM, Fariss CJ, Marlow C, Fowler JH. Inferring tie strength from online directed behavior. PLoS ONE 2013; 8(1): e52168.

[10] Bolte J, Kemer J. Quantum graphs with two-particle contact interactions. J Phys A: Math Theor 2013; 46: 045207.

[11] Balaban AT. Chemical graphs. XXXIV. Five new topological indices for the branching of tree-like graphs. Theor Chim Acta 1979; 53: 355-75.

[12] Randic M, Zupan J, Vikic-Topic D. On representation of proteins by star-like graphs. J Mol Graph Model 2007: 290-305.

[13] Yook SH, Oltvai ZN, Barabasi AL. Functional and topological characterization of protein interaction networks. Proteomics 2004; 4(4): 928-42.

[14] Nilsson D, Andersson B. A graphical tool for parasite genome annotation. Comput Methods Programs Biomed 2004; 73(1): 55- 60.

[15] Concu R, Podda G, Ubeira FM, Gonzalez-Diaz H. Review of QSAR Models for Enzyme Classes of Drug Targets: Theoretical Background and Applications in Parasites, Hosts, and other Organisms. Curr Pharm Des 2010; 16(24): 2710-23.
[16] Altmann M. Reinterpreting network measures for models of disease transmission. Soc Networks 1993; 15(1): 1-17.
[17] Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, Chou KC, González-Díaz H. Stochastic molecular descriptors for polymers. 4. Study of complex mixtures with topological indices of mass spectra spiral and star networks: The blood proteome case. Polymer 2008; 49: 5575-87.
[18] Read RC, Wilson RJ. An Atlas of Graphs. Clarendon Press: 1998.
[19] Balaban AT. Topological indices based on topological distances in molecular graphs. Pure Appl Chem 1983; 55(2): 199-206.
[20] Bharucha-Reid AT. McGraw-Hill Series in Probability an Statistic. New York, McGraw-Hill Book Company 1960;167-434.
[21] Barigye SJ, Marrero-Ponce Y, Santiago OM, Lopez YM, Perez- Giménez F, Torrens F. Shannon's, mutual, conditional and joint entropy information indices: generalization of global indices defined from local vertex invariants. Curr Comput Aided Drug Des 2013; 9(2): 164-83.
[22] Todeschini R, Consonni V. Handbook of Molecular Descriptors. Wiley-VCH: 2002.
[23] Hahn HK, Schoenberger K. The ordered distribution of natural numbers on the square root spiral. 2007.
[24] González-Díaz H, Perez-Montoto LG, Duardo-Sanchez A, Paniagua E, Vázquez-Prieto S, Vilas R, Dea-Ayuela MA, Bolas- Fernández F, Munteanu CR, Dorado J, Costas J, Ubeira FM. Generalized lattice graphs for 2D-visualization of biological information. J Theor Biol 2009; 261(1): 136-47.
[25] Randic M, Balaban AT. On a four-dimensional representation of DNA primary sequences. J Chem Inf Comput Sci 2003; 43(2): 532- 9.
[26] Randic M. A Graph Theoretical Characterization of Proteomics Maps. Int J Quant Chem 2002 90: 848-58.
[27] Randic M. Graphical representations of DNA as 2-D map. Chem Phys Lett 2004; 386(4): 468-71.
[28] Berger B, Leighton T. Protein folding in the hydrophobic- hydropbilic (HP) model is NP-complete. J Comput Biol 1998; 5(1): 27-40.
[29] Gates MA. A simple way to look at DNA. J Theor Biol 1986; 119: 319-28.
[30] Randic M, Guo X, Basak SC. On the characterization of DNA primary sequences by triplet of nucleic acid bases. J Chem Inf Comput Sci 2001; 41(3): 619-26.
[31] Nandy A. Novel Method for Discrimination of Conserved Genes through Numerical Characterization of DNA Sequences. Int E J Mol Design 2003; 2: 000-000.
[32] Roy A, Raycbaudbur C, Nandy A. Novel techniques of graphical representation and analysis of DNA sequences-A review. J Biosci 1998; 23(1): 55-71.
[33] 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.
[34] Agüero-Chapin G, González-Díaz 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 letters 2006; 580(3): 723-30.
[35] González-Díaz H, Pérez-Bello A, Cruz-Monteagudo 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 Systs 2007; 85: 20-6.
[36] Yuan Z. Prediction of protein subcellular locations using Markov chain models. FEBS letters 1999; 451(1): 23-6.
[37] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank Nucleic Acids Res 2000; 28: 235-42.
[38] Munteanu CR, González-Díaz H. S2SNet - Sequence to Star Network, Reg. No. 03 / 2008 / 1338, Santiago de Compostela, Spain. Santiago de Compostela, Spain 2008.
[39] Pérez Montoto LG, Prado-Prado FJ, Munteanu CR, González Díaz H. CULSPIN - Compute ULam SPiral INdices, Register No.: 03/2009/1199 (SC-207-09), Spain. Santiago de Compostela 2009.
[40] Munteanu CR, González-Díaz H. MInD-Prot - Markov Indices for Drugs and Proteins, Register No.: 03/2012/1 051 (SC-228-l2). Santiago de Compostela, Spain 2012.
[41] Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (COK): an open- source lava library for Chemo- and Bioinformatics. J Chem Inf Comput Sci 2003; 43(2): 493-500.
[42] Guha R. Chemical Informatics Functionality in R. J Stat Software 2007; 18: 1-16.
[43] Spjuth O, Helmus T, Willighagen EL, *et al.* Bioclipse: an open source workbench for chemo- and bioinformatics. BMC Bioinformatics 2007; 8: 59.

[44] Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C. KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics 2013; 14: 257.
[45] Lawson KR, Lawson J. LICSS - a chemica1 spreadsheet in microsoft excel. J Cheminform 2012; 4(1): 3.
[46] Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. MATCH, communications in mathematical and in computer chemistry 2006; 56: 237-48.
[47] Tetko IV, Gasteiger J, Todeschini R, *et al.* Virtual computational chemistry laboratory - design and description. J Comput Aided Mol Des 2005; 19: 453-63.
[48] Ponce YM. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. Bioorg Med Chem 2004; 12(24): 6351-69.
[49] Casanola-Martin GM, Marrero-Ponce Y, Khan MT, Ather A, Khan KM, Torrens F, Rotondo R. Dragon method for finding novel tyrosinase inhibitors: Biosilico identification and experimental *in vitro* assays. Eur J Med Chem 2007; 42(11-12): 1370-81.
[50] Pérez-Garrido A, Helguera AM, Rodríguez FG, Cordeiro MN. QSAR models to predict mutagenicity of acrylates, methacrylates and alpha,beta-unsaturated carbonyl compounds. Dent Mater 26(5): 397-415.
[51] Estrada E, Quincoces JA, Patlewicz G. Creating mo1ecular diversity from antioxidants in Brazilian propolis. Combination of TOPS-MODE QSAR and virtual structure generation. Mol Divers 2004; 8(1): 21-33.
[52] Cabrera-Pérez MA, Bermejo-Sanz M, Ramos-Torres L, Grau- Ávalos R, Pérez-González M, González-Díaz H. A topological sub- structural approach for predicting human intestinal absorption of drugs. Eur J Med Chem 2004; 39: 905-16.
[53] Molina-Ruiz R, Saiz-Urra L, Rodriguez-Borges JE, *et al.* A TOPological Sub-structural Molecular Design (TOPS-MODE)- QSAR approach for modeling the antiproliferative activity against murine leukemia tumor cellline (L1210). Bioorg Med Chem 2009; 17(2): 537-47.
[54] Casanola-Martin GM, Marrero-Ponce Y, Tareq Hassan Khan M, Torrens F, Pérez-Giménez F, Rescigno A. Atom- and bond-based 2D TOMOCOMD-CARDD approach and ligand-based virtual screening for the drug discovery of new tyrosinase inhibitors. J Biomol Screen 2008; 13(10): 1014-24.
[55] González-Díaz H, Romaris F, Duardo-Sánchez A, Pérez-Montoto LG, Prado-Prado F, Patlewicz G, Ubeira FM. Predicting drugs and proteins in parasite infections with topological indices of complex networks: theoretica1 backgrounds, applications, and legal issues. Curr Pharm Des 2010; 16(24): 2737-64.
[56] Casanola-Martin GM, Khan MT, Marrero-Ponce Y, Ather A, Sultankhodzhaev MN, Torrens F. New tyrosinase inhibitors selected by atomic linear indices-based classification models. Bioorg Med Chem Lett 2006; 16(2): 324-30.
[57] Castillo-Garit JA, Marrero-Ponce Y, Escobar J, Torrens F, Rotondo R. A novel approach to predict aquatic toxicity from molecular structure. Chemosphere 2008; 73(3): 415-27.
[58] Castillo-Garit JA, Marrero-Ponce Y, Torrens F, Garcia-Domenech R. Estimation of ADME properties in drug discovery: predicting Caco-2 cell permeability using atom-based stochastic and non- stochastic linear indices. J Pharm Sci 2008; 97(5): 1946-76.
[59] Ponce YM, Marrero RM, Castro EA, Ramos de Armas R, Diaz HG, Zaldivar VR, Torrens F. Protein quadratic indices of the "macromolecular pseudograph's alpha-carbon atom adjacency matrix". 1. Prediction of Are repressor alanine-mutant's stability. Molecules 2004; 9(12): 1124-47.
[60] Marrero-Ponce Y, Medina-Marrero R, Castro AE, Ramos de Armas R, González-Díaz H, Romero-Zaldivar V, Torrens F. Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. Molecules 2004; 9: 1124-47.
[61] Marrero-Ponce Y, Nodarse O, González-Díaz H, Ramos de Armas R, Romero-Zaldivar V, Torrens F, Castro EA. Nucleic Acid Quadratic Indices of the "Macromolecular Graph's Nucleotides Adjacency Matrix". Modeling of Footprints after the Interaction of Paromomycin with the HIV-l Ψ-RNA Packaging Region. Int J Mol Sci 2004; 5: 276-93.
[62] Kier LB, Hall LH. Molecular Structure Description: The Electrotopological State. Academic Press: 1999.
[63] Katritzky AR, Oliferenko A, Lomaka A, Karelson M. Six- membered cyclic ureas as HIV-1 protease inhibitors: a QSAR study based on CODESSA PRO approach. Quantitative structure-activity relationships. Bioorg Med Chem Lett 2002; 12(23): 3453-7.
[64] Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M. Antimalarial activity: a QSAR modeling using CODESSA PRO software. Bioorg Med Chem 2006; 14(7): 2333-57.

[65] Katritzky AR, Dobchev DA, Tulp I, Karelson M, Carlson DA. QSAR study of mosquito repellents using Codessa Pro. Bioorg Med Chem Lett 2006; 16(8): 2306-11.
[66] Katritzky-AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M. Antimalarial activity: a QSAR modeling using CODESSA PRO software. Bioorganic & medicinal chemistry 2006; 14(7): 2333-57.
[67] Rappin N, Dunn R. wxPython in Action. Manning Publications Co.: Greenwich, CT 2006.
[68] González-Díaz H, Molina-Ruiz R, Hernández I. MARCH-INSIDE version 3.0 *(MARkov CHains INvarianls for Slmulalion & DEsign);* Windows supported version under request to the main author contact email: gonzalezdiazh@yahoo.es. 3.0 ed 2007.
[69] Koschützki D. CentiBiN Version 1.4.2. 2006:CentiBiN Version 1.4.2, Centralities in Biological Networks © 2004-6 Dirk Koschützki Research Group Network Analysis, IPK Gatersleben, Germany.
[70] Batagelj V, Mrvar A. Pajek, Program for Large Network Analysis (ver. 1.15), http://vlado.fmf.uni-lj.silpub/networks/pajek/. 1.15 ed 2006.
[71] González-Díaz H, Duardo-Sánchez A, Ubeira FM, *et al.* Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 2010; 11(4): 379-406.
[72] González-Díaz H, Prado-Prado F, Ubeira FM. Predicting antimicrobial drugs and targets with the MARCH-INSIDE approach. Curr Top Med Chem 2008; 8(18): 1676-90.
[73] Hou TJ, Xu XJ, ADME evaluation in drug discovery. 2. Prediction of partition coefficient by atom-additive approach based on atom- weighted solvent accessible surface areas. J Chem Inf Comput Sci 2003; 43(3): 1058-67.
[74] 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- 8.
[75] González-Díaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sánchez-González A. Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. J Comput Chem 2007; 28(6): 1042-8.
[76] González-Díaz H, Molina R, Uriarte E. Recognition of stable protein mutants with 3D stochastic average electrostatic potentials. FEBS Lett 2005; 579(20): 4297-301.
[77] Concu R, Podda G, Uriarte E, González-Díaz H. Computational chemistry study of 3D-structure-function relationships for enzymes based on Markov models for protein electrostatic, HINT, and van der Waals potentials. J Comput Chem 2009; 30: 1510-20.
[78] 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: 1990-5.
[79] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IA. The WEKA Data Mining Software: An Update. SIOKDD Explorations 2009; 11(1).
[80] Overington J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL- EBI). Interview by Wendy A. Warr. J Comput Aided Mol Des 2009; 23(4): 195-8.
[81] Gaulton A, Bellis LJ, Bento AP, Chambers J, *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2011; 40(Database issue): DI 100-7.
[82] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison OR. Open Babel: An open chemical toolbox. J Cheminform 2011; 3: 33.
[83] Barigye SJ, Marrero-Ponce Y, Pérez-Giménez F, Bonchev D. Trends in information theory-based chemical structure codification. Mol Divers 2013.
[84] Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Wiley-VCH: Weinheim 2009.
[85] Barigye SJ, Marrero-Ponce Y, Martínez López Y, *et al.* Event- based criteria in OT -ST AF information indices: theory, exploratory diversity analysis and QSPR applications. SAR QSAR Environ Res 2013; 24(1): 3-34.
[86] Marrero-Ponce Y, Santiago OM, López YM, Barigye SJ, Torrens F. Derivatives in discrete mathematics: a novel graph-theoretical invariant for generating new 2/3D molecular descriptors. I. Theory and QSPR application. J Comput Aided Mol Des 2012; 26(11): 1229-46.
[87] Martínez-Santiago O, Millán-Cabrera R, Marrero-Ponce Y, *et al.* Discrete Derivatives for Atom-Pairs as a Novel Graph-Theoretical lnvariant for Generating New Molecular Descriptors: Orthogonality, Interpretation and QSARs/QSPRs on Benchmark Databases. Molecular Informatics 2014; 33(5): 343-68.
[88] Munteanu CR, Fernández-Blanco E, Seoane JA, *et al.* Drug Discovery and Design for Complex Diseases through QSAR Computational Methods. Curr Pharm Des 2010; 16(24): 2640-55.

[89] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in *silico* discovery of anti-colorectal cancer agents. Bioorg Med Chem 2012; 20(15): 4848-55.
[90] Yerma RP, Hansch C. QSAR modeling of taxane analogues against colon cancer. Eur J Med Chem 2010; 45(4): 1470-7.
[91] Speck-Planche A, Kleandrova VV, Luan F, Cordeiro MN. Chemoinformatics in anti-cancer chemotherapy: multi-target QSAR model for the in *silico* discovery of anti-breast cancer agents. Eur J Pharm Sci 2012; 47(1): 273-9.
[92] Prado-Prado F, García-Mera X. QSAR models for computer-aided drug design and molecular docking for disorders of the central nervous system and other diseases. Curr Top Med Chem 2012; 12(16): 1731-3.
[93] Luan F, Borges F, Cordeiro MN. Recent advances on A(3) adenosine receptor antagonists by QSAR tools. Curr Top Med Chem 2012; 12(8): 878-94.
[94] Speck-Planche A, Kleandrova VV, Rojas-Vargas JA. QSAR model toward the rational design of new agrochemical fungicides with a defined resistance risk using substructural descriptors. Mol Diversity 2011; 15(4): 901-9.
[95] Prado-Prado FJ, García I, García-Mera X, González-Díaz H. Entropy multi-target QSAR model for prediction of antiviral drug complex networks. Chemom Intell Lab Syst 2011; 107(2): 227-33.
[96] Marzaro a, Chilin A, Guiotto A, Uriarte E, Brun P, Castagliuolo 1, Tonus F, González-Díaz H. Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors. Eur J Med Chem 2011; 46(6): 2185-92.
[97] Jeong JA, Cho H, Jung SY, Kang HB, Park JY, Kim J, Choo D, J., Lee JY. 3D QSAR studies on 3,4-dihydroquinazolines as T-type calcium channel blocker by comparative molecular similarity indices analysis (CoMSIA). Bioorg Med Chem Lett 2010; 20(1): 38-41.
[98] McLachlan OJ, Do K-A, Ambroise C. Analyzing microarray gene expression data. Wiley: 2004.
[99] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattem Recogn Lett 2009; 30(1): 27-38.
[100] Witten I, Frank E. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann: 2005.
[101] Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics : collected papers. Lawrence Erlbaum Associates: Mahwah, NJ 1996.
[102] Teetor P. R Cookbook. O'Reilly: 2011.
[103] StatSoft.lnc. STATISTICA, (data analysis software system), version 6.0, www.statsoft.com. 6.0 ed 2002.
[104] Caballero J, Fernández M. Artificial neural networks from MATLAB in medicinal chemistry. Bayesian-regularized genetic neural networks (BRONN): application to the prediction of the antagonistic activity against human platelet thrombin receptor (PAR-I). Curr Top Med Chem 2008; 8(18): 1580-605.
[105] Liu H, Setiono R. A probabilistic approach to feature selection - A filter solution. 13th Intemational Conference on Machine Learning; 1996; Bari, ltaly. Year; pp. 319-27.
[106] Bishop CM. Neural Networks for Pattem Recognition. Oxford University Press, USA: 1995.
[107] Bishop CM. Pattern recognition and machine learning. Springer: 2006.
[108] John OH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. 11 th Conference on Uncertainty in Artificial Intelligence; 1995 August 18-20, 1995; Montreal, Quebec. City: Morgan Kaufman Year; pp. 338-45.
[109] Breiman L. Random Forest. Machine Learning 2001; 45: 5-32.
[110] Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res 2008; 9: 1871-4.
[111] Vapnik VN. Statistical learning theory. Wiley: New York ; Chichester 1998.
[112] González-Díaz H, Ferino a, Munteanu CR, Vilar S, Uriarte E. In: Oncoproteomics. Cho WK, Ed. Springer 2009.
[113] Vázquez JM, Aguiar V, Seoane JA, *et al.* Star graphs of protein sequences and proteome mass spectra in cancer prediction. Curr Proteomics 2009; 6(4): 275-88.
[114] González-Díaz H, Ferino a, Prado-Prado FJ, Vilar S, Uriarte E, Pazos A, Munteanu CR. In: An Omics Perspective on Cancer Research. Cho WCS, Ed. Springer 2010.
[115] Martínez-Romero M, Vázquez-Naya JM, Rabuñal JR, *et al.* Artificial intelligence techniques for colorectal cancer drug metabolism: ontologies and complex networks. Curr Drug Metab 2010; 11(4): 347-68.
[116] Munteanu CR, Magalhaes AL, Uriarte E, Gonzalez-Diaz H. Multi- target QPDR classification model for human breast and colon cancer-related proteins using star graph topological indices. J Theor Bio12009; 257(2): 303-11.

[117] Munteanu CR, González-Díaz H, Magalhaes AL. Enzymes/non- enzymes classification model complexity based on composition, sequence, 3D and topological indices. J Theor Biol 2008; 254(2): 476-82.
[118] Munteanu CR, González-Díaz H, Borges F, de Magalhaes AL. Natural/random protein classification models based on star network topological in dices. J Theor Biol 2008; 254(4): 775-83.
[119] Wang ORL, Dunbrack J. PISCES: a protein sequence culling server. Bioinformatics (Oxford, England) 2003; 19: 1589-91.
[120] Van Waterbeemd H. In: Chemometric methods in molecular design. Van Waterbeemd H, Ed. New York, Wiley-VCH 1995; 265-82.
[121] Aguiar-Pulido V, Munteanu CR, Seoane JA, Fernández-Blanco E, Pérez-Montoto LG, González-Díaz H, Dorado J. Naive Bayes QSDR classification based on spiral-graph Shannon entropies for protein biomarkers in human colon cancer. Mol BioSyst 2012; 8(6): 1716-22.
[122] Cruz-Monteagudo M, Munteanu CR, Borges F, Cordeiro MN, Uriarte E, González-Díaz H. Quantitative Proteome-Property Relationships (QPPRs). Part 1: finding biomarkers of organic drugs with mean Markov connectivity indices of spiral networks of blood mass spectra. Bioorg Med Chem 2008; 16(22): 9684-93.
[123] González-Díaz H, Torres-Gomez LA, Guevara Y, Almeida MS, Molina R, Castanedo N, Santana L, Uriarte E. Markovian chemicals *"in silico"* design (MARCH-INSIDE), a promising approach for computer-aided molecular design III: 2.5D indices for the discovery of antibacterials. J Mol Mode12005; 11(2): 116-23.
[124] Cruz-Monteagudo M, González-Díaz H, Borges F, Domínguez ER, Cordeiro MN. 30-MEONEs: an alternative *"in silico"* technique for chemical research in toxicology. 2. quantitative proteome-toxicity relationships (QPTR) based on mas s spectrum spiral entropy. Chem Res Toxicol2008; 21(3): 619-32.
[125] González-Díaz H, Marrero Y, Hernández I, Bastida I, Tenorio E, Nasco O, Uriarte E, Castanedo N, Cabrera MA, Aguila E, Marrero O, Morales A, Pérez M. 3D-MEONEs: an alternative *"in silico"* technique for chemical research in toxicology. 1. prediction of chemically induced agranulocytosis. *Chem Res Toxicol 2003;* 16(10): 1318-27.
[126] Pérez-Bello A, Munteanu CR, Ubeira FM, De Magalhaes AL, Uriarte E, González-Díaz H. Alignment-free prediction of mycobacterial ONA promoters based on pseudo-folding lattice network or star-graph topological indices. J Theor Biol 2009; 256(3): 458-66.
[127] Vilar S, González-Díaz H, Santana L, Uriarte E. QSAR model for alignment-free prediction of human breast cancer biomarkers based on electrostatic potentials of protein pseudofolding HP-lattice networks. J Comput Chem 2008; 29(16): 2613-22.
[128] Sjoblom T, Jones S, Wood LO, Parsons OW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SO, Willis r, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parrnigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. The consensus coding sequences of human breast and colorectal cancers. Science 2006; 314(5797): 268-74.
[129] Concu R, Dea-Ayuela MA, Pérez-Montoto LG, *et al.* 3D entropy and moments prediction of enzyme classes and experimental- theoretic study of peptide fingerprints in Leishmania parasites. Biochim Biophys Acta 2009; 1794(12): 1784-94.
[130] González-Díaz H, Prado-Prado FJ, García-Mera X, Alonso N, Abeijón P, Caamano O, Yáñez M, Munteanu CR, Pazos Sierra A, Dea-Ayuela MA, Gómez-Muñoz MT, Garijo MM, Sansano J, Ubeira FM. MIND-BEST: web server for drugs & target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretic- experimental study of G3PD protein from Trichomona gallineae. J Proteome Res 2010; 10(4): 1698-718.
[131] González-Díaz H, Prado-Prado F, Sobarzo-Sánchez E, Haddad M, Chevalley SM, Valentin A, Quetin-Leclercq J, Dea-Ayuela MA, Teresa Gómez-Munos M, Munteanu CR, José Torres-Labandeira J, García-Mera X, Tapia RA, Ubeira FM. NL MINO-BEST: A web server for ligands and proteins discovery- Theoretic-experimental study of proteins of Giardia lamblia and new compounds active against Plasmodium falciparum. J Theor Biol 2011; 276(1): 229- 49
[132] Rodríguez-Soca Y, Munteanu CR, Dorado J, Pazos A, Prado-Prado FJ, González-Díaz H. Trypano-PPI: a web server for prediction of unique targets in trypanosome proteome by using electrostatic parameters of protein-protein interactions. J Proteome Res 2010; 9(2): 1182-90.
[133] González-Díaz H, Muino L, Anadon AM, Romaris F, Prado-Prado FJ, Munteanu CR, Dorado J, Sierra AP, Mezo M, González- Warleta M, Garate T, Ubeira FM. MISS-Prot: web server for self/non-self discrimination of protein residue networks in parasites; theory and experiments in Fasciola peptides and Anisakis allergens. Mol BioSyst 2011; 7(6): 1938-55.