



Transworld Research Network
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Topological Indices for Medicinal Chemistry, Biology, Parasitology, Neurological and
Social Networks, 2010: 145-161 ISBN: 978-81-7895-489-9
Editor: Humberto González-Díaz and Cristian Robert Munteanu

8. QSPR models for human Rhinovirus surface networks

Santiago Vilar¹ and Humberto González-Díaz²

¹*National Institutes of Health, DHHS, Bethesda, Maryland 20892, USA*

²*Department of Microbiology & Parasitology, Faculty of Pharmacy
University of Santiago de Compostela, Santiago de Compostela, 15782, Spain*

1. Introduction

Quantitative Structure-Property/Activity Relationship (QSPR/QSAR) [1] techniques based in different indices have a wide variety of applications in bioorganic chemistry research to connect the chemical structure of small-sized molecules with antiviral activity of drugs [2]. The interest in the application of QSAR has steadily increased in recent decades and in a very recent work Verma and Hansch pointed out that it may be useful in the search for anti-HRV (Human Rhinoviruses) agents [3]. In this paper, they have discussed QSAR models to predict new anti-HRV drugs taking into consideration the chemical structure of the drug candidate. In our opinion, QSAR approaches can be interesting for the computational study of not only small molecules but large biopolymers or not-molecular biological systems. It opens, for instance, the possibility for a QSAR view of the HRVs problem from the side of the virus instead of the drug side, which can be seen as the second part of Verma and Hansch work.

Correspondence/Reprint request: Dr. González-Díaz H, Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Santiago de Compostela (USC), Santiago de Compostela, 15782, Spain
E-mail: humberto.gonzalez@usc.es or gonzalezdiazh@gmail.com

In fact, the knowledge about the 3-D structure including the surface of proteins increases our understanding of its function and interaction with other proteins. However, this knowledge of the sequence and its 3-D structure does not provide a clear relationship with its biological properties. For this reason, is important to search for novel protein 3D indices derived from new protein molecular graphics representations useful to seek QSAR models able predict the biological functions of proteins. Some 3D molecular descriptors used to codify molecular structures of polymers include Arteca's mean crossing-over number, the Flory radius of gyration and the I3 index amongst others [4].

In addition, the creation of databases of DNA/RNA and protein sequences without 3D structure determined has led to significant developments in this search of molecular graphics of 2D network/graph type based methodologies that characterize DNA and protein sequences. Molecular graphics representations for DNA or RNA sequences have been reported by different authors [5, 6]. In the case of proteins, Hydrophobicity-Polarity Lattices (HP-Lattices) are one of the more used types of molecular graphics to model protein structure-properties relationships and folding dynamics in 2D or 3D spaces (pseudo-folding) [7, 8]. A new protein pseudo-folding molecular graphics or network type representation have been introduced by Fernández and Caballero *et al.* recently [9].

In any case, using different type of numerical indices derived from these proteins or DNA/RNA 2D molecular graphics to perform QSAR studies is simpler than when we need to know 3D structure. These indices describe graph/network topology, connectivity, or branching and are often referred to as the graph Topological Indices (TIs) or Network Connectivity Indices (CIs). Often, CIs/TIs are enough efficient to codify important amounts of information in a timely way with respect to 3D indices [10]. The uses of these indices to seek Structure-Function relationships in Cellular Biochemistry are diverse [11]. In two recent reviews, we revised in-depth the applications in Theoretical Biology and Bioinformatics [12] and CIs/TIs derived from network/graph type molecular graphics of small-sized molecules, macromolecules, and more complex sources of information including whole Proteome Mass Spectrums, Genomes, or Protein Interaction Networks [13].

Recently, the MARCH-INSIDE approach introduced by our group has been generalised to encode structural features of DNA/RNA and proteins. In these works we included 2D-RNA secondary structure graphs, pseudo-3D proteins molecular graphics, and other types of graphics or networks, [14] including the HP-Lattice type of complex networks [15].

The study described here concerns to the protein sequences present in the capsid of HRV. In each case the sequence of viral proteins of 19 strains of HRV is represented by an HP-Lattice network. Later, we calculated by the

first time up to 11 different classes of TIs for these HP-Lattice networks. The TIs include total indices of the full network and local indices for specific groups of aminoacids. Next, we used these TIs as input parameters of a Linear Discriminant Analysis (LDA) in order to construct 11 different QSAR models. These QSAR models are discriminant functions that may classify a HRV as a major or minor group virus. The mechanisms are the above referred ICAM-1-mediated and the LDL-mediated mechanism. We compared the QSAR models based on the different types of TIs. We also compared these models with 3D Topographic Indices (TGIs) derived from virus surface road maps. These are graphs equivalent to Viral SCN or the same Complex Networks of aminoacid vicinity at the viral surface. Some of the models QSAR based on TGIs of SCN were previously published and other were developed in these work in order to make a more rigorous comparative study [13, 16, 17].

2. Materials and methods

2.1. Statistical analysis

Once the different descriptors had been calculated for all the Human Rhinoviruses (HRVs) in the databases we proceed with statistical analysis. For it, the variables were standardised and a Linear Discriminant Analysis (LDA) was carried out using the STATISTICA 6.0 software package [18] to develop three classification functions that are capable of differentiating between different groups. The formula for the LDA classification function is:

$$S = a_0 + \sum_{c,k,g}^{3,m,21} b_{kcg} \cdot {}^k TI_c(g) \quad (1)$$

The variable S is a real value score for the biological property under investigation: the tendency of the virus to bound the LDLR receptor instead of ICAM. The values a_k and b_{kcg} are the coefficients obtained for the LDA classification functions (QSAR model) [19, 20]. The statistical quality of the models was assessed using parameters such as Wilks' statistic (λ), the Fisher ratio (F), the square of the Mahalanobis distance (D^2) and the percentage of good classification for the training set as well as for the cross-validation procedure. The classification of cases was carried out by considering the subsequent classification probabilities, which are the probabilities that the respective case belongs to a particular group, i.e., active or inactive. The different discriminant functions were obtained using the forward-stepwise

method with a minimum tolerance of 0.01 [19, 20]. As an alternative to TIs in we also explored TGIs using the same notation:

$$S = a_0 + \sum_{c,k,g}^{3,m,21} b_{kcg} \cdot {}^k TGI_c(g) \quad (2)$$

3. Results and discussion

3.1. QSAR models based on TIs of 2D-lattice surface complex networks

Human Rhinoviruses (HRVs) are the single most important cause of common colds. We used the same HRV series recently used by Vlasak *et al.*[21] a total of 19 HRVs were studied: 10 belonging to the minor group and the other 9 to the major group. The widespread nature of this affliction, the economic consequences, and the well-known impracticality of vaccine development due to the large number of HRV serotypes (>100) have justified the search for antiviral chemotherapeutic agents. Rhinoviruses belong to the *Picornaviridae* family and represent a type of RNA virus with a small size. HRVs are naturally occurring polymers-composed systems present in the form of small icosahedral particles (~30 nm) composed of 60 copies of viral capsid proteins VP1, 2, 3 and 4 and a positive-strand (messenger sense) RNA. In terms of mechanism of cellular infection they may be classified into two groups: the major group (viruses binding intracellular adhesion molecule 1, ICAM-1) and the minor group (viruses binding low-density lipoprotein receptor, LDLR). The structure of the VP surface has been found to be involved in the receptor specificity of HRVs. In this sense, the prediction of the mechanism of infection of new viral mutants with theoretical models becomes of the major importance to assist bioorganic and medicinal chemists on the design new effective drugs [22].

However, in many cases knowing a DNA/RNA or protein sequence does not provide information about relationships with biological properties. For this reason it is necessary to codify sequence information by constructing 2D graphical representations [23] or other representations for DNA[24]; Liao's RNA graphs [25] or HP-Lattice proteins network representations [26] through which it is possible to calculate TIs with the aim of finding relationships between a protein structure and its biological activity. Liao and other authors have successfully used Lattice like representations of DNA viral sequences to model biological properties [27].

Randic *et al.* [28] demonstrated that 2D-Lattice representations are the result of projecting in the 2D plane the view of a 3D-Tetrahedral

representation of the DNA sequence. The results also apply to HP-Lattice representations of proteins of course; we only have to substitute the four classes of nucleotides by the four groups of aminoacids grouped according to polarity and hydrophobicity. In this 2D-Lattice projection, we draw only the nodes in the 3D- space with tetrahedral coordinates that are closer to the plane (more external nodes from this plane view with respect to the center) and all the rest of nodes with the same coordinates in the projecting axis have to be overlapped in the same node. In this sense, we interpret here 2D-Lattices as the projection of the surface of the protein represented in the 3D-Tetrahedral coordinates. By this reason, we call HP-Lattice here as 2D-Lattice-Surface Complex Networks (2DL-SCN) in opposition to real 3D-SCN. In consequence, 2DL-SCN are Pseudo-folding graph representations of protein whereas 3D-SCN are representations of realistic 3D protein Folding.

Initially, the sequences of the different HRVs were introduced into the programme MARCH-INSIDE 2.0[29] and this was used to generate one HP-Lattice network representation per sequence. Specifically in this work, we interpreted HP-Lattice Networks as 2D-Lattice Surface Complex Networks (2DL-SCN); see results and discussion section. In the **Figure 1** we superposed the 2DL-SCN of all HRV strains to illustrate the similarities and dissimilarities between the two groups and how difficult may be to discriminate between them. The aim of the method proposed here is to overcome the 10D-amino acid space bottleneck by grouping the twenty natural amino acids into only four groups:

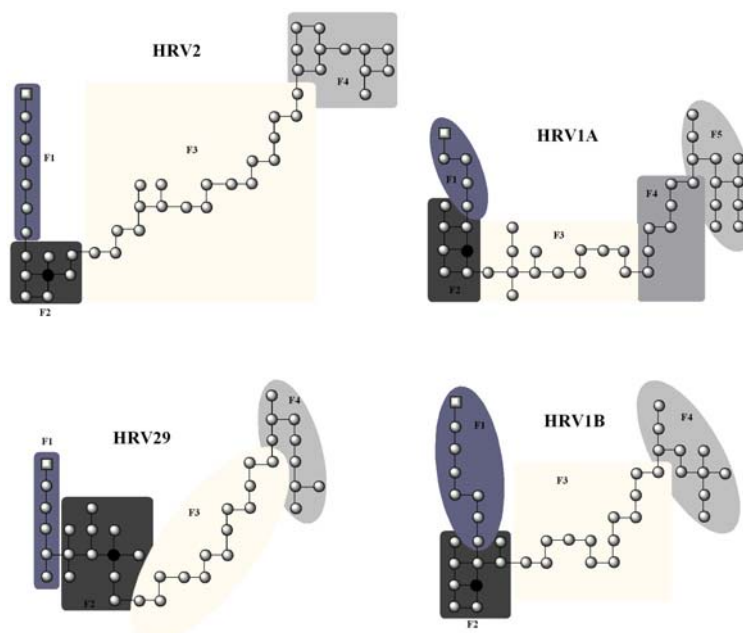


Figure 1. 2DL-SCN for HRVs 2, 1A, 29 and 1B and BPA of the different fragments (F1-F5).

1. The coordinates of abscissa axis increases in +1 for an acidic amino acid (rightwards-step) or:
2. The coordinates of abscissa axis decreases in -1 for a basic amino acid (leftwards-step) or:
3. The coordinates of ordinate axis increases in +1 for a polar amino acid (upwards-step) or:
4. The coordinates of ordinate axis decreases in -1 for a non-polar amino acid (downwards-step).

Table 1. Names, symbols, formula, and network type for TIs and/or TGIs in this work.

Name	notation ^a	Classic Symbol & Formula ^b
Entropies	${}^k\text{TI}_\Theta(\mathbf{g})$	$\Theta_k(\mathbf{g}) = -\sum_j^{\xi} p_k(j) \cdot \log p_k(j)$
Spectral Moments	${}^k\text{TI}_\pi(\mathbf{g})$	$\pi_k(\mathbf{g}) = \sum_j^{\xi} p_{jj}^k$
Electrostatic Potentials	${}^k\text{TI}_\xi(\mathbf{g})$	$\xi_k(\mathbf{g}) = \sum_j^{\xi} p_k(j) \cdot \left(\frac{q_j}{d_{j0}}\right)$
Balaban Index	${}^0\text{TGI}_J(\mathbf{g})$	$J(\mathbf{g}) = \frac{1}{2} \cdot C \cdot (\mathbf{d}^t \cdot \mathbf{A} \cdot \mathbf{d}^{tT})_{\mathbf{g}}$
Wiener Index	${}^0\text{TI}_W(\mathbf{g})$	$W(\mathbf{g}) = \frac{1}{2} (\mathbf{u} \cdot \mathbf{D} \cdot \mathbf{u}^T)_{\mathbf{g}}$
Mol. Topol. Index	${}^0\text{TGI}_{MTI}(\mathbf{g})$	$MTI(\mathbf{g}) = \sum_{i=1}^n [(\mathbf{A} + \mathbf{D})\mathbf{v}]_{\mathbf{g}}$
Randic Index	${}^0\text{TI}_\chi(\mathbf{g})$	$\chi(\mathbf{g}) = \left(\prod_{i=1}^n \text{deg}(j)\right)_{\mathbf{g}}^{-1/2}$
Sum of Degrees	${}^0\text{TI}_\delta(\mathbf{g})$	$SoD(\mathbf{g}) = \sum_j^n \text{deg}(j)_{\mathbf{g}}$
Diameter	${}^0\text{TI}_D(\mathbf{g})$	$D(\mathbf{g}) = \max(\text{dist}_{ij}(\mathbf{g}))$
Radius	${}^0\text{TI}_R(\mathbf{g})$	$R(\mathbf{g}) = \min(\text{dist}_{ij}(\mathbf{g}))$
Shape Coefficient	${}^0\text{TI}_{I_2}(\mathbf{g})$	$I_2(\mathbf{g}) = \frac{D(\mathbf{g}) - R(\mathbf{g})}{R(\mathbf{g})}$
Eccentricity	${}^0\text{TI}_{\text{Cecc}}(\mathbf{g})$	$C_{\text{ecc}}(\mathbf{g}) = \frac{1}{\max(\text{dist}_{ij}(\mathbf{g}))}$
Closeness	${}^0\text{TI}_{\text{Clo}}(\mathbf{g})$	$C_{\text{clo}}(\mathbf{g}) = \frac{1}{\sum \text{dist}_{ij}(\mathbf{g})}$

^a Is the notation of TIs used in this work. ^b The parameters ${}^A p_k(j)$, ${}^k p_{jj}$, denotes the absolute probabilities of finding an aminoacid with charge q or the probabilities of self-return to an aminoacid with charge q after a loop type random walk of length k within the 2DL-SCN. The indices R and D are the topological radius and the topological diameter obtained from the distance matrix.

In this work, LDA was used to link the TIs of the sequence 2DL-SCN networks with the cellular entry route of a series of Human Rhinoviruses and discriminate the HRV strains. We developed in total 11 different QSAR classification models, one for each class of TIs. These four groups characterise the physicochemical nature of the amino acids as polar, non-polar, acidic or basic in essence Hydrophobicity or Polarity (HP). This kind of classifications has been used for the annotation of protein fragment patterns and motifs or generate HP-Lattice networks or the same 2DL-SCN [30]. Classification as acidic or basic prevails over the polar/non-polar classification in such a way that the four groups do not overlap each other. Subsequently, each amino acid in the sequence is placed in a Cartesian 2D space starting with the first amino acid at the (0, 0) coordinates. The coordinates of the successive amino acids are calculated as follows (in a similar manner as for DNA spaces) [31]. The names, symbols, and notation of all these TIs that entered into the 11 models after statistical analysis appear in the **Table 1**.

The equations of the models appear at follows. Once the different 2DL-SCN had been generated for the proteins in the different viruses, a series of total and local molecular descriptors for the whole sequence and the different amino-acids in the sequence were calculated with the aforementioned programme MARCH-INSIDE 2.0.[29] We calculated these TIs for the whole

Table 2. Experiment 1: Coefficients of QSAR models using TIs of 2DL-SCN vs. TGIs of 3D-SCN.

TGIs of VSCNs					
Markov Chain indices					
VSCNs			HP-Lattice network		
TGIs type	TGIs	QSAR Coefficient	QSAR Coefficient	TIs	TIs type
Entropy	${}^5\text{TGI}_\theta(\text{T})$	0.84	0.09	${}^5\text{TI}_\theta(\text{T})$	Entropy
	${}^0\text{TGI}_\theta(\text{L})$	289.9	48.6	${}^0\text{TI}_\theta(\text{L})$	
	a_0	-22.53	-8.77	a_0	
Spectral Moment	${}^2\text{TGI}_\pi(\text{T})$	0	0.24	${}^2\text{TI}_\pi(\text{T})$	Spectral Moment
	${}^0\text{TGI}_\pi(\text{L})$	17.3	31.14	${}^0\text{TI}_\pi(\text{L})$	
	a_0	-15.39	-14.61	a_0	
Classic TIs					
Wiener	${}^0\text{TGI}_w(\text{T})$	$9.6 \cdot 10^{-4}$	$8.0 \cdot 10^{-4}$	${}^0\text{TI}_w(\text{T})$	Wiener
	${}^0\text{TGI}_w(\text{L})$	0.09	3.24	${}^0\text{TI}_w(\text{L})$	
	a_0	23.42	-4.68	a_0	
Randic	${}^0\text{TGI}_\chi(\text{T})$	$5.04 \cdot 10^{-4}$	-1.72	${}^0\text{TI}_\chi(\text{T})$	Randic
	${}^0\text{TGI}_\chi(\text{L})$	-14.97	0.72	${}^0\text{TI}_\chi(\text{L})$	
	a_0	15.12	-5.56	a_0	
Sum of Node Degrees	${}^0\text{TGI}_\delta(\text{T})$	-0.11	0.06	${}^0\text{TI}_\delta(\text{T})$	Sum of Node Degrees
	${}^0\text{TGI}_\delta(\text{L})$	-14.97	3.4	${}^0\text{TI}_\delta(\text{L})$	
	a_0	15.12	-18.63	a_0	
Diameter	${}^0\text{TGI}_D(\text{T})$	0.58	-1.72	${}^0\text{TI}_D(\text{T})$	Diameter
	${}^0\text{TGI}_D(\text{L})$	-0.75	0.72	${}^0\text{TI}_D(\text{L})$	
	a_0	-7.07	4.53	a_0	

2DL-SCN and for local groups of amino-acids. In this work we used the uniform notation ${}^k\text{TI}_c(g)$ for all TIs; where is the classic symbol of the TI and refers to one of the 11 classes of TIs calculated, k is the order of the TI within the class, and g to the local group of amino-acids. When a TI is calculate for the entire lattice the local group $g = T$, indicating that it is a Total and not local TI. The classes of TIs considered were 11. If the TI belong to a class without TIs of different order we use $k = 0$. The groups of TIs include the $g = T$ and other 20 groups for each kind of amino-acid. The calculations of these and other TIs for different graphs/networks have been explained in detail before; consequently we give herein only general formulae in **Table 2** [12].

Classic TIs models

Balaban TIs model:

$$S = -7.97 \cdot 10^{-8} \cdot {}^0\text{TI}_J(T) + 0.09 \cdot {}^0\text{TI}_J(L) - 0.92 \quad (3)$$

Wiener TIs model:

$$S = 9.6 \cdot 10^{-4} \cdot {}^0\text{TI}_W(T) - 0.92 \cdot {}^0\text{TI}_W(L) + 23.47 \quad (4)$$

MTI TIs model:

$$S = -7.0 \cdot 10^{-5} \cdot {}^0\text{TI}_{MTI}(T) - 1.50 \cdot {}^0\text{TI}_{MTI}(L) + 21.94 \quad (5)$$

Randic Connectivity TIs model:

$$S = 5.04 \cdot 10^{-4} \cdot {}^0\text{TI}_\chi(T) - 14.97 \cdot {}^0\text{TI}_\chi(L) + 15.12 \quad (6)$$

Lattice network nodes Sum of Degrees TIs model:

$$S = -0.11 \cdot {}^0\text{TI}_\delta(T) - 9.78 \cdot {}^0\text{TI}_\delta(L) + 28.10 \quad (7)$$

Shape coefficient TIs models

Radius TIs model:

$$S = 1.54 \cdot {}^0\text{TI}_R(T) - 0.86 \cdot {}^0\text{TI}_R(L) - 12.7 \quad (8)$$

Diameter TIs model:

$$S = 0.58 \cdot {}^0\text{TI}_D(T) - 0.75 \cdot {}^0\text{TI}_D(L) - 7.07 \quad (9)$$

Shape coefficient type TIs model:

$$S = 3.07 \cdot {}^0TI_{I_2}(T) + 11.04 \cdot {}^0TI_{I_2}(L) - 0.92 \quad (10)$$

Markov Chain TIs Models

Entropy TIs model:

$$S = 0.09 \cdot {}^3TI_{\ominus}(T) + 48.60 \cdot {}^0TI_{\ominus}(L) - 8.77 \quad (11)$$

Spectral Moments TIs model:

$$S = 0.24 \cdot {}^0TI_{\pi}(T) + 31.14 \cdot {}^1TI_{\pi}(L) - 14.61 \quad (12)$$

2DL-Electrostatic Potential TIs model:

$$S = -22.21 \cdot {}^0TI_{\xi}(T) + 3.47 \cdot {}^0TI_{\xi}(L) + 32.68 \cdot {}^5TI_{\xi}(L) + 16.04 \quad (13)$$

Almost all models have p -level values <0.05 and proved to have very good predictability in training series. The Wiener indices model is the only one that correctly classified all the HRV strains in training and cross validation series. The models with the Balaban, MTI, Lattice Electrostatic Potential, and Spectral Moment HP-Lattice network descriptors correctly evaluated 100% of the viruses in training but misclassified some strains in validation. However, the Balaban model has a p -level relative high for a statistical significant model considering that $p = 0.05$ is just the threshold value for the test. The Shape coefficient indices model showed the lowest discriminatory power, with only 60% of the LDLR group correctly evaluated and 88.9% of ICAM-1 recognised by the theoretical model.

3.2. Back Projection Analysis of 2DL-SCN based models

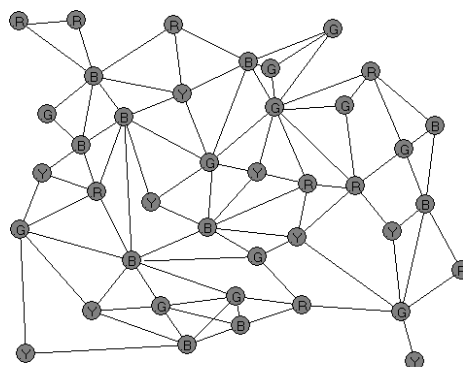
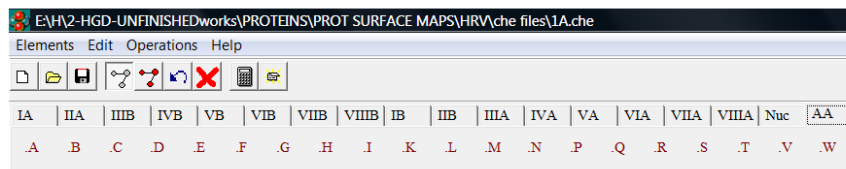
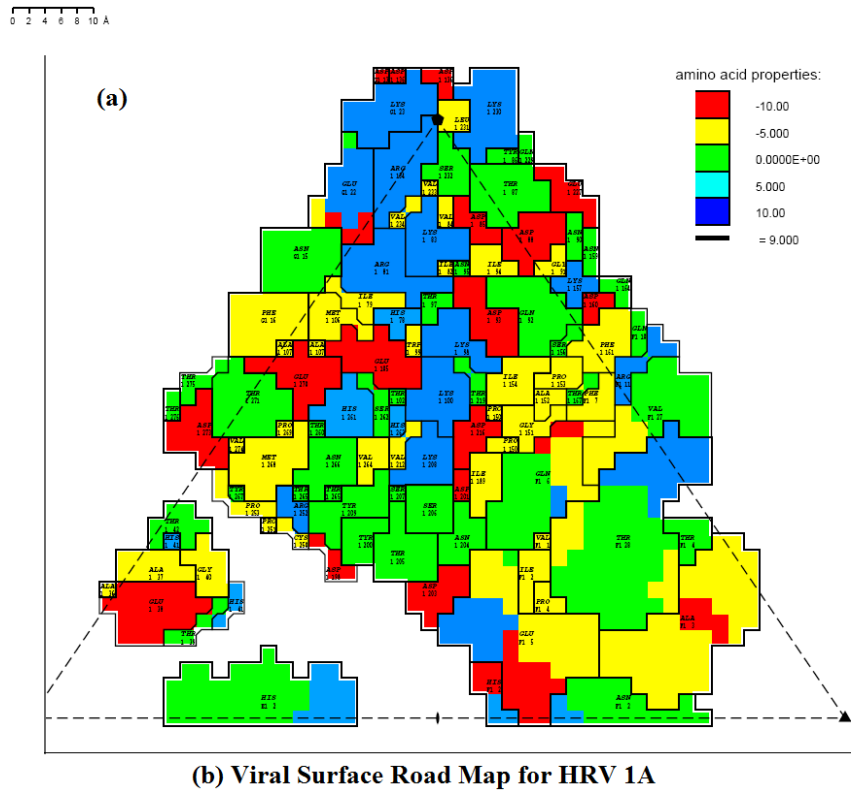
In the context of QSAR the so called Back-Projection Analysis is the process of drawing a map that depicts the influence of every molecular substructure on the property under investigation [32]. Some of the descriptors used in this study, such as entropy or moment, allow this type of approach to be used, a fact that is extremely useful in terms of interpreting the results. These descriptors and the application of this type of back-projection analysis approach have been reported in previous publications [33]. In fact, one of the most significant advantages of the QSAR approach reported here concerns to the interpretation of the results in terms of the influence that each network

sub-structure has over the property in question. This information can be obtained by applying a BPA, which consists of the projection of the QSAR model backwards onto the 2DL-SCN network. In this sense, we may first to make a partition of the network or graph into nodes. Then, calculate the local TIs of these network nodes. Later we can substitute the value of these local TIs or node centralities into the QSAR model, and finally sum the contribution of nodes re-grouped into sub-networks to map this fragment contributions over the network in a colour scale [34]. This kind of analysis has been largely reported for sub-graphs in the QSAR study of small-sized molecules [35].

We extended BPA to map the function of the different fragment of a protein backwards over the network representation of the large secondary structure of the corresponding RNA [33]. In a very recent study, the contributions obtained in this way can be matched against the degree of conservation of this sequences fragments by BLAST-based sequence alignment [36]. In this work, the BPA of the QSAR model was carried out by the first time using the node TIs on selected 2DL-SCN structures and the results are shown in **Figure 2**. In this figure, the fragments that contribute most to the interaction with the viral receptor are represented by darker colours and those that contribute least by lighter colours. The node containing the Lys of the HI loop appears as a black dot. The TIs of a node for small-sized graphs and large complex networks (also known as node centralities) formally differ but they are essentially local TIs of the same graph/network nature [13]. For instance, see the case of Closeness-vitality, $C_{clv}(j) = W(G) - W(G/j)$; a node centrality of complex networks derived as the difference between the Wiener index of the network with and without the node j [37].

The 2DL-SCN of the protein sequences of HRVs 2, 1A, 29 and 1B were partitioned into 4 or 5 relevant fragments (F1-F5) depending on the structures studied. The contributions of the different fragments to the interaction with the viral receptor were then calculated. This calculation was carried out with the spectral moment and the entropy model. The calculation of fragment contributions with spectral moments based QSAR models is one of the most extended in the literature [38]. We also perform the calculation with the Entropy model to illustrate that the results obtained for both models are very similar, and validate the consistency of the method. Interestingly, for all models the statistical procedure selected local descriptors of the region of HI loop. This region of the virus presents the higher contribution to binding the low-density lipoprotein receptor (LDLR). This is the region in which the lysine is maintained in the HI loop located in fragment F2. It appears that this amino acid is very important in influencing the entry route of the virus, although it is possible that a range of factors could be responsible for the

ability of the virus to penetrate the cell through various mechanisms [39]. Fragment F1 also appears to make a significant contribution in the four proteins studied, although the contribution is markedly lower than that of fragment F2.



Position: 1439-753

(b) MARCH-INSIDE view of network for HRV 1A

Figure 2. Surface road map and 3D-SCN for HRV 1A

3.3. QSAR models based on TIs of 3D Surface Complex Networks

In general Complex Networks other than the Lattice-like networks above treated are of wide use in modern science including proteins as well [13]. Different types of protein contact maps or protein structural Complex Networks can be used to represent spatial protein structure information in the form of 2D graph/network representations. In general, in these networks two amino-acids (nodes) are connected by an edge if they are spatial neighbours or the nodes or edges of the network are weighted with 3D structure dependent labels. Consequently the TIs derived for these classes of networks depend on the 3D structure of the protein. Consequently, several researches prefer to call these TIs as the graph Topographic Indices (TGIs). In previous works, we investigated the road maps of HRV surface. These road maps are in mathematical terms protein 3D-Surface Complex Networks (3D-SCN) with viral surface exposed amino-acids playing the role of nodes. In the 3D-SCN two amino-acids (nodes) are connected by an edge (arc) if they are spatial neighbours (adjacent) in the virus surface. The reader should be aware that being chemically connected (continuous amino-acids in the protein sequence or S-S bridge connected amino-acids) is not a condition necessary not sufficient to be a neighbour in the 3D-SCN. The construction of 3D-SCN or viral road maps has been given in detail before, so we refer to the original work and omit here detailed explanations. In **Figure 3** we illustrate a road map of the HRV surface and the corresponding 3D-SCN for the viral strain HRV 1A.

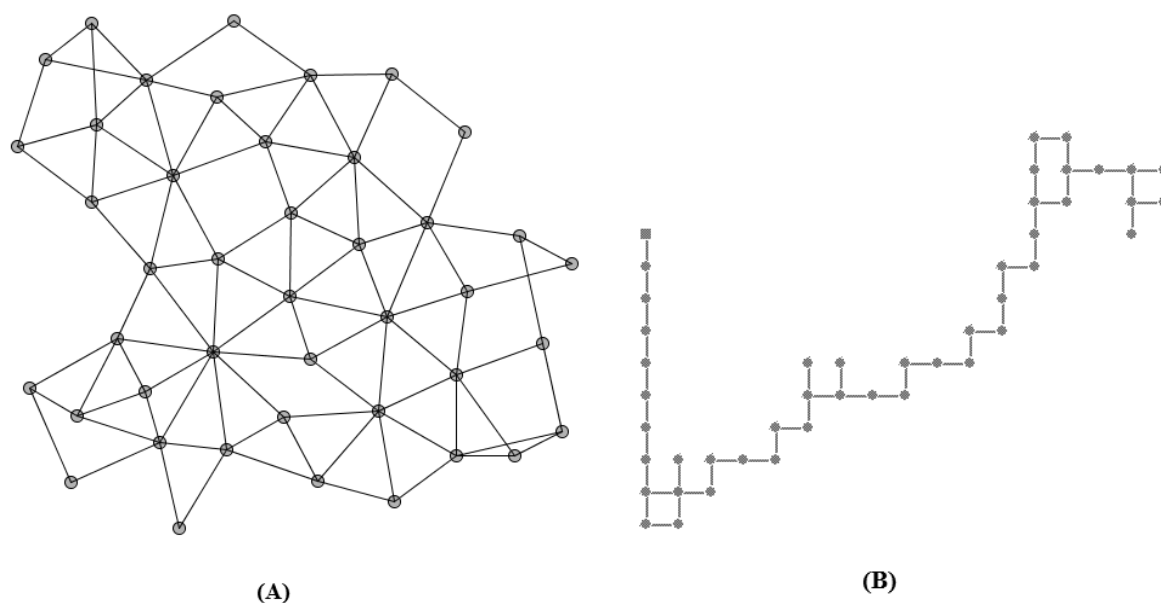


Figure 3. (A) 3D-SCN, and (B) 2DL-SCN for HRV2 strain.

In this work, we perform a comparison between the QSAR models obtained with the TIs of HP-Lattice network (see previous section) and the TGIs of a 3D-SCN. Two of the QSAR models based on TGIs of 3D-SCN have been reported before (the Markov Chain Entropy and Electrostatic Potential) [16, 17]. The other models based on TGIs of 3D-SCN and used in the comparison are being reported here by the first time:

Markov Chain TGIs Models:

Spectral Moments TGIs model (previously reported) [16]:

$$S = 3.11 \cdot {}^2TGI_{\pi}(Bs) + 17.30 \cdot {}^0TGI_{\pi}(L) - 15.39 \quad (14)$$

Entropy TGIs model (previously reported) [17]:

$$S = 0.84 \cdot {}^5TGI_{\ominus}(T) + 289.9 \cdot {}^0TGI_{\ominus}(L) + 24.07 \cdot {}^0TGI_{\ominus}(Bs) - 22.53 \quad (15)$$

Absolute Probability TGIs model (reported in this work):

$$S = 24.27 \cdot {}^0TGI_{pa}(Bs) + 384.21 \cdot {}^0TGI_{pa}(L) - 17.22 \quad (16)$$

Models based on SCN TGIs analogues of Classic TIs

Wiener TGIs model (reported in this work):

$$S = 8.0 \cdot 10^{-4} \cdot {}^0TGI_w(T) + 3.24 \cdot {}^0TGI_w(L) - 4.68 \quad (17)$$

SCN Sum of Degrees model (reported in this work):

$$S = 0.06 \cdot {}^0TI_{\delta}(T) + 3.4 \cdot {}^0TI_{\delta}(L) - 18.63 \quad (18)$$

Diameter TGIs model (reported in this work):

$$Vr(LDLR / ICAM) = -1.72 \cdot {}^0TGI_D(T) + 0.72 \cdot {}^0TGI_D(L) + 4.53 \quad (19)$$

Randic TGIs model (reported in this work):

$$S = -4.86 \cdot {}^0TGI_{\chi}(T) + 23.2 \cdot {}^0TGI_{\chi}(L) - 5.56 \quad (20)$$

Models based on SCN TGIs analogues of Complex Networks Centralities

Node Eccentricity TGIs model (reported in this work):

$$S = 286.9 \cdot {}^0TGI_{Cecc}(T) + 3729.9 \cdot {}^0TGI_{Cecc}(L) - 59.59 \quad (21)$$

Node Closeness TGIs model (reported in this work):

$$S = -630.3 \cdot TGI_{C_{clo}}(T) + 61379.7 \cdot TGI_{C_{clo}}(L) - 2.393 \quad (22)$$

We have to be aware that the 2DL-SCN like in the case of Nandy type lattices for nucleotides is a 2D projection of a pseudo-folding (raw approximation) to protein or nucleic acid folding in the 3D space [40]. On the other hand, the 3D-SCN presupposes a detailed knowledge of the 3D folding of the protein to determine which amino acids are surface neighbours. Consequently, the QSAR models based on TGIs of 3D-SCN were statistically significant and based on a more realistic network than the 2DL-SCN model. However, the 2DL-SCN demonstrated to be enough rigorous to produce accurate QSAR models based on them. These difference in the type of information encode determine that the TGIs of 3D-SCN and the TIs of 2DL-SCN are essentially different even when they are based on the same type of invariant. In order to visually illustrate the differences between of 3D-SCN vs. 2DL-SCN we shown both type of networks for the virus strain HRV2 in **Figure 4**.

In general, we should not expect the same behaviour in the coefficients of the QSAR model even for the same class of invariants. For instance, QSAR coefficients of the same type of entropy differ from one network to the other but the entropy QSAR models are both accurate. It coincides with the successful application of entropy type measures to codify information content at different structural scale of the system reported by Graham *et al* [41, 42]. Interestingly, we can determine the exact QSAR coefficients of some TIs of 2DL-SCN and their more rigorous TGIs analogues based on 3D-SCN; demonstrating the previous statement. Thence, we can confirm here that the utility of a type of graph invariant depend not only on the type of problem we are trying to solve. It depends on the database, or the invariant formula but also on the type of graph representation used, which justifies the recent search by many authors of new graph or network type representations for nucleic acids, [43, 44] proteins, [45, 46] or proteomic maps [47].

4. Conclusions

We demonstrated that TIs of SCN are indices of general use at different structure organization levels. In particular, we show that both TIs of realistic folding 3D-SCN and TGIs of pseudo-folding 2DL-SCN of viral capsids predict HRVs-receptor Interactions.

Acknowledgments

González-Díaz H. acknowledges tenure track research position funded by Program Isidro Parga Pondal, Xunta de Galicia.

References

1. Balaban AT, Beteringhe A, Constantinescu T, Filip PA, Ivanciuc O. Four new topological indices based on the molecular path code. *Journal of chemical information and modeling*. 2007 May-Jun;47(3):716-31.
2. Marrero-Ponce Y. Linear indices of the "molecular pseudograph's atom adjacency matrix": definition, significance-interpretation, and application to QSAR analysis of flavone derivatives as HIV-1 integrase inhibitors. *J Chem Inf Comput Sci*. 2004 Nov-Dec;44(6):2010-26.
3. Verma RP, Hansch C. Understanding human rhinovirus infections in terms of QSAR. *Virology*. 2007 Mar 1;359(1):152-61.
4. Estrada E. Characterization of the folding degree of proteins. *Bioinformatics*. 2002;18:697-704.
5. Nandy A, Nandy P. Graphical analysis of DNA sequence structure: II. Relative abundances of nucleotides in DNAs, gene evolution and duplication. *Curr Sci*. 1995;68:75-85.
6. Liao B, Ding K, Wang T. On A Six-Dimensional Representation of RNA Secondary Structures. *J Biomol Struc Dynamics* 2005;22:455-64.
7. Chikenji G, Fujitsuka Y, Takada S. Shaping up the protein folding funnel by local interaction: lesson from a structure prediction study. *Proc Natl Acad Sci U S A*. 2006 Feb 28;103(9):3141-6.
8. Jiang M, Zhu B. Protein folding on the hexagonal lattice in the HP model. *J Bioinform Comput Biol*. 2005 Feb;3(1):19-34.
9. Fernández M, Caballero F, Fernández L, Abreu JI, Acosta G. Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. *Proteins*. 2008;70(1):167-75.
10. 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.
11. Chou KC, Cai YD. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. *J Cell Biochem*. 2003 Dec 15;90(6):1250-60.
12. González-Díaz H, Vilar S, Santana L, Uriarte E. Medicinal Chemistry and Bioinformatics – Current Trends in Drugs Discovery with Networks Topological Indices. *Curr Top Med Chem*. 2007;7(10):1025-39.
13. González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E. Proteomics, networks and connectivity indices. *Proteomics*. 2008;8:750-78.
14. González-Díaz H, Saiz-Urra L, Molina R, Gonzalez-Diaz Y, Sanchez-Gonzalez A. Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *J Comput Chem*. 2007 Jan 31;28(6):1042-8.
15. Agüero-Chapin G, González-Díaz H, Molina R, Varona-Santos J, Uriarte E, González-Díaz Y. Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from *Psidium guajava* L. *FEBS Lett*. 2006;580 723-30.

16. González-Díaz H, Uriarte E. Biopolymer stochastic moments. I. Modeling human rhinovirus cellular recognition with protein surface electrostatic moments. *Biopolymers*. 2005 Apr 5;77(5):296-303.
17. González-Díaz H, Molina, R.R., Uriarte, E. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. *Polymer*. 2003(45):3845-53.
18. STATISTICA-6.0. 6.0 ed. Tulsa, OK, U.S.A.: StatSoft Inc. 2002.
19. Marrero-Ponce Y, Khan MT, Casanola Martin GM, Ather A, Sultankhodzhaev MN, Torrens F, et al. Prediction of Tyrosinase Inhibition Activity Using Atom-Based Bilinear Indices. *ChemMedChem*. 2007 Apr 16;2(4):449-78.
20. Castillo-Garit JA, Marrero-Ponce Y, Torrens F, Rotondo R. Atom-based stochastic and non-stochastic 3D-chiral bilinear indices and their applications to central chirality codification. *J Mol Graph Model*. 2007 Jul;26(1):32-47.
21. Vlasak M, Blomqvist S, Hovi T, Hewat E, Blaas D. Sequence and structure of human rhinoviruses reveal the basis of receptor discrimination. *J Virol*. 2003;77:6923-30.
22. Herz J. Deconstructing the LDL receptor--a rhapsody in pieces. *Nat Struct Biol*. 2001;8:476-8.
23. Liao B, Wang TM. New 2D graphical representation of DNA sequences. *J Comput Chem*. 2004 Aug;25(11):1364-8.
24. Zhang Y, Chen W. Analysis of similarity/dissimilarity of long DNA sequences based on three 2DD-curves. *Comb Chem High Throughput Screen*. 2007 Mar;10(3):231-7.
25. Liao B, Wang T, Ding K. On A Seven-Dimensional Representation of RNA Secondary Structures. *Molecular Simulation*. 2005;31(14):1063-71.
26. Thachuk C, Shmygelska A, Hoos HH. A replica exchange Monte Carlo algorithm for protein folding in the HP model. *BMC Bioinformatics*. 2007 Sep 17;8(1):342.
27. Liao B, Xiang X, Zhu W. Coronavirus phylogeny based on 2D graphical representation of DNA sequence. *J Comput Chem*. 2006;27(11):1196-202.
28. Randić M, Vracko M, Nandy A, Basak SC. On 3-D graphical representation of DNA primary sequences and their numerical characterization. *J Chem Inf Comput Sci*. 2000 Sep-Oct;40(5):1235-44.
29. Gonzales-Diaz H, Molina R, Hernandez I. MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design). 2.0 ed: Chemicals Bio-actives Center, Central University of Las Villas, Cuba. 2006:MARCH-INSIDE version 2.0 (Markovian Chemicals In Silico Design). This is a preliminary experimental version, a future professional version shall be available to the public. For any information about it, sends and e-mail to the corresponding author gonzalezdiazh@yahoo.es or humbertogd@uclv.edu.cu.
30. Berger B, Leighton T. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J Comput Biol*. 1998 Spring;5(1):27-40.
31. Randić M. Graphical representations of DNA as 2-D map. *Chem Phys Lett* 2004;386(4-6):468-71.
32. Stiefl N, Baumann K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure-Activity Relationship Technique. *J Med Chem*. 2003;46 1390-407.
33. González-Díaz H, Agüero-Chapin G, Varona-Santos J, Molina R, de la Riva G, Uriarte E. 2D RNA-QSAR: assigning ACC oxidase family membership with

- stochastic molecular descriptors; isolation and prediction of a sequence from *Psidium guajava* L. *Bioorg Med Chem Lett*. 2005 Jun 2;15(11):2932-7.
34. Gia O, Marciani Magno S, González-Díaz H, Quezada E, Santana L, Uriarte E, et al. Design, synthesis and photobiological properties of 3,4-cyclopentenepsoralens. *Bioorg Med Chem*. 2005 Feb 1;13(3):809-17.
 35. Vilar S, Estrada E, Uriarte E, Santana L, Gutierrez Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds through the use of TOPS-MODE and 2D/3D connectivity indices. 2. Purine derivatives. *Journal of chemical information and modeling*. 2005 Mar-Apr;45(2):502-14.
 36. González-Díaz H, Agüero-Chapin G, Varona J, Molina R, Delogu G, Santana L, et al. 2D-RNA-Coupling Numbers: A New Computational Chemistry Approach to Link Secondary Structure Topology with Biological Function. *J Comput Chem*. 2007;28:1049–56.
 37. Junker BH, Koschuetzki D, Schreiber F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*. 2006 Apr 21;7(1):219.
 38. Estrada E, Vilar S, Uriarte E, Gutierrez Y. In silico studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. *J Chem Inf Comput Sci*. 2002 Sep-Oct;42(5):1194-203.
 39. Vlasak M, Roivainen M, Reithmayer M, Goesler I, Laine P, Snyers L, et al. The minor receptor group of human rhinovirus (HRV) includes HRV23 and HRV25, but the presence of a lysine in the VP1 HI loop is not sufficient for receptor binding. *J Virol*. 2005 Jun;79(12):7389-95.
 40. Randić M, Vračko M, Nandy A, Basak SC. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. *J Chem Inf Comput Sci*. 2000;40:1235-44.
 41. Graham DJ. Information Content in Organic Molecules: Brownian Processing at Low Levels. *Journal of chemical information and modeling*. 2007;47(2):376-89.
 42. Graham DJ. Information Content in Organic Molecules: Structure Considerations Based on Integer Statistics. *J Chem Inf Comput Sci*. 2002;42:215.
 43. Nandy A, Harle M, Basak SC. Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*. 2006;9:211-38.
 44. Raychaudhury C, Nandy A. Indexing Scheme and Similarity Measures for Macromolecular Sequences. *J Chem Inf Comput Sci*. 1999;39 243-7.
 45. Randić M, Butina D, Zupan J. Novel 2-D graphical representation of proteins. *Chem Phys Lett*. 2006;419 528-32.
 46. Fernández M, Caballero J, Fernández L, Abreu JI, Garriga M. Protein radial distribution function (P-RDF) and Bayesian-Regularized Genetic Neural Networks for modeling protein conformational stability: Chymotrypsin inhibitor 2 mutants. *J Mol Graph Model*. 2007;26(4):748-59.
 47. Randić M, Estrada E. Order from chaos: observing hormesis at the proteome level. *Journal of proteome research*. 2005 Nov-Dec;4(6):2133-6.