Comparative Study of Topological Indices of Macro/Supramolecular RNA Complex Networks

Guillermín Agüero-Chapín,[†,‡,§] Agostinho Antunes,[§] Florencio M. Ubeira,[†] Kuo-Chen Chou,[||] and Humberto González-Díaz*,[†,||]

Department of Microbiology & Parasitology, Faculty of Pharmacy, University of Santiago de Compostela, Santiago de Compostela 15782, Spain, CAP, CEQA, and CBQ, Faculty of Chemistry and Pharmacy, Universidad Central Marta Abreu de la Villas, Santa Clara 54830, Cuba, CIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas 177, 4050-123 Porto, Portugal, and Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130

Received June 3, 2008

RNA function annotation is often based on alignment to a previously studied template. In contrast to the study of proteins, there are not many alignment-free methods to predict RNA functions if alignment fails. The use of topological indices (TIs) of RNA complex networks (CNs) to find quantitative structure−activity relationships (QSAR) may be an alternative to incorporate secondary structure or sequence-to-sequence similarity. Here, we introduce new QSAR-like techniques using RNA macromolecular CNs (mmCNs), where nodes are nucleotides, or RNA supramolecular CNs (smCNs), where nodes are RNA sequences. We studied a data set of 198 sequences including 18S-rRNAs (important phylogenetic molecular biomarkers). We constructed three types of RNA mmCNs: sequence-linear (SL), Cartesian-lattice (CL), and sequence-folding CNs (SF-CNs) and two smCNs: sequence-sequence disagreement CN (SSD) and sequence-sequence similarity (SSS-smCN). We reported the first comparative QSAR study with all these CIs and CNs, which includes: (i) spectral moments (${}^{i}\mu_{d}(w)$) of SL-mmCNs (accuracy = 75.3%), (ii) electrostatic CIs (ξ_d) of CL-mmCNs (>90%), (iii) thermodynamic parameters (ΔG, ΔH, ΔS, and T_m) of SF-mmCNs (64.7%), (iv) disagreement-distribution moments (M_k) of the SSD-smCN (79.3%), and (v) node centralities of the SSD-smCN (78.0%). Furthermore, we reported the experimental isolation of a new RNA sequence from *Psidum guajava* leaf tissue and its QSAR and BLAST prediction to illustrate the practical use of these methods. We also investigated the use of these CNs to explore rRNA diversity on bacteria, plants, and parasites from the *Dactylogyrus* genus. The HPL-mmCNs model was the best of all found. All the CNs and TIs, except SF-mmCNs, were introduced here by the first time for the QSAR study of RNA, which allowed a comparative study for RNA classification.

1. INTRODUCTION

A gene is not simply a DNA sequence; it is information that is converted to a useful product, a protein, or functional RNA molecule. Genes that encode RNA as their final products are often harder to identify than are protein-encoding genes, but even the latter can be very difficult to spot in a vertebrate genome. In this sense, bioinformatics techniques have illuminated the identification and retrieval of relevant sequences. The most common methodologies for sequence function annotation are based on sequence homologies to a template gene previously studied in another or the same species being the function entirely or partly defined by that relationship.[1] These informational techniques have limitations when there is low nucleotide identity between the query sequence and others already recorded with similar functions (remote homologues) or there is not a similar sequence recorded in the database.

Thus, the development of alignment independent tools for function prediction based on the calculation of numerical parameters derived from chemical structure of nucleic acids and proteins represented as complex networks (CNs) could be a solution. One way is extending the calculation of 1D, 2D, and 3D classic structural parameters for small-sized molecular graphs to nucleic acids and proteins macromolecular CNs (mmCNs). A recent review[2] highlighted the growing importance of machine learning methods for prediction of protein functional class, independent of sequence similarity. It is possible to derive from the mmCNs of proteins different numerical parameters to carry out structure−function studies, commonly called quantitative structure−activity relationship (QSAR). The QSAR models have been applied to small-sized and macromolecules to relate the structural parameters of mmCNs with biological function.[3] These methods often use as input 1D protein sequence numerical parameters specifically defined to seek sequence−function relationships. The indices called pseudo-amino acid compositions based on the 1D coupling numbers of the macromolecular sequence protein graph have been previously explored.[4,5] These indices can be

* To whom correspondence should be addressed. Email: humberto.gonzalez@usc.es. Tel: +34-981-563100. Fax: +34-981 594912.

[†] University of Santiago de Compostela.

[‡] Universidad Central Marta Abreu de la Villas (UCLV).

[§] Universidade do Porto.

[||] Gordon Life Science Institute.

10.1021/ci8001809 CCC: $40.75
Published on Web 10/21/2008

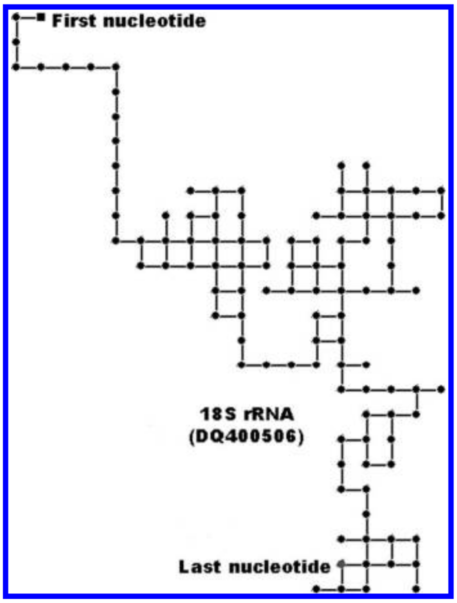


Figure 1. CL-mmCN for the new 18S-rRNA sequence isolated from *Psidum guajava* L (DQ400506).

classified as topological indices (TIs) or connectivity indices (CIs) of sequence linear (SL) graphs or mmCNs (SL-mmCNs). The concept of TIs and CIs are not considered as synonyms by some authors, but considering that they codify essentially information related to the patterns of connectivity or connectedness in the network, we refer all of them as connectedness or connectivity measures.[6−9] Many authors have introduced 2D or higher dimension graph theory or mmCN-like representations of protein sequences to calculate CIs. This constitutes an important step to uncover higher-order useful information not encoded by 1D sequence parameters.[10−13] However, the use of 3D descriptors can be limiting because it requires knowledge of the 3D structure of the biomacromolecules in details.[14,15] Despite the fact that 2D CIs are often derived from nonrealistic mmCNs representation for DNA and proteins, they are easier to obtain and contain important 2D information. We successfully introduced 2D sequence-coupling numbers associated to hydrophobicity-polarity (HP) 2D-lattice or Cartesian-type mmCNs representations for proteins related to plant metabolism.[11]

In a very recent review,[16] we showed that several QSAR studies have been done in proteins, but just a few were devoted to RNA.[17,18] Particularly, 2D-RNA mmCNs have been used for mRNAs, which are folded on its possible secondary structures following chemical−physical properties by accurate methods.[19] In RNA mmCNs, the nodes represent the nucleotides of one molecule of the nucleic acid and the edges of the network represent the main backbone chemical bonds, the hydrogen bonds, the spatial vicinity in 3D contact-map-like mmCNs, or both. We have reported the use of new CIs of folded structures (FS) of the RNA represented as mmCNs (FS-mmCNs), instead of proteins, to predict the function without reliance on alignment tools. Marrero-Ponce et al. have also used CIs derived from algebraic matrix linear and quadratic forms of mmCNs of proteins and RNA FS-mmCNs to predict properties without using alignment techniques.[18,20] Recent studies [21,22] have proposed the use of quantum chemical descriptors and weighted structural descriptors, closely related to molecular Randic CIs and Balaban distance indices as a distinctive characteristic of each RNA structure. Such works were also based on FS-mmCNs.

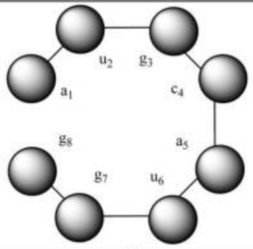
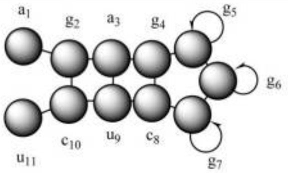
The CIs/TIs can be also calculated for supramolecular CNs (smCNs). As described above, in the case of small-sized molecular structure networks (ssmCNs), the nodes represent atoms, and the edge of the network represent the chemical bonds. In mmCNs the nodes are monomers (amino acids or nucleotides), and the edges are monomer−monomer interactions (peptidic bonds, intramolecular hydrogen bonds, or S−S bridges). In the case of smCNs, the nodes are whole molecules such as drugs, proteins, DNA, or RNA molecules as in the proteins interaction networks (PIN)[23] or the gene coexpression networks.[24] In mathematical terms, smCNs are very similar to mmCNs and ssmCNs. Although new, TIs are still being defined and introduced at the present time;[25] the classic TIs such as the Wiener index (W) and the Zagreb group connectivity indices (M1 and M2) are very useful because they can be calculated with software implemented to manipulate small-sized molecular networks of drugs, such as DRAGON,[26] or software specific for smCNs and mmCNs, such as CentiBin[27,28] and Pajek.[29,30] The CIs of mmCNs describe the connectivity of the macromolecular chemical structure, whereas the CIs of smCNs describe the relationships between the molecules and their neighbors in these CNs.[16,31] In a recent work, we showed that CIs of smCNs such as Wiener, closeness centrality, graph diameter, index of aggregation, assortative mixing coefficient, and connectivity differed clearly between PINs characterizing malignant tissue and PINs derived from randomly selected protein lists.[32]

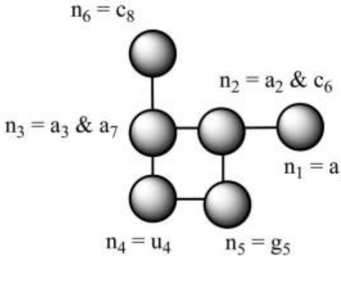
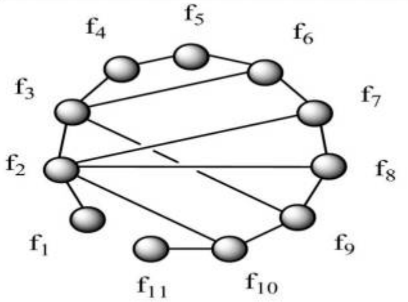
However, despite the large number of types of CIs for mmCNs and smCNs that can be used for QSAR studies of polymers, there are not many alternatives reported for the QSAR study of RNAs apart from CIs of SF-mmCNs. Here, we studied a data set of 198 control sequences and 18S-rRNAs, which are specific phylogenetic molecular biomarkers.[33,34] In this work, we developed and compare five QSAR classification models to identify if a sequence is as 18S-rRNA. We introduced herein new CIs and CNs to build up the QSAR studies of RNAs. Specifically we introduced the following RNA CNs: Cartesian-lattice mmCNs, sequence-linear mmCNs (SL-mmCNs), sequence-sequence disagreement smCN (SSD-smCN), and sequence-sequence similarity smCNs (SSS-smCNs). The QSAR model based on CIs of the CL-mmCN was the best of all found. Afterward, we reported in this work the experimental isolation of a new 18S-rRNA sequence from *Psidum guajava* leaf tissue, as well as its QSAR and BLAST prediction. This work introduce by the first time a series of CIs and CNs for the alignment-free QSAR study of RNAs and perform a comparative study of these techniques.

2. MATERIALS AND METHODS

2.1. Computational Methods. We downloaded 93 nucleotide sequences of 18S rRNAs (cDNA) from GenBank, belonging to different species of plants genus.[35] Each sequence was labeled by its accession number; see Table 1SM of Supporting Information. We set a control group composed by a heterogeneous group of 106 sequences comprehending diverse plant mRNAs; 5.8S, 26S, 28S

Table 1. Examples of RNA Sequence Fragments, Graph Representations, Matrices, and Some Parameters[a]

sequence	f1 $a_1u_2g_3c_4a_5u_6g_7g_8$	f2: $a_1g_2a_3g_4g_5g_6g_7c_8u_9c_{10}u_{11}$
name	SL-mmCN for f1	FS-mmCN for f2
matrix	$\begin{array}{c\|cccccccc} & a_1 & u_2 & g_3 & c_4 & a_5 & u_6 & g_7 & g_8 \\ a_1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ u_2 & 1 & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ g_3 & 0 & 1 & 3 & 1 & 0 & 0 & 0 & 0 \\ c_4 & 0 & 0 & 1 & 3 & 1 & 0 & 0 & 0 \\ a_5 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 \\ u_6 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 \\ g_7 & 0 & 0 & 0 & 0 & 0 & 1 & 3 & 1 \\ g_8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 3 \end{array}$	$\begin{array}{c\|ccccccccccc} & a_1 & g_2 & a_3 & g_4 & g_5 & g_6 & g_7 & c_8 & u_9 & c_{10} & u_{11} \\ a_1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ g_2 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ a_3 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ g_4 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ g_5 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ g_6 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ g_7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ c_8 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ u_9 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ c_{10} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ u_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$
picture		 a1 g2 a3 g4 g5 g6 c10 u9 c8 u11 g7
TIs	$^A\mu_0/L = 2/8 = 0.25$, $^G\mu_0/L = 3/8 = 0.375$	$\Delta G = \Delta H = -28.50$, $\Delta S = -91.89$
sequence	f3: $a_1a_2a_3u_4g_5c_6a_7c_8$	f4: $a_1a_2a_3a_4a_5a_6a_7a_8$
name	CL-mmCN for f3	SSS-smCN for fragments f1 to f11
matrix	$\begin{array}{c\|cccccc} & n_1 & n_2 & n_3 & n_4 & n_5 & n_6 \\ n_1 & p_{11} & p_{12} & 0 & 0 & 0 & 0 \\ n_2 & p_{21} & p_{22} & p_{23} & 0 & p_{25} & 0 \\ n_3 & 0 & p_{32} & p_{33} & p_{34} & 0 & p_{36} \\ n_4 & 0 & 0 & p_{43} & p_{44} & p_{45} & 0 \\ n_5 & 0 & p_{52} & 0 & p_{54} & p_{55} & 0 \\ n_6 & 0 & 0 & p_{63} & 0 & 0 & p_{66} \end{array}$	$\begin{array}{c\|ccccccccccc} & f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 & f_8 & f_9 & f_{10} & f_{11} \\ f_1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ f_2 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ f_3 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ f_4 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ f_5 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ f_6 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ f_7 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ f_8 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ f_9 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ f_{10} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ f_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{array}$

picture	 n6 = c8
	n2 = a2 & c6
	n3 = a3 & a7
	n1 = a1
	n4 = u4
	n5 = g5
		 f4
		f5
		f6
		f3
		f7
		f2
		f8
		f1
		f9
		f11
		f10

TIs	$\xi_0 = 0.29$, $\xi_1 = 0.54$, $\xi_2 = 0.40$	$C_{\delta 2} = 5$, $C_{\delta 4} = 2$, $C_{\delta 9} = 3$

[a] In the FS-mmCNs, we can express base composition as a function of the 0th spectral moments taking into consideration that: $^{total}\mu_0 = {}^A\mu_0 + {}^G\mu_0 + {}^C\mu_0 + {}^U\mu_0 = L = 8$ is the length of the sequence, For the supramolecular complex network, the indices $C_{\delta i}$, reported in this example are the node degrees centrality; this represents the number of RNA sequences similar to the fragment sequence represented by the ith node.

plant rRNAs; chloroplast 4.5S rRNA; 23S and 16S bacteria rRNAs and other RNA sequences from diverse sources was set as control group; see Table 2SM, Supporting Information. We used a MM to encode information about the RNA sequences. We generated different classes of CNs to represent the sequences of all sequences. The CIs specific for each class of CNs were calculated to codify numerically information related to sequence, sequence secondary structure (the mmCNs case), or sequence-to-sequence similarity (the smCNs case). These CIs can be used afterward as sequence structure numeric indices to seek the QSAR models.

2.1.1. mmCNs. 2.1.1.1. Model 1: Electrostatic CIs (ξ_d) of CL-mmCNs. We recently applied these numerical indices to proteins of interest in plant genetic engineering by our group in previous report.[11,36] This time, we propose by the first time to fold an RNA sequence into a 2D Cartesian lattice system that assigns each Cartesian axis to each one of the four RNA nucleotides. This graph groups purine and pyrimidine bases on the abscissas and ordinates axis respec-

Table 2. Concise Definitions of the CIs for mmCN and smCN Used in This Work

CI type	formula[a]	CI type	formula[a]
thermodynamic	$\Delta G = \Delta H - T\Delta S$	moments	$M_k = \frac{1}{n}\sum_{i}^{n} \mathrm{SSD}_{ij}^{\ k}$
spectral	${}^{i}\mu_d(w) = \mathrm{Tr}(A^k)$	moments	$\mathrm{MM}_k = \frac{1}{n}\sum_{i}^{n} (\mathrm{SSD}_{ij}^{\ k} - M_1)^k$
lattice electrostatics	$\xi_d = \sum_{j=1}^{n} {}^{A}p_d(j) \cdot \theta_j$	degree	$C_{\mathrm{deg}}(v) = \deg(v)$
current-flow betweenness	$C_{\mathrm{cfb}}(v) = \sum_{s,t \in V} \tau_{st}(v)/(n-1)(n-2)$	Katz status index	$C_{\mathrm{katz}} = \sum_{k=1}^{\infty} \alpha^k \cdot (\mathbf{A}^t)^k \cdot \mathbf{u}$
eccentricity	$C_{\mathrm{ecc}}(v) = \max\{\mathrm{dist}(v, w): w \in V\}^{-1}$	eigenvector	$C_e(v) = e_1(v)$
closeness	$C_{\mathrm{clo}}(v) = 1/\left(\sum_{w \in V} \mathrm{dist}(v, w)\right)$	Hubbell index	$C_{\mathrm{hubbell}} = \vec{E} + WC_{\mathrm{hubbell}}$
radiality	$C_{\mathrm{rad}}(v) = \sum_{w \in V} (\Delta_G + 1 - \mathrm{dist}(v, w))/(n-1)$	bargaining	$C_{\mathrm{brg}} = \alpha \cdot (I - \beta\mathbf{A})^{-1} \cdot \mathbf{A} \cdot \mathbf{u}$
centroid values	$C_{\mathrm{cen}}(v) = \min\{f(v, w): w \in V \setminus \{v\}\}$	page rank	$C_{\mathrm{pagerank}} = dPC_{\mathrm{pagerank}} + (1-d) \cdot \mathbf{u}$
stress	$C_{\mathrm{str}} = \sum_{s \notin v \in V} \sum_{t \notin v \in V} \delta_{st}(v)$	HITS-authority	$C_{\mathrm{auths}} = \mathbf{A}^T C_{\mathrm{hubs}}$
shortest-path betweenness	$C_{\mathrm{spb}} = \sum_{s \notin v \in V} \sum_{t \notin v \in V} \delta_{st}(v)$	HITS-hubs	$C_{\mathrm{hubs}} = \mathbf{A} \cdot C_{\mathrm{auths}}$
current-flow closeness	$C_{\mathrm{cfc}}(v) = (n-1)/\left(\sum_{t \notin V} p_{vt}(v) - p_{vt}(v)\right)$	closeness vitality	$C_{\mathrm{clv}}(v) = W(G) - W(G \setminus \{v\})$

[a] $G = (V, E)$, undirected or directed graph with $n = |V|$ vertices; deg(v), degree of the vertex v in an undirected graph; dist(v, w), length of a shortest path between the vertices v and w; σ_{st}, number of shortest paths from s to t; $\sigma_{st}(v)$, the number of shortest path from s to t that use the vertex v. **A** is the adjacency matrix of the graph G. All the CIs labeled in the for C_x are smCN CentiBin node Centralities.

tively. Figure 1 depicts the 2D Cartesian-like or lattice (CL-mmCN) representation for nucleotide sequence of 18S rRNA (DQ400506) from *Psidium guajava* L. The representation was carried out following Randic's method[37] by adding to the coordinates (x, y) of the kth nucleotide the values (1, 0) if the (k + 1)th nucleotide is Guanine (right-step), (−1, 0) if it is adenine (left-step), (0, 1) if it is cytosine (upward-step), or (0, −1) if the (k + 1)th nucleotide is uracil (downward-step).

Using Markov chain theory, we calculated the ξ_d values in the 2D-lattice considering isolate (noninteracting $d = 0$) nodes or taking into consideration the effect of neighbor nodes at different topological distances within the 2D Lattice ($d > 0$).[11] In Table 1, we depict an example of the calculation of ξ_d values. The method uses essentially three matrix magnitudes:[15]

(a) The matrix $^{1}\mathbf{\Pi}$ (see eq 1). This matrix is built up as a square matrix ($n \times n$).[38] Note that the number of nodes (n) in the 2D-lattice network may be equal or smaller than the number of nucleotides in the polynucleotide sequence.[12] The matrix $^{1}\mathbf{\Pi}$ contains the probabilities $^{1}p_{ij}$ to

Table 3. Summary of QSAR Models Based on Different Networks

features[a]	lattice network	sequence network	folding network	distance network	similarity network
CIs	ξ_d	${}^i\mu_d(w)$	ΔG, ΔH, ΔS, T_m	M_k	C_g
type	macromolecular	macromolecular	macromolecular	supramolecular	supramolecular
RNA-dim	1D	0D	2D	1D	1D
connectivity	partial	sequence	partial	total	partial
edges	fold step	BB	BB + HB	ever	SSS
weights	no	no	no	SSD	no
nodes	n bases	1 base	1 base	1 sequence	1 sequence
weights	φ_i	HB	no	no	no
degree	1−4	0	1−3	198	1−25
			train (%)		
series[b]	18S-rRNAs	other RNAs	18S-rRNAs	other RNAs	average
18S-rRNAs	97,1	84,3	72,9	67,1	61,4
other RNAs	95,0	67,5	57,5	90,0	92,5
total	96,0	75,3	64,7	79,3	78,0
			CV (%)		
series[b]	18S-rRNAs	other RNAs	18S-rRNAs	other RNAs	average
18S-rRNA	100	82,6	69,6	52,2	56,5
other RNAs	84,0	68,0	44,0	88,0	96,0
total	91,7	75,0	56,3	70,8	77,1

[a] Notes: RNA-dim, dimension of the chemical structure of the RNA necessary to calculate the descriptors; BB, backbone bond; HB, hydrogen bonds; SSD, sequence-sequence pair composition disagreement distance; SSS, sequence-sequence 250bp similarity; CV, cross-validation; connectivity partial, many but not all nodes connect each other; total, all nodes connect each other. [b] Rows are observed values and columns predicted ones.

reach a node n_i with charge q_i moving within the 2D-Lattice throughout a walk of length $k = 1$ from a node n_j with

$$ {}^1p_{ij} = \frac{\alpha_{ij}\cdot\left(\frac{q_j}{d_{0j}}\right)}{\sum_{m=1}^{n}\alpha_{im}\cdot\left(\frac{q_m}{d_{0m}}\right)} = \frac{\alpha_{ij}\cdot\theta_j}{\sum_{m=1}^{n}\alpha_{im}\cdot\theta_m} \quad (1) $$

$$ {}^Ap_0(j) = \frac{\left(\frac{q_j}{d_{0j}}\right)}{\sum_{m=1}^{n}\left(\frac{q_j}{d_{0j}}\right)} = \frac{\theta_j}{\sum_{m=1}^{n}\theta_j} \quad (2) $$

In these equations, q_j is the charge of the node n_j (q was standardized to a positive value, representing negative charge as the lowest positive values). The parameter α_{ij} equals 1 if the nodes n_i and n_j are adjacent in the graph and equals 0 otherwise. The distances d_{0j} or d_{0m} are the Euclidean distances from the corresponding nodes to the center of coordinates. Consequently, the ratios between q_j and d_{0j} can be interpreted as the electrostatic potential θ_j measured in the corresponding node n_j with respect to the center because of the additive effect of overlapping amino acids.[39]

(b) The charge vector ${}^0\mathbf{Q}$. The method considers that a total charge or weight (q_i) can be assigned to each node.

(c) The zero-order vector (${}^A\mathbf{\Pi}_0$) (see eq 3). This vector lists the absolute initial probabilities ${}^Ap_0(j)$, with which a node selected at random presents a given charge q_j, and it is presented as a row vector.

$$ {}^A\mathbf{\Pi}_d = {}^A\mathbf{\Pi}_0 \times {}^d\mathbf{\Pi} = {}^A\mathbf{\Pi}_0 \times ({}^1\mathbf{\Pi})^d \quad (3) $$

Last, we can use MARCH INSIDE to calculate the average ξ_d for any node n_j that one can reach in the 2D-lattice network by moving from any node n_i throughout the entire graph using walks of length d. Considering that the ξ_d are discrete average values, we determine them as the sum of two-term products. These ξ_d values encode in a stochastic manner the properties (charge) of the nucleotides or nodes placed at different distances in the sequence arranged in a 2D-lattice network. It is remarkable that these average

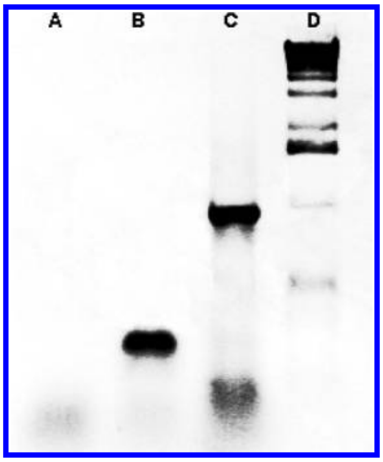


Figure 2. Isolation of 18S-rRNA fragment from *Psidium guajava* L: (A) negative control, (B) RT-PCR reaction with degenerated primers on total RNA, (C) positive control for Ready To Go RT-PCR Beads system kit, (D) 1 kb ladder (Gibco BLR).

potentials can be written as the product of ${}^{0}\xi$ and the natural powers of the matrix ${}^{1}\mathbf{\Pi}$ based on the Chapman—Kolgomorov equations:[40]

$$\xi_d = \sum_{j=1}^{n} {}^{A}p_d(j)\cdot\theta_j \;=\; {}^{0}\xi^{T}\cdot({}^{1}\mathbf{\Pi})^{d}\cdot{}^{0}Q \tag{4}$$

All calculations of ξ_d values for RNA sequences in both groups including sequence representation were carried out with our in-house software MARCH INSIDE.[41] The evaluation of expression 4 for $d = 0$ gives the order ξ_0; for $d = 1$ the short-range ξ_1, for $d = 2$ the middle-range ξ_2, and for $d > 2$ the long-range electrostatic potential values $\xi_{d>2}$.[41] This expansion have been reported for 2D-lattices and pseudo-3D networks[49] or 3D structures of proteins[39] and is reported herein by the first time for the linear graph $n_1-n_2-n_3$ characteristic of the RNA sequence (UUUC). Note that the central node contains both uracil and cytosine nucleotides

$$\begin{aligned}\xi_0 &= [{}^{A}p_0(n_1), {}^{A}p_0(n_2), {}^{A}p_0(n_3)]\cdot\begin{bmatrix}1\,0\,0\\0\,1\,0\\0\,0\,1\end{bmatrix}\cdot\begin{bmatrix}\theta_1\\\theta_2\\\theta_3\end{bmatrix}\\ &= {}^{A}p_0(n_1)\cdot\theta_1 + {}^{A}p_0(n_2)\cdot\theta_2 + {}^{A}p_0(n_3)\cdot\theta_3\end{aligned} \tag{5a}$$

$$\begin{aligned}\xi_1 &= [{}^{A}p_0(n_1), {}^{A}p_0(n_2), {}^{A}p_0(n_3)]\cdot\begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}\cdot\begin{bmatrix}\theta_1\\\theta_2\\\theta_3\end{bmatrix}\\ &= {}^{A}p_1(n_1)\cdot\theta_1 + {}^{A}p_1(n_2)\cdot\theta_2 + {}^{A}p_1(n_3)\cdot\theta_3\end{aligned} \tag{5b}$$

$$\begin{aligned}\xi_2 = [{}^{A}p_0(n_1), {}^{A}p_0(n_2), {}^{A}p_0(n_3)]&\cdot\begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}\cdot\\ &\begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}\cdot\begin{bmatrix}\theta_1\\\theta_2\\\theta_3\end{bmatrix}\end{aligned} \tag{5c}$$

$$\begin{aligned}\zeta_3 = [{}^{A}p_0(n_1), {}^{A}p_0(n_2), {}^{A}p_0(n_3)]&\cdot\begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}\cdot\\ \begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}&\cdot\begin{bmatrix}{}^{1}p_{11} & {}^{1}p_{12} & 0\\ {}^{1}p_{21} & {}^{1}p_{22} & {}^{1}p_{23}\\ 0 & {}^{1}p_{32} & {}^{1}p_{33}\end{bmatrix}\cdot\begin{bmatrix}\theta_1\\\theta_2\\\theta_3\end{bmatrix}\end{aligned} \tag{5d}$$

2.1.1.2. Model 2: Spectral Moment CIs (${}^{i}\mu_d(w)$) of SL-mmCNs. The μ_i values were calculated as the sum of the values in the main diagonal for the nucleotide−nucleotide adjacency matrix for the RNA sequence (**A**). This is a square matrix of $n \times n$ row and files associated to a path graph. In this mmCN, each nucleotide is represented by a node; the nucleotide in the dth position of the sequence is connected (adjacent) only to the nucleotides in the $(d - 1)$th and $(d + 1)$th positions of the sequence. The dth natural powers of $\mathbf{A}^d$ count the number of paths connecting the different pair of nodes (nucleotides) at distance d within the sequence. The trace operator (Tr) sum of the values in the main diagonal of $\mathbf{A}^d$ for a given type of nucleotides are called the nucleotide-specific local spectral moments, ${}^{i}\mu_d(w)$ of the sequence and are introduced herein as CIs in QSAR study of RNAs. We can use different properties or weights of the nucleotide (w) in the main diagonal entries, such as charge, hydrophobicity, mass, or the number of possible hydrogen bonds (HB) to differentiate the type of nucleotide. In

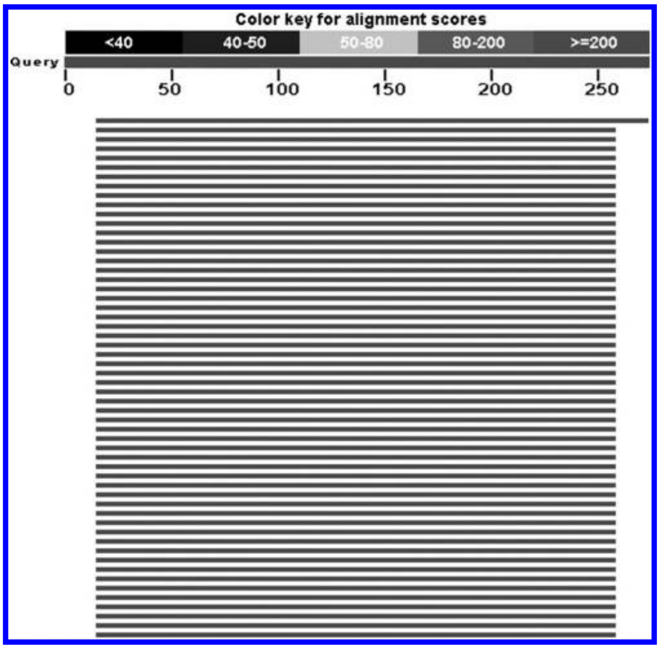


Figure 3. Results for the nBLAST experiment. The top rule scales number of bp; black bars indicate conserved regions for different sequences. Sequences names were not depicted.

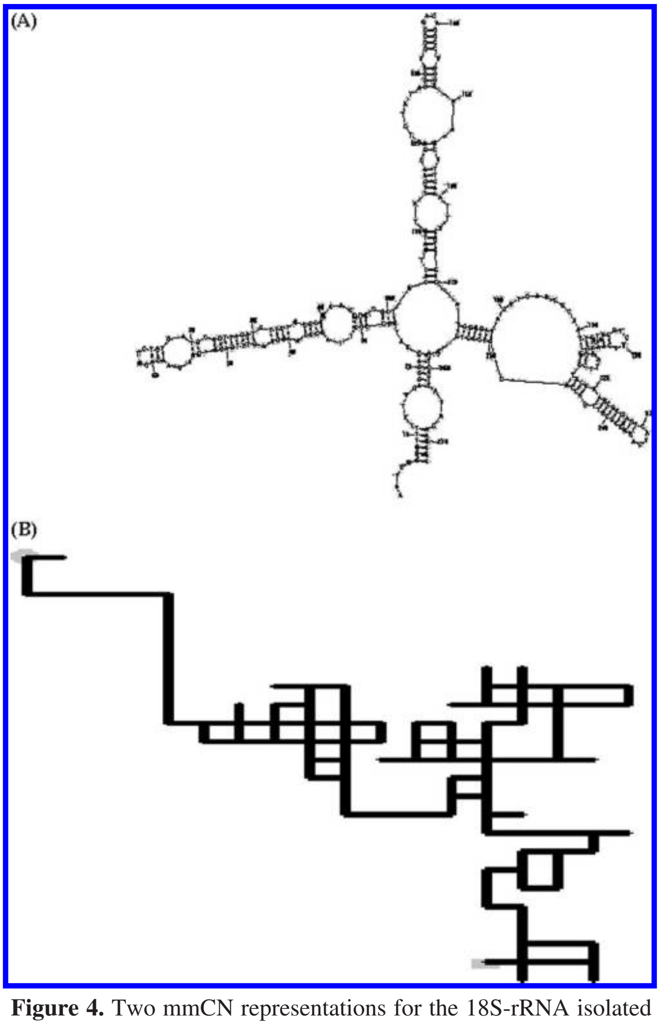


Figure 4. Two mmCN representations for the 18S-rRNA isolated from *P. guajava* L: (A) FS-mmCN and (B) CL-mmCN.

particular, the 0-order moments, ${}^{i}\mu_0(w)$, for any w coincide with the number of nucleotides of each type in the sequence and the 1st-order moments, ${}^{i}\mu_1(w)$, are equal to the sum of the w values for the nucleotides of type i. In the Table 1, we depict a short example on the calculation of ${}^{i}\mu_d(w)$ values. Below we illustrate the calculation of these indices using the HB as weights (3 for G and C or 2 for A and U) for the fragment UAU

$$^{U}\mu_0(\mathrm{HB}) = \mathrm{Tr}\begin{bmatrix} 2\,1\,0 \\ 1\,3\,1 \\ 0\,1\,2 \end{bmatrix}^0 \Bigg|_{ii=U} = \mathrm{Tr}\begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix} \Bigg|_{ii=U} = 2 \qquad (6a)$$

$$^{A}\mu_0(\mathrm{HB}) = \mathrm{Tr}\begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,1\,1 \end{bmatrix} \Bigg|_{ii=A} = 1 \qquad (6b)$$

$$^{U}\mu_1(\mathrm{HB}) = \mathrm{Tr}\begin{bmatrix} 2\,1\,0 \\ 1\,3\,1 \\ 0\,1\,2 \end{bmatrix} \Bigg|_{ii=A} = 4 \qquad (6c)$$

$$^{G}\mu_0(\mathrm{HB}) = \mathrm{Tr}\begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix} \Bigg|_{ii=A} = 0 \qquad (6d)$$

$$^{A}\mu_1(\mathrm{HB}) = \mathrm{Tr}\begin{bmatrix} 2\,1\,0 \\ 1\,3\,1 \\ 0\,1\,2 \end{bmatrix}^1 \Bigg|_{ii=A} = 3 \qquad (6e)$$

2.1.1.3. Model 3: Thermodynamic CIs (ΔG, ΔH, ΔS, and T_m) of SF-mmCNs. The list of all sequences was introduced as input of the RNA secondary structure prediction server DINAMelt. We used the fast-fold option named Quickfold (http://www.bioinfo.rpi.edu/applications/hybrid/quikfold.php). This server calculated the secondary structure of all the sequences, as well as the folding thermodynamic parameters free energy (ΔG), enthalpy (ΔH), entropy (ΔS), and melting temperature (T_m). The job parameters were as follows: energy rule RNA (2:3), at 37 °C, $[\mathrm{Na}^+] = 1$ M, $[\mathrm{Mg}^{2+}] = 0$ M, sequence type linear, structures 5% suboptimal, window size as default, maximum of 1 folding, and no limit to maximum distance between paired bases. Under these conditions, the computation took only a few seconds.[42] The equations used by definition are (see also Table 1, where we depict a short example on the calculation of thermodynamic CIs)

$$\Delta G = \Delta H - \mathrm{T}\Delta S \qquad (7a)$$

$$T_\mathrm{m} = 1000\left(\frac{\Delta H}{\Delta S}\right)\Bigg|_{\Delta G=0} \qquad (7b)$$

2.1.1.4. Model 4: CIs Based on Moments of the Distribution Function for the 250bp SSD-smCN. CIs based on moments of the SSD distribution function for the 250bp SSD-smCN were calculated on the basis of the similarity between sequences. First we calculated the percentage of disagreement (100 × number of no-identical bp/250 bp) for all pairs of sequences. We used only the first 250 bp and did not perform sequence alignment. For this calculation, each sequence letter or bp was introduced into one cell of the software STATISTICA in such a way that each sequence lies within one row but occupies 250 columns. In this way, each input variable contains the bp in one specific position of the sequence. For sequences with length (L) lower than 250 bp, we leave empty the last $250 - L$ columns. Next, we used STATISTICA cluster analysis module to calculate the matrix of sequence-sequence disagreement ($\mathrm{SSD}(\%)_{ij}$) between every pair of ith and jth sequences. This matrix can be represented by the SSD-smCN, which is a totally connected smCN, where the nodes are the sequences and all pairs of nodes are connected by an arc or edge-weighted with the $\mathrm{SSD}(\%)_{ij}$ value. With Microsoft Excel, it is very easy to calculate, for each sequence, the moments of the distribution function for $\mathrm{SSD}(\%)_{ij}$, which can be used as CIs of smCN in the QSAR study. We give at follow the formulas for the calculation of these indices including the mean value, ${}^{1}M$, and the variance, ${}^{2}\mathrm{MM}$, of the distribution function of $\mathrm{SSD}(\%)_{ij}$ for the jth sequence with respect to all the other n sequences.

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{SSD}(\%)_{ij})^k \qquad (8a)$$

$$\mathrm{MM}_k = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{SSD}(\%)_{ij} - {}^{1}M_{ij})^k \qquad (8b)$$

2.1.1.5. Model 5: CIs Based on Node Centralities of the 250bp SSS-smCN. First, we calculated the sequence-sequence similarity matrix based on the $\mathrm{SSD}(\%)_{ij}$ value calculated above. The values of this Boolean matrix are $\mathrm{SSS}_{ij} = 1$ if the ith and jth sequences have a $\mathrm{SSD}(\%)_{ij}$ lower than certain cutoff and $\mathrm{SSS}_{ij} = 0$ otherwise. It means that $\mathrm{SSS}_{ij} = 1$ indicates that the first 250bp of two sequences are similar according to the cutoff and without performing alignment. This new matrix can be represented in the form of a partially connected SSS-smCN. In the SSS-smCN, like in the SSD-smCN, the nodes are the sequences, but here, two nodes are connected only if the two respective sequences are similar ($\mathrm{SSD}(\%)_{ij}$ lower than cutoff). Using the software CentiBin,[28] we were able to calculate several node centralities for each sequence. Each node centrality is a parameter that numerically characterizes the connectivity or topology of the node with respect to all the neighbors in the network. These centralities were used as input CIs of a smCN in the QSAR studies. In Table 2, we summarize the calculation of the CIs for mmCN and smCN used in this work, including the node centralities mentioned here.[27]

2.1.2. Statistical Analysis. Finally, we uploaded to statistic analysis software, the STATISTICA 6.0, a row data files with all CIs for each sequence and a dependent variable indicating if it is an 18S rRNA sequence.[43] We used linear discriminant analysis (LDA) as modeling technique to build up the QSAR model. Best Subset or Forward Stepwise implemented in the LDA module of STATISTICA was the method used for feature selection. The statistical significance of the LDA model was determined with a Fisher's test by examining the canonical regression coefficient (Rc) and the respective p-level (p), which represents the overall significance of the variables included in the model.[44,45] We also inspected the percentage of good classification, cases/variables ratios (ρ parameter), and number of variables to be explored to avoid overfitting or chance correlation.[46] The training of the QSAR model was carried out by selecting at random 150 (~75%) out of 199 available sequences. The remaining 49 sequences (~25%) which were not used for training were then employed to test the predictive ability of the model.

2.2. Experimental Section. *2.2.1. Total RNA Extraction.* Dwarf Guava leaf tissue was ground in liquid nitrogen using a precooled and DEPC treated mortar and pestle. Total RNA was isolated from leaves, following the protocol described by Wadsworth et al.[47] The pellet was resuspended in 50 μL of RNase-free water at 50 °C. The total RNA concentration was measured using GENESYS 10 spectrophotometer, as well as its purity, at 260:280 ratios. RNA integrity was also checked by agarose gel electrophoresis.

2.2.2. Primer Design and RT-PCR. A couple of degenerated primers were designed after aligning 12 sequences of 18s rRNA from individuals belonging to the Rosidae subclass using Clustal × software. The forward and reverse primers were 5′-AAY TGG GGH TTY TTT GAG-3′ and 5′-GGY TTG TAY AAN GAN GC-3′ respectively. RT-PCR reaction was carried out using Ready To Go RT-PCR Beads system (AmershanPharmaciaBiotech), 2 μg of total RNA, and 1.0 μM of both primers in a total reaction volume of 50 μL. Reaction was completed in a single step using a Perkin-Elmer 2400 thermocycler programmed as follows: 30 min 42 °C (RT); 5 min 95 °C; 1 min 95 °C, 1 min 50 °C, and 1 min 72 °C (PCR) during 32 cycles plus final incubation at 7 min a 72 °C.

2.2.3. PCR Reaction. An additional PCR reaction was carried out using as template an aliquot of the RT-PCR reaction. The parameters of the PCR were 2 min 95 °C, 1 min 94 °C, 1 min 52 °C, 1 min 72 °C during 30 cycles, plus 72 °C during 7 min.

2.2.4. Cloning and Sequencing. The product of PCR was purified using GEL Band Purification kit (GE Healthcare, Piscataway, NJ). The band was cloned in a pGEM^{R}-TEasy vector (Promega, U.S.A.), and recombinant selection followed white and blue colony criteria with viable cells XLI-Blue in transformation. Sequencing of cloned fragment was performed using the ABI 3700 sequencer (Applied Biosystems).

3. RESULTS AND DISCUSSION

3.1. QSAR Studies. *3.1.1. Model Based on Electrostatic CIs of CL-mmCNs.* We performed LDA to validate the utility of CL-mmCNs electrostatic CIs in RNA QSAR-based biological function annotation. In this type of LDA-QSAR equations based on TIs/CIs,[48,49] N is the number of cases (RNA sequences used), U the Wilk's lambda statistical parameter, R_c is the canonical regression coefficient, and p the level of error. The equation showing the highest performance in discriminating 18S-rRNA sequences from other RNA sequences is described below.

Model 1: Using total sequence electrostatic CIs of CL-mmCNs

$$\text{18S rRNA score} = 19.319\xi_1 - 14.309\xi_2 - 0.38$$
$$N = 150 \quad R_c = 0.87 \quad U = 0.24 \quad p < 0.001 \qquad (6.1)$$

This model was able to classify correctly 146 out of 150 sequences (accuracy = 96.0%) in the training set using only two predictive variables. It also displayed very satisfactory values of accuracy (91.7%) in cross-validation with external prediction series. The present classifier has a high canonical regression coefficient (R_c =0.87) with a p-level lower than 0.001, which prove the statistical significance of the model.

3.1.2. Comparison with other QSAR models. We developed other LDA-QSAR models, in addition to the model mentioned above based on CL-mmCNs. We used other types of CIs for different classes of mmCNs and smCNs. This kind of comparative QSAR study of RNAs using different classes of CIs of mmCNs and smCNS have not been reported before in the literature. In total, we investigated four additional models using exactly the same training and CV series of 18S rRNA and non-18S rRNA sequences. Below we depict the equations and the main statistical parameters for these models.

Model 2: Using spectral moments of SL-mmCNs

$$\text{18S rRNA score} = 0.0005 \cdot {}^{\text{Total}}\mu_0 - 15.2\left(\frac{{}^{C}\mu_0}{{}^{\text{Total}}\mu_0}\right) + 2.602$$
$$N = 150 \quad R_c = 0.36 \quad U = 0.87 \quad p < 0.001 \quad \rho = 25.0 \qquad (6.2)$$

Model 3: Using thermodynamic parameters of FS-mmCNs

$$\text{18S rRNA score} = -0.82\Delta G + 0.06\Delta H - 0.9T_m + 55.85$$
$$N = 150 \quad R_c = 0.34 \quad U = 0.89 \quad p < 0.001 \quad \rho = 18.75 \qquad (6.3)$$

Model 4: Using distance distribution function moments of the 250bp SSD-smCN

$$\text{18S rRNA score} = 132.84M_1 - 180.1\text{MM}_1 + 1.6$$
$$N = 150 \quad R_c = 0.62 \quad U = 0.61 \quad p < 0.001 \quad \rho = 25.0 \qquad (6.4)$$

Model 5: Using node centralities of 250bp SSS-smCN

$$\text{18S rRNA score} = -1.657C_{\text{CFC}} - 95.847C_{\text{CFB}} + 11.602C_{\text{KS}} - 24.92C_{\text{EV}} - 9.897$$
$$N = 150 \quad R_c = 0.55 \quad U = 0.70 \quad p < 0.001 \quad \rho = 15.0 \qquad (6.5)$$

In all cases, we found statistically significant models, but the strength of the correlation between the input CIs and the classification of the sequences varies notably. In Table 3, we summarize the more important characteristics of the QSAR models presented above. For instance, the model based on distance-distribution function moments of the 250bp SSD-smCN present the higher regression coefficient R_c = 0.70. Conversely, the correlation for the model derived using thermodynamic parameters of FS-mmCNs is the lowest-found R_c = 0.34. These thermodynamic parameters of RNA folding are considered here as CIs taking into consideration their high dependence on the final connectivity of the RNA FS-mmCN. These CIs have been previously demonstrated to be useful in QSAR.[50] Interestingly, the two smCNs models have notably higher correlation than the other mmCNs models studied in this section. However, the model based on the CL-mmCNs electrostatic CIs is still the one with the higher regression coefficient R_c = 0.87 (see the previous section). Its high accuracy in recognition of the 18S rRNA could be the result of the codification of information about the connectivity of the nucleotides in the macromolecules of rRNAs (linear information) and the distribution of its bases purine and pyrimidine bases according its chemical structure (higher-order information). The 2D-RNA lattice also accounts for the content of purines (GA) and pyrimidine (TC) in the rRNA molecules, which can be observed in the

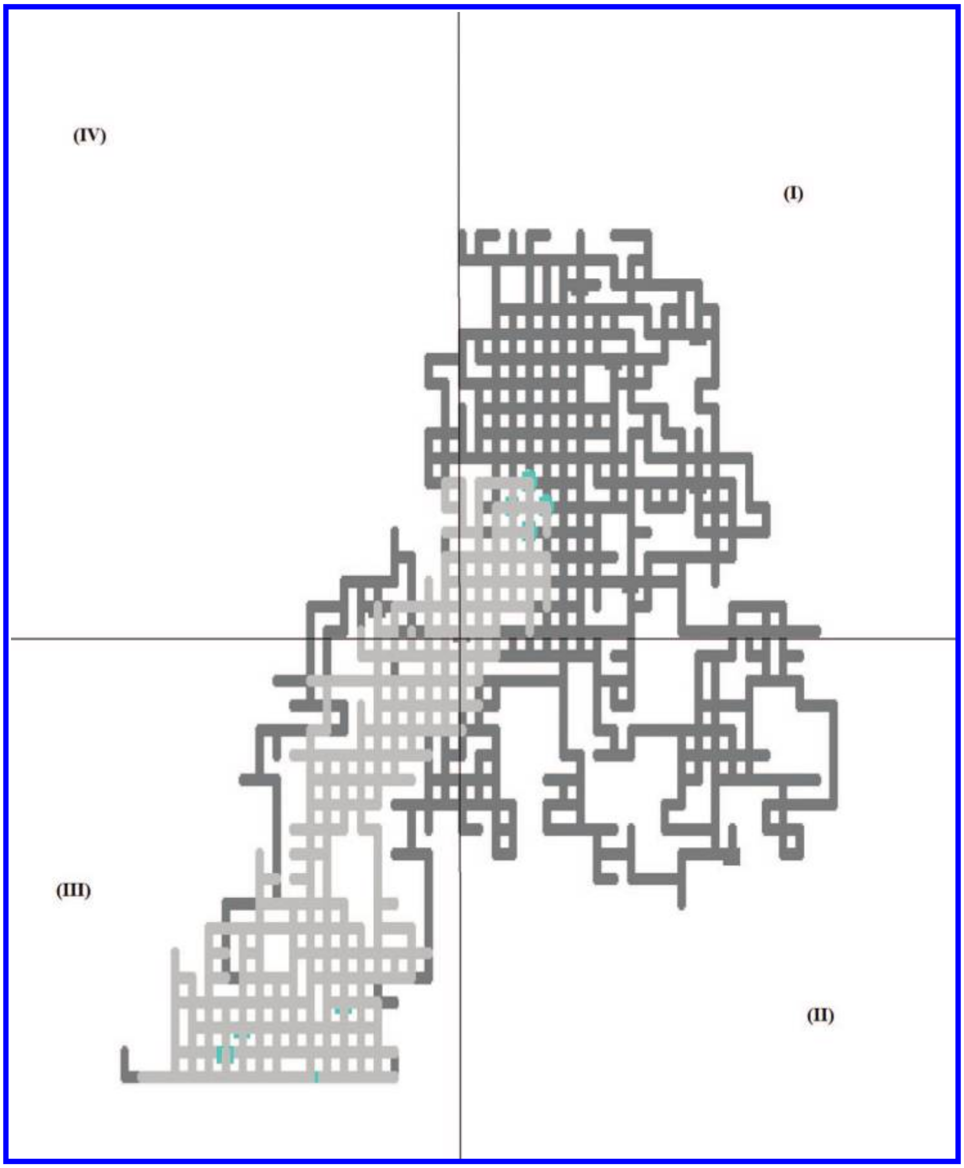


Figure 5. Superposition of CL-mmCNs for bacterial 16S-rRNAs (dark) vs plant 18S-rRNAs (light).

tendency of occupying certain quadrant in the Cartesian system. The variations in the content of nucleotides have been successfully applied in the recognition of non-protein-coding RNAs in genomes.[51] In this sense, we can conclude here that the quality of the final QSAR model depends strongly on the type of CI or CN selected but not necessarily on macromolecular or supramolecular nature of the CN used. This first comparative study of CIs of mmCNs versus smCNs is a very interesting and new field on Computational Chemistry and Bioinformatics, which have been identified by Bonchev[52] and others as an area that opens new opportunities to the integration of both molecular and supramolecular scientists.

3.2. QSAR Prediction of a New Experimental Outcome. *3.2.1. Isolation and QSAR Prediction of a Novel 18S rRNA.* We isolated, cloned, and sequenced a new 18S-rRNA fragment from *Psidium guajava* L. (GenBank accession number DQ400506). The protocol used for total RNA extraction successfully yielded a total RNA with suitable quality, no visual RNA degradation in electrophoresis, and the ratio of 260:280 was higher than 1.80. RT-PCR showed a single band around 300 bp (see Figure 2), which was amplified and later sequenced to produce a 272 bp fragment. Afterward, we analyzed our query sequence fragment using MARCH INSIDE methodology (MM) to predict with which probability this sequence may be annotated correctly as an 18S rRNA. We used only the QSAR model based on the electrostatic CIs of CL-mmCNs (model 1), because of the higher accuracy of this model compared with the others. This particular case was included in the validation subset to be predicted. We first constructed the CL-mmCNs for the sequence DQ400506 (see Figure 1 above), second calculated the values of ξ_1 and ξ_2, and third predicted the probability with the LDA model. The model successfully classified our sequence as rRNA with a high probability ($p = 0.999$).

3.2.2. Comparing RNA Networks, CIs, and BLAST Analysis for DQ400506. We have extended MM to identify RNA functions starting from sequence information using the MARCH INSIDE methodology. The method could be used as a complement to another tools based on sequence alignment and prediction of secondary structure. The importance of the present method with respect to BLAST, one of the more well-known alignment methods, is that we can obtain the same result but without reliance upon the alignment. In the Figure 3, we illustrate the results for BLASTn[53] analysis using DQ400506 as the query sequence to compare both methods and give a more accurate prediction of this sequence. In fact, our model classified the query sequence with a high value of probability (0.999, see previous section) and BLASTn analysis showed nucleotide identity of 99−100% with well-known sequences of the rRNA of plants. Consequently, both methods coincide in the annotation of DQ400506.

On the other hand, the present method has advantages with respect to RNA secondary structure methods based on FS-mmCNs, as is the case of model 3. One advantage is that we do not need to calculate several probable RNA structures and predict the biological function depending on the selection of the optimized structure based on folding energy (lowest ΔG). In addition, because of the superposition of several bp in the same nodes the CL-mmCNs used here present a notably lower number of nodes than the FS-mmCNs used to represent RNA secondary structures. As explained above, the FS-mmCNs is a graphic representation that assigns one node to each bp of a folded RNA secondary structure derived with optimization algorithms.[54] As a result, the matrices associated to the CL-mmCNs have a lower dimension, and QSAR prediction becomes faster than with FS-mmCNs. We illustrate this fact for the sequence DQ400506 in Figure 4.

3.3. Notes on CNs Approach to rRNAs Diversity.

3.3.1. Example Applied to the Diversity of rRNAs on Bacteria and Plants. Last, we would like to give some additional notes on the biological diversity of rRNAs expressed in terms of the present types of CNs. RNA is the predominant product of transcription, constituting some 80−90% of the total mass of cellular RNA in both eukaryotes and prokaryotes. Each ribosome is composed of two subunits containing rRNAs of different lengths, as well as a different set of proteins. These two subunits are named "large and small subunit", which contain two major rRNA molecules (23S and 16S rRNA in bacteria; 28S and 18S rRNA in eukaryotes, respectively) and a 5S rRNA. The large subunit of vertebrate ribosome also contains a 5.8S rRNA base-paired to the 28S rRNA. The same parts of each type of rRNA theoretically can form base-paired stem-loops, which would generate a similar three-dimensional structure for each rRNA in all organisms in spite of the considerable variability of the primary nucleotide sequences of these rRNAs.[1] In this sense, an important approach to analysis of the diversity of rRNAs is to compare the sequences of corresponding rRNAs in related organisms. Taking this perspective in view, it is natural to expect that new CNs introduced to study these types of RNAs have to describe the diversity of these biopolymers describing both similarity/dissimilarity regions. We would like to illustrate first the capacity of the smCNs introduced here to account for the above-mentioned diversity. Figure 5 depicts the superposition of several CL-smCNs known rRNA sequences isolated from bacteria and plants.

The figure illustrates that despite the existence of an overlapping or common zone above the center of coordinates of the Cartesian system many of the rRNAs of plants tend to dwell preferably at the third quadrant (III), the rRNAs of bacteria lie within the first quadrant (I), and quadrants II and IV are relatively unoccupied. Interestingly, we find similar behavior when plotting the 18S rRNA of parasites. In the Figure 6, we depict the superposition chart for parasites vs bacteria (section A) and parasite vs plants (section B). For this study, we used sequences of 18S rRNA from fish host parasites of the genus *Dactylogyrus*.[55] The chart illustrates that despite of having an overlapping or common zone above the center of coordinates of the Cartesian system many of the rRNAs tend to dwell preferably at the third quadrant (III) and the first quadrant (I), whereas quadrants II and IV are relatively unoccupied. However, the more important fact is that the representation accounts for rRNAs diversity differentiating the organisms with different phylogenetic relationships. It may become the TIs of CL-mmCns in an alignment-free alternative tool for future phylogenetic-property relationships. For instance, application to prediction of specificity of host−parasite relationships from RNA or protein sequences, which often rely upon phylogeny tree derived by sequence alignment.[56,57]

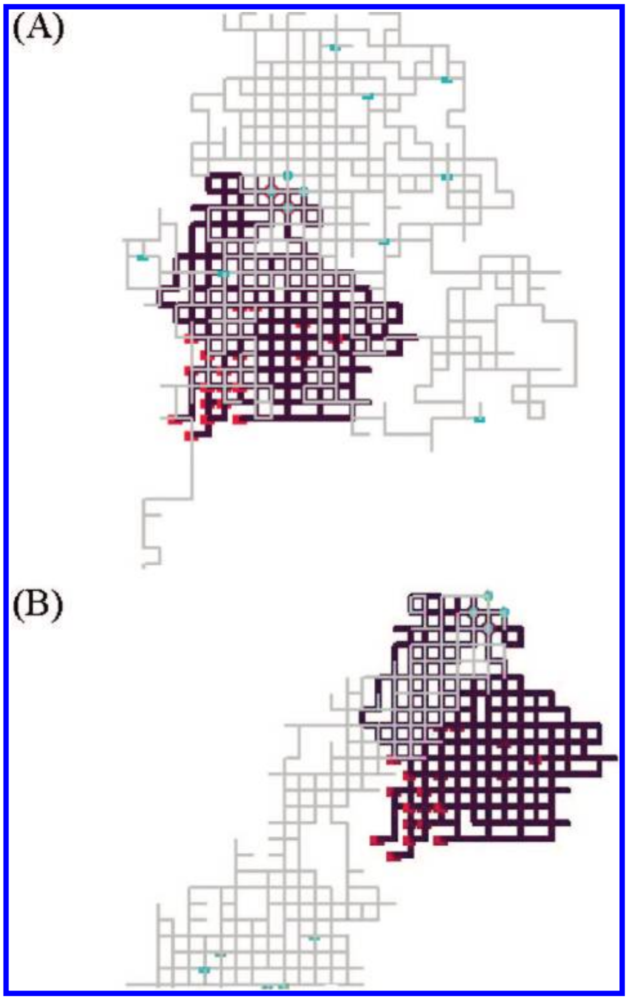
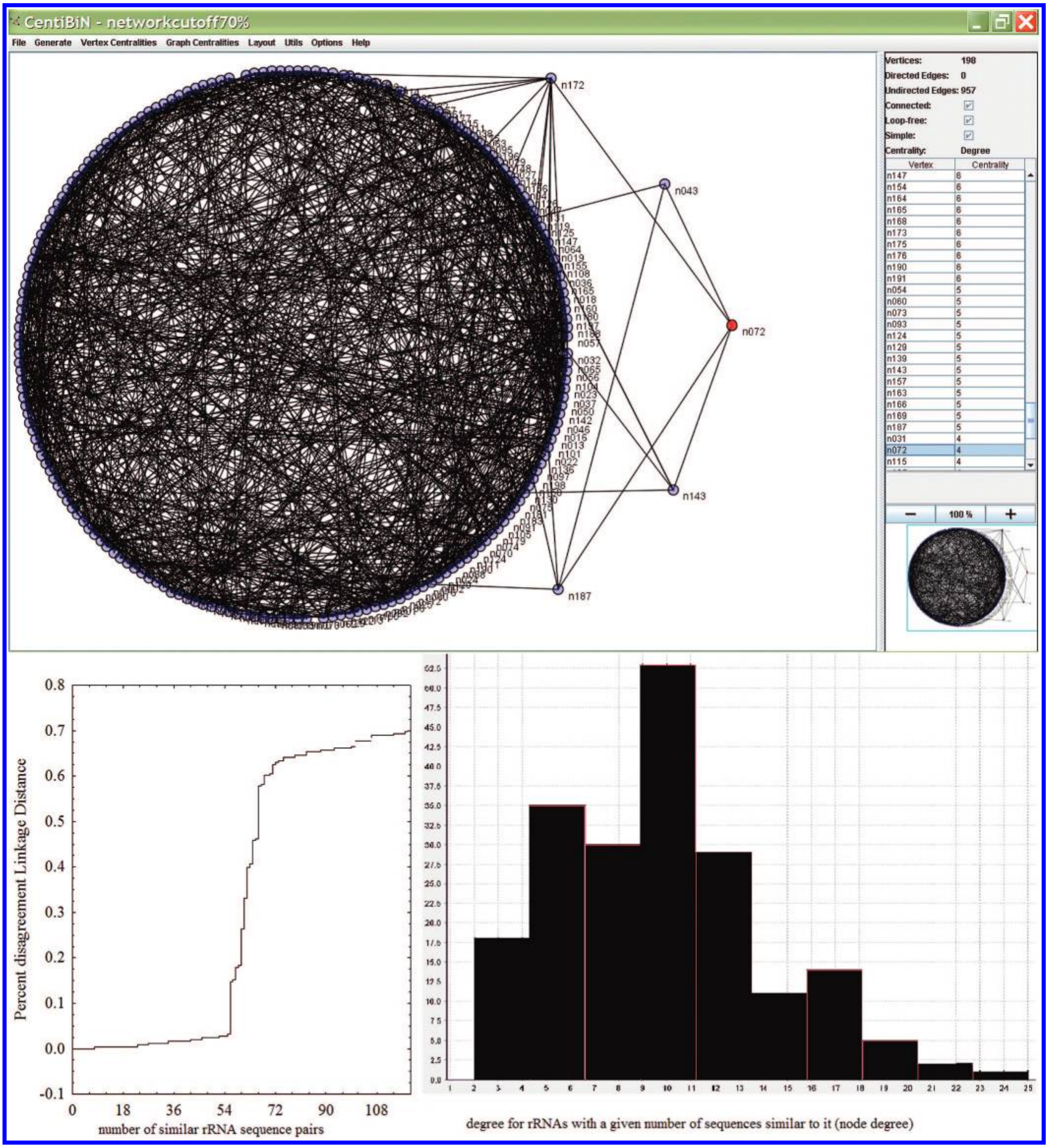


Figure 6. Superposition of CL-mmCNs for (A) parasite 18S-rRNAs (dark) vs bacterial 16S-rRNAs (light) and (B) parasite 18S-rRNAs (dark) vs plant 18S-rRNAs (light).

Last we would like to visualize the 250bp SSS-smCN for all the sequences used here to show how it accounts for the diversity on rRNAs similarity/dissimilarity regions. This network was introduced here and used to derive the QSAR model 5. This smCN is a simplified version of the SSD-smCN used to derive QSAR model 4 obtained after removal of pairs of sequences with low similarity. Notably these two


CentiBiN - networkcutoff70%
File Generate Vertex Centralities Graph Centralities Layout Utils Options Help
n172
n043
n072
n143
n187
− 100 % +
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
Percent disagreement Linkage Distance
0 18 36 54 72 90 108
number of similar rRNA sequence pairs
52.5 50.0 47.5 45.0 42.5 40.0 37.5 35.0 32.5 30.0 27.5 25.0 22.5 20.0 17.5 15.0 12.5 10.0 7.5 5.0 2.5 0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
degree for rRNAs with a given number of sequences similar to it (node degree)

Vertices: 198
Directed Edges: 0
Undirected Edges: 957
Connected: ☑
Loop-free: ☑
Simple: ☑
Centrality: Degree

Vertex	Centrality
n147	6
n154	6
n164	6
n165	6
n168	6
n173	6
n175	6
n176	6
n190	6
n191	6
n054	5
n060	5
n073	5
n093	5
n124	5
n129	5
n139	5
n143	5
n157	5
n163	5
n166	5
n169	5
n187	5
n031	4
n072	4
n115	4

Figure 7. (top) CentiBin representation of the SSS-smCN with sequence-agreement cutoff >30% for the first 250 bp. (bottom-left) Percent disagreement linkage distance vs linking steps. (bottom-right) Node degree distribution of the SSS-smCN.

models have also adequate accuracy values. It means that both CNs account for relevant information on sequences similarity/dissimilarity inherent to biological function. As can be noted the SSS-smCN could give us a quantitative approach to the total degree of diversity of one specific sequence or overall for all sequences. For it, we should inspect the superior part of the Figure 7 and note that, using the software CentiBin, we can click on a sequence and immediately withdrawn all sequences that are similar to it. In addition, in the bottom-left part, we can note that we need to link more than 100 pairs of sequences (linking steps) to reach similarity over 70% being the larger variation between 50 and 70 steps (change from 0−60%). Finally, in the bottom-right part of the Figure 7, we can see how it describes the diversity on the similarity of rRNAs inspecting the distribution of the degree of similarity (node degree) or number of sequences with a given number of similar sequences. Notably, this diversity resembles a normal distribution with the center or mean above 10 and extreme values moving from 2 to 25 (see Figure 7). Finally, both kind of rRNA CNs introduced here (mmCNs and smCNs) obey the necessary condition of accounting for the diversity of RNAs to codify information related to sequence−function relationships.

ACKNOWLEDGMENT

The authors are grateful to Editor Prof. Anton J. Hopfinger and unknown referees for their kind attention and useful comments. H.G.-D. acknowledges a research contract funded by the program Isidro Parga Pondal of the Xunta of Galicia. H.G.-D. also thanks Xunta of Galicia (IN89A2007/84-0) for funding of a postdoctoral position under supervision of K.-C.C. at Gordon Life Science Institute, San Diego, CA. This study was also supported by Grants PGIDIT05RAG20301PR from the Regional Ministry of Innovation and Industry (Consellería de Innovación e Industria, Xunta de Galicia, Spain) and AGL2006-13936-C02-01 from the Ministry of Education and Science (Ministerio de Educación y Ciencia, Spain). The authors also want to acknowledge the Project PTDC/BIA-BDE/69144/2006 by the Portuguese Foundation for Science and Technology (FCT).

Supporting Information Available: Tables of accession numbers and predicted scores for rRNA sequences and for no rRNA sequences with all models and sequence id's, groups, and ξ_1 and ξ_2 values. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lehninger, A. L. *Biochemistry*, 4th ed; 2005.

(2) Han, L.; Cui, J.; Lin, H.; Ji, Z.; Cao, Z.; Li, Y.; Chen, Y. Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* **2006**, *6* (14), 4023–37.

(3) 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.

(4) Chou, K. C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 246−255. (Erratum: **2001**, *44*, 601).

(5) Chou, K. C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10–19.

(6) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: 2002.

(7) Bonchev, D. Overall connectivity−A next generation molecular connectivity. *J. Mol. Graphics Modell.* **2001**, *20* (1), 65–75.

(8) Bonchev, D. The overall Wiener index−A new tool for characterization of molecular topology. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 582–92.

(9) Bonchev, D. Overall connectivities/topological complexities: a new powerful tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 934–41.

(10) Randic, M.; Balaban, A. T. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 532–9.

(11) 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*, 723–730.

(12) Randic, M.; Vracko, M. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 599–606.

(13) Nandy, A.; Basak, S. C. Simple numerical descriptor for quantifying effect of toxic substances on DNA sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (4), 915–9.

(14) 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.

(15) Ramos de Armas, R.; González-Díaz, H.; Molina, R.; Uriarte, E. Markovian backbone negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. *Proteins* **2004**, *56* (4), 715–23.

(16) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics* **2008**, *8*, 750–778.

(17) González-Díaz, H.; Aguero-Chapin, G.; Varona-Santos, J.; Molina, R.; Delogu, G.; Santana, L.; Uriarte, E.; Podda, G. 2D-RNA-coupling numbers: A new computational chemistry approach to link secondary structure topology with biological function. *J. Comput. Chem.* **2007**, *28* (6), 1049–56.

(18) Marrero-Ponce, Y.; Castillo-Garit, J. A.; Nodarse, D. Linear indices of the “macromolecular graph's nucleotides adjacency matrix” as a promising approach for bioinformatics studies. Part 1: Prediction of paromomycin's affinity constant with HIV-1 W-RNA packaging region. *Bioorg. Med. Chem.* **2005**, *13*, 3397–3404.

(19) Hoen, P. A.; Out, R.; Commandeur, J. N.; Vermeulen, N. P.; van Batenburg, F. H.; Manoharan, M.; van Berkel, T. J.; Biessen, E. A.; Bijsterbosch, M. K. Selection of antisense oligodeoxynucleotides against glutathione S-transferase Mu. *RNA* **2002**, *8* (12), 1572–83.

(20) Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. Protein linear indices of the “macromolecular pseudograph α-carbon atom adjacency matrix” in bioinformatics. Part 1: Prediction of protein stability effects of a complete set of alanine substitutions in Arc repressor. *Bioorg. Med. Chem.* **2005**, *13* (8), 3003–15.

(21) Bermudez, C. I.; Daza, E. E.; Andrade, E. Characterization and comparison of *Escherichia coli* transfer RNAs by graph theory based on secondary structure. *J. Theor. Biol.* **1999**, *197* (2), 193–205.

(22) Galindo, J. F.; Bermudez, C. I.; Daza, E. E. tRNA structure from a graph and quantum theoretical perspective. *J. Theor. Biol.* **2006**, *240* (4), 574–82.

(23) LaCount, D. J.; Vignali, M.; Chettier, R.; Phansalkar, A.; Bell, R.; Hesselberth, J. R.; Schoenfeld, L. W.; Ota, I.; Sahasrabudhe, S.; Kurschner, C.; Fields, S.; Hughes, R. E. A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* **2005**, *438* (3), 103–107.

(24) Carlson, M. R.; Zhang, B.; Fang, Z.; Mischel, P. S.; Horvath, S.; Nelson, S. F. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* **2006**, *7*, 40.

(25) Lu, C. H.; Jalbout, A. F.; Adamowicz, L.; Wang, Y.; Yin, C. S. Prediction of gas chromatographic retention indices of benzene dicarboxylic diesters using novel topological indices. *Bull. Environ. Contam. Toxicol.* **2006**, *77* (6), 793–8.

(26) Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Khan, K. M.; 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), 1370−1381.

(27) Junker, B. H.; Koschuetzki, D.; Schreiber, F. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* **2006**, *7* (1), 219.

(28) Koschützki, D. *CentiBiN*, version 1.4.2; 2006.

(29) Batagelj, V.; Mrvar, A. *Pajek*, version 1.15; 2006.

(30) Ludemann, A.; Weicht, D.; Selbig, J.; Kopka, J. PaVESy: Pathway visualization and editing system. *Bioinformatics* **2004**, *20* (16), 2841–4.

(31) Estrada, E. Virtual identification of essential proteins within the protein interaction network of yeast. *Proteomics* **2006**, *6* (1), 35–40.

(32) Platzer, A.; Perco, P.; Lukas, A.; Mayer, B. Characterization of protein interaction networks in tumors. *BMC Bioinformatics* **2007**, *8* (1), 224.

(33) Liu, D. W.; Kato, H.; Sugane, K. The nucleotide sequence and predicted secondary structure of small subunit (18S) ribosomal RNA from *Spirometra erinaceieuropaei*. *Gene* **1997**, *184* (2), 221–7.

(34) Abouheif, E.; Zardoya, R.; Meyer, A. Limitations of metazoan 18S rRNA sequence data: implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the Cambrian explosion. *J. Mol. Evol.* **1998**, *47* (4), 394–405.

(35) Benson, D. A.; Boguski, M. S.; Lipman, D. J.; Ostell, J.; Ouellette, B. F.; Rapp, B. A.; Wheeler, D. L. GenBank. *Nucleic Acids Res.* **1999**, *27* (1), 12–7.

(36) González-Díaz, H.; de Armas, R. R.; Molina, R. Markovian negentropies in bioinformatics. 1. A picture of footprints after the interaction of the HIV-1 Psi-RNA packaging region with drugs. *Bioinformatics* **2003**, *19* (16), 2079–87.

(37) Randic', M. Graphical representation of DNA as a 2-D map. *Chem. Phys. Lett.* **2004**, (386), 468–471.

(38) González-Díaz, H.; Uriarte, E.; Ramos de Armas, R. Predicting stability of Arc repressor mutants with protein stochastic moments. *Bioorg. Med. Chem.* **2005**, *13* (2), 323–31.

(39) González-Díaz, H.; Uriarte, E. Proteins QSAR with Markov average electrostatic potentials. *Bioorg. Med. Chem. Lett.* **2005**, *15* (22), 5088–94.

(40) Saiz-Urra, L.; González-Díaz, H.; Uriarte, E. Proteins Markovian 3D-QSAR with spherically-truncated average electrostatic potentials. *Bioorg. Med. Chem.* **2005**, *13* (11), 3641–7.

(41) González-Díaz, H.; Molina-Ruiz, R.; Hernandez, I. *MARCH-INSIDE*, version 3.0 (Markov chains invariants for simulation and design) 3.0; 2007. MARCH-INSIDE v3.0 Windows supported version available upon request to the main author. E-mail: gonzalezdiazh@yahoo.es.

(42) Markham, N. R.; Zuker, M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* **2005**, *33*, W577−W581.

(43) *STATISTICA for Windows*, release 6.0; Statsoft Inc.: 2001.

(44) Stewart, J.; Gill, L. *Econometrics*, 2nd ed.; Prentice Hall: London, 1998.

(45) Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; Li, W. Standardized multiple regression model. In *Applied Linear Statistical Models*, 5th ed.; McGraw Hill: New York, 2005; pp 271−277.

(46) Kowalski, R. B.; Wold, S. Pattern recognition in chemistry. In *Handbook of Statistics*; Krishnaiah, P. R.; Kanal, L. N. Eds.; North Holland Publishing Company: Amsterdam, 1982; pp 673−697.

(47) Wadsworth, G. J.; Redinbaugh, M. G.; Scandalios, J. G. A procedure for the small-scale isolation of plant RNA suitable for RNA blot analysis. *Anal. Biochem.* **1998**, *172* (1), 279–83.

(48) Marrero-Ponce, Y.; Khan, M. T.; Casanola Martin, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F.; Rotondo, R. Prediction of tyrosinase inhibition activity using atom-based bilinear indices. *ChemMedChem* **2007**, *2* (4), 449–478.

(49) Casanola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Ather, A.; Sultan, S.; Torrens, F.; Rotondo, R. TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: Evaluation of different classification model combinations using bond-based linear indices. *Bioorg. Med. Chem.* **2007**, *15* (3), 1483–503.

(50) González-Díaz, H.; Vilar, S.; Santana, L.; Podda, G.; Uriarte, E. On the applicability of QSAR for recognition of miRNA bioorganic structures at early stages of organism and cell development: embryo and stem cells. *Bioorg. Med. Chem.* **2007**, *15* (7), 2544–50.

(51) Schattner, P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.* **2002**, *30* (9), 2076–82.

(52) Bonchev, D.; Buck, G. A. From molecular to biological structure and back. *J. Chem. Inf. Model.* **2007**, *47* (3), 909–17.

(53) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.

(54) Mathews, D. H. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* **2005**, *21* (10), 2246–53.

(55) Simkova, A.; Morand, S.; Jobet, E.; Gelnar, M.; Verneau, O. Molecular phylogeny of congeneric monogenean parasites (*Dactylogyrus*): A case of intrahost speciation. *Evol. Int. J. Org. Evol.* **2004**, *58* (5), 1001–18.

(56) Simkova, A.; Verneau, O.; Gelnar, M.; Morand, S. Specificity and specialization of congeneric monogeneans parasitizing cyprinid fish. *Evol. Int. J. Org. Evol.* **2006**, *60* (5), 1023–37.

(57) Chou, K. C. Review: Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* **2004**, *11*, 2105–2134.

CI8001809