

Nucleic Acid Quadratic Indices of the “Macromolecular Graph’s Nucleotides Adjacency Matrix”. Modeling of Footprints after the Interaction of Paromomycin with the HIV-1 Ψ -RNA Packaging Region

Yovani Marrero Ponce ^{1,2,*}, Delvin Nodarse ¹, Humberto González Díaz ^{2,3}, Ronal Ramos de Armas ², Vicente Romero Zaldivar ⁴, Francisco Torrens ⁵ and Eduardo A. Castro ⁶

¹ Department of Pharmacy, Faculty of Chemical-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

² Department of Drug Design, Chemical Bioactive Center. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba

³ Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, 15706, Spain

⁴ Faculty of Informatics. University of Cienfuegos, Cienfuegos, Cuba

⁵ Institut Universitari de Ciència Molecular, Universitat de València, Dr. Moliner 50, E-46100 Burjassot (València), Spain

⁶ INIFTA, División Química Teórica, Suc.4, C.C. 16, La Plata 1900, Buenos Aires, Argentina

* Author to whom correspondence should be addressed; Fax: (+53)-42-281130/281455; Telephone: (+53)-42-281192/281473; E-mail: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es

Received: 28 January 2004; in revised form: 9 November 2004 / Accepted: 10 November 2004 / Published: 30 November 2004

Abstract: This report describes a new set of macromolecular descriptors of relevance to nucleic acid QSAR/QSPR studies, nucleic acids’ quadratic indices. These descriptors are calculated from the macromolecular graph’s nucleotide adjacency matrix. A study of the interaction of the antibiotic Paromomycin with the packaging region of the RNA present in type-1 HIV illustrates this approach. A linear discriminant function gave rise to excellent discrimination between 90.10% (91/101) and 81.82% (9/11) of interacting/non-interacting sites of nucleotides in training and test set, respectively. The LOO cross-validation procedure was used to assess the stability and predictability of the model.

Using this approach, the classification model has shown a LOO global good classification of 91.09%. In addition, the model's overall predictability oscillates from 89.11% until 87.13%, when n varies from 2 to 3 in leave- n -out jackknife method. This value stabilizes around 88.12% when n was > 3 . On the other hand, a linear regression model predicted the local binding affinity constants [$\log K (10^{-4}\text{M}^{-1})$] between a specific nucleotide and the aforementioned antibiotic. The linear model explains almost 92% of the variance of the experimental $\log K$ ($R = 0.96$ and $s = 0.07$) and LOO press statistics evidenced its predictive ability ($q^2 = 0.85$ and $s_{cv} = 0.09$). These models also permit the interpretation of the driving forces of the interaction process. In this sense, developed equations involve short-reaching ($k \leq 3$), middle-reaching ($4 < k < 9$) and far-reaching ($k = 10$ or greater) nucleotide's quadratic indices. This situation points to electronic and topologic nucleotide's backbone interactions control of the stability profile of Paromomycin-RNA complexes. Consequently, the present approach represents a novel and rather promising way to chem & bioinformatics research.

Keywords: Footprinting, Paromomycin, RNA HIV-1, TOMOCOMD-CANAR approach, Nucleic Acid Quadratic Index, QSPR/QSAR

Introduction

High throughput genome sequencing projects are producing an enormous amount of raw sequence data. All this data begs for methods that are able to synthesize the information into biological knowledge [1]. Public databases such as GenBank are growing in size at an exponential rate [2]. A significant proportion of the data corresponds to genomic sequences containing the structures not only of many genes but also of RNA.

The amount of new genome data has dramatically increased in recent years and it has once again brought to the forefront the question of protein and nucleic acid functions [3]. In this respect, the use of footprint techniques has proven to be an important experimental method for the discovery of significant processes in molecular biology and the field of genomics [4-8]. These experimental techniques permit quantitatively analyze D(R)Nase footprinting data for drugs interacting with D(R)NA obtaining apparent binding constants from the spot intensities appearing on the footprinting autoradiogram [9]. The study of the interactions of drugs with biomolecules is now the hot topic in modern bioinformatics. This kind study constitutes a significant step towards rational drug design.

The interactions between aminoglycosides and the packaging region of type-1 HIV (Human Immunodeficiency Virus) appear to represent a promising route for antiviral discoveries [10]. Aminoglycoside drugs are cationic natural products that interact with RNA [11]. The bactericidal effects inherent in these compounds stem from their ability to block protein synthesis by binding to the A-site on ribosomal RNA [12]. In fact, aminoglycoside analogues can be used to treat certain diseases. For example, the genetic information in human immunodeficiency virus and various tumour viruses is in the form of RNA [13]. Since the genomes of these viruses are likely to have unique structures, it may be possible to design agents that selectively block virus proliferation by targeting a specific site on RNA [14].

One of the present authors has recently introduced the novel computer-aided molecular design scheme *TOMOCOMD* (acronym of *TO*pological *MO*lecular *COM*puter *D*esign). It calculates several new 2D/3D families of total and local (atom and atom-type) topologic and stochastic molecular descriptors, such as quadratic and linear indices; defined by analogy with the quadratic and linear mathematical maps [15, 16]. This point of view was very recently successfully applied to the prediction of physical properties and Caco-2 permeability of organic compounds and drugs, respectively [15-18]. Interestingly, molecular quadratic indices can be generalized to allow the codification of 3D-structural features [19].

Therefore, describing an extended TOMOCOMD-CANAR approach to account for RNA structure constitutes the main aim of this paper. In the present study, we propose a total and local definition of nucleic acid quadratic indices of the “macromolecular graph’s nucleotides adjacency matrix”. The other objective of the present work focused on deriving quantitative structure property relationships to predict the probability and the affinity with which paromomycin bind to the HIV-1 Ψ -RNA packaging region.

Materials and methods

Computational Methods

A nucleic acid is a long, unbrached polynucleotide – that is, a polymer consisting of nucleotides. Each nucleotide has the three following components: 1) A cyclic five-carbon sugar, 2) a purine or pyrimidine base attached to the 1'-carbon atom of sugar by N-glycoside bond, and 3) A phosphate attached to the 5'-carbon of the sugar by a phosphoester linkage. The nucleotides in nucleic acids are covalently linked by a second phosphoester bond that joins the 5'-phosphate of one nucleotide and the 3'-OH group of the adjacent nucleotides. The purine and pyrimidine bases are not engaged in any covalent bonds to each other. Thus, a polynucleotide consists of an alternating sugar-phosphate backbone and each nucleotide is characterized by the base attached to it, which can be either adenine (*A*), cytosine (*C*), guanine (*G*) or thymine (*T*) [RNA molecule contains the base uracil (*U*) instead of *T*]. Consequently, a RNA molecule is uniquely determined by the sequence of bases along its chain, and it has a definite orientation [20-23].

In particular, a typical RNA is the single-stranded polyribonucleotide. This macromolecule has a folded 3D conformation that is held together in part by noncovalent base-pairing interactions like those that hold together the two stands of the DNA helix. In the single-stranded RNA molecule, however, the complementary bases pairs form between nucleotides residues in the same chain, which causes the RNA molecule to fold up in a unique way that is important for its biochemical activity. In this sense, the RNA structure contains several sets of unpaired nucleotide residues. Most of the weak interactions (hydrogen bonds) form between Watson-Crick complementary bases (between pairs of non-consecutive bases), i.e., between *A* and *U* and between *C* and *G*, but a far from negligible amount of bonds also form between other pairs of bases, as for example the *G*·*U* wobble pairs [20-23].

On the other hand, the general principles of the molecular quadratic indices of the “molecular pseudograph’s atom adjacent matrix” for small-to-medium sized organic compounds have been explained in some detail elsewhere [15-19]. However, this work gives an extended overview of this approach.

First, in analogy to the molecular vector X used to represent organic molecules, we introduce here the macromolecular vector (X_m). The components of this vector are numeric values, which represent a certain nucleotide residues (DNA-RNA bases) properties. These properties characterize each kind of nucleotides (purine and pyrimidine bases) within the nucleic acid, because the only uncommon part of these nucleotides is these bases. Such properties can be experimental molar absorption coefficient ϵ_{260} at 260 nm and PH = 7.0, first (ΔE_1) and second (ΔE_2) single excitation energies in eV, and first (f_1) and second (f_2) oscillator strength values (of the first singlet excitation energies) of the nucleotide DNA-RNA bases, and so on [24]. For instance, the $f_{1(B)}$ property of the DNA-ARN bases B takes the values $f_{1(A)} = 0.28$ for adenine, $f_{1(G)} = 0.20$ for guanine, $f_{1(U)} = 0.18$ for uracil and so on [24]. Table 1 depicts nucleotides (bases) descriptors properties for the DNA-RNA bases.

Table 1. Five properties of DNA-RNA bases using as labels to characterize each nucleotide. Experimental molar absorption coefficient ϵ_{260} at 260 nm and pH=7.0, first (ΔE_1) and second (ΔE_2) single excitation energies in eV, and first (f_1) and second (f_2) oscillator strength values (of the first singlet excitation energies) of the nucleotide DNA-RNA bases [24].

Purine and pyrimidine bases (RNA/ADN)	f_1	f_2	$\epsilon_{260}/1000$	ΔE_1	ΔE_2
Adenine (A)	0.28	0.54	15.4	4.75	5.99
Guanine (G)	0.20	0.27	11.7	4.49	5.03
Uracil (U)	0.18	0.3	9.9	4.81	6.11
Thymine (T)	0.18	0.37	9.2	4.67	5.94
Cytosine (C)	0.13	0.72	7.5	4.61	6.26

Thus, a RNA having 5, 10, 15, ..., n nucleotides can be represented by means of vectors, with 5, 10, 15, ..., n components, belonging to the spaces \mathfrak{R}^5 , \mathfrak{R}^{10} , \mathfrak{R}^{15} , ..., \mathfrak{R}^n , respectively. Where n is the dimension of these real sets (\mathfrak{R}^n). This approach allows us encoding RNA sequences such as AGUCACGUA through out the macromolecular vector $X_m = [0.28, 0.20, 0.18, 0.13, 0.28, 0.13, 0.20, 0.18, 0.28]$, in the f_1 -scale (see Table 1). This vector belongs to the product space \mathfrak{R}^9 . The use of other AND-ARN bases properties defines alternative macromolecular vectors.

For a given nucleic acid composed of nucleotides (*vector of \mathfrak{R}^n*), the “macromolecular vector” (X_m) is constructed and the k^{th} nucleic acid’s total quadratic indices, $q_k(x_m)$ are calculated as quadratic forms as shown in Eq. 1:

$$q_k(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^k a_{ij} {}^m X_i {}^m X_j \quad (1)$$

where, ${}^k a_{ij} = {}^k a_{ji}$ (symmetric square matrix), n is the number of nucleotides of the nucleic acid, and ${}^m X_1, \dots, {}^m X_n$ are the coordinates or components of the macromolecular vector (X_m) in a system of canonical basis vectors of \mathfrak{R}^n . In this case, the canonical (‘natural’) base of \mathfrak{R}^n $\{e_1, \dots, e_n\}$ is used as the form’s base. Thereafter, the coordinates of any vector X_m coincide with the components of this vector. For that reason, such coordinates can be considered as weights of the vertices (ADN-ARN

bases) of the graph of the nucleic acid's backbone. The coefficients $^k a_{ij}$ are the elements of the k^{th} power of the macromolecular matrix $M(G_m)$ of the nucleic acid's graph (G_m). Here, $M(G_m) = [a_{ij}]$, where n is the number of bases (nucleotides) in sugar-phosphate's backbone. The elements a_{ij} are defined as follows:

$$a_{ij} = P_{ij} \text{ if } i \neq j \text{ and } e_k \in E(G_m) \\ = 0 \text{ otherwise} \tag{2}$$

Table 2. A close up to the mathematical definition of total (RNA fragment) and local (nucleotide) nucleic acid quadratic indices of the “macromolecular graph’s nucleotide adjacency matrix” of a RNA fragment.

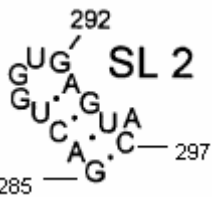
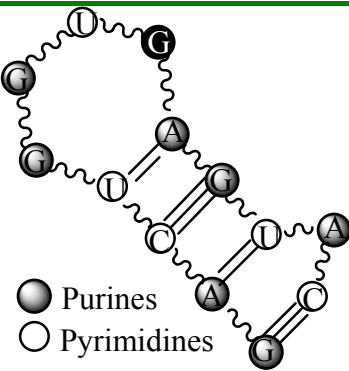
 <p>Secondary structure of an RNA fragment of the SL 2 motif (see Figure 1)</p>	 <p>● Purines ○ Pyrimidines</p> <p>Macromolecular graph's: an undirected graph with multiple edges G_m</p>	$X_m = [G A C U G G U G A G U A C]; X_m \in \mathfrak{R}^{13}$																																																																																																																																																																																																			
$q_0(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^0 a_{ij}^m X_i^m X_j^m = [{}^m X]^t M^0(G_m) [{}^m X] = 0.5662$	<p>In the definition of X_m, as macromolecular vector, the symbol of the bases is used to indicate the corresponding AND-RNA bases property, for instance, f_1. That is: if we write A it means $f_{1(A)}$, adenine first oscillator strength values or some bases property, which characterizes each nucleotide in the nucleic acid molecule. So, if we use the canonical bases of \mathfrak{R}^{13}, the coordinates of any macromolecular vector X_m coincide with the components of that macromolecular vector.</p>	$[X_m]^t = [0.20 \ 0.28 \ 0.13 \ 0.18 \ 0.20 \ 0.20 \ 0.18 \ 0.20 \ 0.28 \ 0.20 \ 0.18 \ 0.28 \ 0.13]$																																																																																																																																																																																																			
$q_1(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^1 a_{ij}^m X_i^m X_j^m = [{}^m X]^t M^1(G_m) [{}^m X] = 1.7124$	<p>$[X_m]^t$: Transposed of $[X_m]$ and it means the vector of the coordinates of X_m in Canonical base of \mathfrak{R}^{13} (a row matrix)</p>	<p>$[X_m]$: vector of the coordinates of X_m in Canonical base of \mathfrak{R}^{13} (a columns matrix)</p>																																																																																																																																																																																																			
$q_2(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^2 a_{ij}^m X_i^m X_j^m = [{}^m X]^t M^2(G_m) [{}^m X] = 6.7533$	<table border="1" data-bbox="882 1462 1417 1955"> <thead> <tr> <th></th> <th>G</th> <th>A</th> <th>C</th> <th>U</th> <th>G</th> <th>G</th> <th>U</th> <th>G</th> <th>A</th> <th>G</th> <th>U</th> <th>A</th> <th>C</th> </tr> </thead> <tbody> <tr> <th>\underline{G}_{285}</th> <td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>3</td> </tr> <tr> <th>\underline{A}_{286}</th> <td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{C}_{287}</th> <td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>3</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{U}_{288}</th> <td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{G}_{289}</th> <td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{G}_{290}</th> <td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{U}_{291}</th> <td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{G}_{292}</th> <td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{A}_{293}</th> <td>0</td><td>0</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{G}_{294}</th> <td>0</td><td>0</td><td>3</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td><td>0</td> </tr> <tr> <th>\underline{U}_{295}</th> <td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td><td>0</td> </tr> <tr> <th>\underline{A}_{296}</th> <td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td><td>1</td> </tr> <tr> <th>\underline{C}_{297}</th> <td>3</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td> </tr> </tbody> </table>		G	A	C	U	G	G	U	G	A	G	U	A	C	\underline{G}_{285}	0	1	0	0	0	0	0	0	0	0	0	0	3	\underline{A}_{286}	1	0	1	0	0	0	0	0	0	0	2	0	0	\underline{C}_{287}	0	1	0	1	0	0	0	0	0	3	0	0	0	\underline{U}_{288}	0	0	1	0	1	0	0	0	2	0	0	0	0	\underline{G}_{289}	0	0	0	1	0	1	0	0	0	0	0	0	0	\underline{G}_{290}	0	0	0	0	1	0	1	0	0	0	0	0	0	\underline{U}_{291}	0	0	0	0	0	1	0	1	0	0	0	0	0	\underline{G}_{292}	0	0	0	0	0	0	1	0	1	0	0	0	0	\underline{A}_{293}	0	0	0	2	0	0	0	1	0	1	0	0	0	\underline{G}_{294}	0	0	3	0	0	0	0	0	1	0	1	0	0	\underline{U}_{295}	0	2	0	0	0	0	0	0	0	1	0	1	0	\underline{A}_{296}	0	0	0	0	0	0	0	0	0	0	1	0	1	\underline{C}_{297}	3	0	0	0	0	0	0	0	0	0	0	1	0
	G	A	C	U	G	G	U	G	A	G	U	A	C																																																																																																																																																																																								
\underline{G}_{285}	0	1	0	0	0	0	0	0	0	0	0	0	3																																																																																																																																																																																								
\underline{A}_{286}	1	0	1	0	0	0	0	0	0	0	2	0	0																																																																																																																																																																																								
\underline{C}_{287}	0	1	0	1	0	0	0	0	0	3	0	0	0																																																																																																																																																																																								
\underline{U}_{288}	0	0	1	0	1	0	0	0	2	0	0	0	0																																																																																																																																																																																								
\underline{G}_{289}	0	0	0	1	0	1	0	0	0	0	0	0	0																																																																																																																																																																																								
\underline{G}_{290}	0	0	0	0	1	0	1	0	0	0	0	0	0																																																																																																																																																																																								
\underline{U}_{291}	0	0	0	0	0	1	0	1	0	0	0	0	0																																																																																																																																																																																								
\underline{G}_{292}	0	0	0	0	0	0	1	0	1	0	0	0	0																																																																																																																																																																																								
\underline{A}_{293}	0	0	0	2	0	0	0	1	0	1	0	0	0																																																																																																																																																																																								
\underline{G}_{294}	0	0	3	0	0	0	0	0	1	0	1	0	0																																																																																																																																																																																								
\underline{U}_{295}	0	2	0	0	0	0	0	0	0	1	0	1	0																																																																																																																																																																																								
\underline{A}_{296}	0	0	0	0	0	0	0	0	0	0	1	0	1																																																																																																																																																																																								
\underline{C}_{297}	3	0	0	0	0	0	0	0	0	0	0	1	0																																																																																																																																																																																								
$q_3(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^3 a_{ij}^m X_i^m X_j^m = [{}^m X]^t M^3(G_m) [{}^m X] = 25.3806$	<p>$M^1(G_m)$: Macromolecular graph's nucleotide Adjacency Matrix</p>																																																																																																																																																																																																				
$q_4(x_m) = \sum_{i=1}^n \sum_{j=1}^n {}^4 a_{ij}^m X_i^m X_j^m = [{}^m X]^t M^4(G_m) [{}^m X] = 105.5649$																																																																																																																																																																																																					

Table 2. Cont.

Nucleotide (N)	$q_{0L}(x_m, N)$	$q_{1L}(x_m, N)$	$q_{2L}(x_m, N)$	$q_{3L}(x_m, N)$	$q_{4L}(x_m, N)$
G285	0.04	0.134	0.666	2.154	9.654
A286	0.0784	0.1932	1.0668	3.5112	17.2256
C287	0.0169	0.1378	0.5369	2.8223	10.1634
U288	0.0324	0.1602	0.5328	2.0844	8.9226
G289	0.04	0.076	0.254	0.748	2.738
G290	0.04	0.076	0.156	0.422	1.136
U291	0.0324	0.072	0.1512	0.3492	1.0872
G292	0.04	0.092	0.232	0.786	2.8
A293	0.0784	0.2128	0.8652	3.3768	12.6308
G294	0.04	0.17	0.996	3.604	18.342
U295	0.0324	0.1872	0.4572	2.6136	8.6328
A296	0.0784	0.0868	0.5376	1.3608	7.4004
C297	0.0169	0.1144	0.3016	1.5483	4.8321
ARN fragment	0.5662	1.7124	6.7533	25.3806	105.5649

where, $E(G_m)$ represents the set of edges of G_m and P_{ij} is the number of edges among the vertices (nucleotides) v_i and v_j . In this adjacency matrix $M(G_m)$ the row i and column i correspond to vertex v_i from G_m . The element a_{ij} of this matrix represents a bond between a nucleotide i and other j . Here, we consider only covalent interaction (phosphodiester bond) and hydrogen bond interaction (between complementary bases). As a first approximation, we considered both interactions equivalent. The matrix $M^k(G_m)$ provides the number of walks of length k linking the nucleotides i and j .

Equation (1) for $q_k(x_m)$ can be written as the single matrix equation:

$$q_k(x_m) = [{}^mX]^t M^k(G_m) [{}^mX] \quad (3)$$

where $[{}^mX]$ is a column vector (a $nx1$ matrix), $[{}^mX]^t$ the transpose of $[{}^mX]$ (a $1xn$ matrix) and $M^k(G_m)$ the k^{th} power of the matrix $M(G_m)$ of the macromolecular pseudograph G_m (mathematical quadratic form's matrix). Table 2 exemplifies the calculation of $q_k(x_m)$ for a secondary structure RNA fragment.

In addition to total quadratic indices, computed for the whole-macromolecule, local-fragment (nucleotide and nucleotide-type) formalisms can be developed. These descriptors are termed local nucleic acid's quadratic indices, $q_{kL}(x_m)$. The definition of these descriptors is as follows:

$$q_{kL}(x_m) = \sum_{i=1}^m \sum_{j=1}^m {}^k a_{ijL} {}^m X_i {}^m X_j \quad (4)$$

where m is the number of nucleotides of the fragment of interest and ${}^k a_{ijL}$ is the element of the file i and column j of the matrix $M^k_L(G_m)$. This matrix is extracted from $M^k(G_m)$ and contains information referred to the vertices of the specific nucleic acid fragments (F_R) and also of the molecular environment. The matrix $M^k_L(G_m) = [{}^k a_{ijL}]$ with elements ${}^k a_{ijL}$ is defined as follows:

$$\begin{aligned} {}^k a_{ijL} &= {}^k a_{ij} \text{ if both } v_i \text{ and } v_j \text{ are vertices (nucleotides) contained within } F_R \\ &= 1/2 {}^k a_{ij} \text{ if } v_i \text{ or } v_j \text{ are contained within } F_R \\ &= 0 \text{ otherwise} \end{aligned} \quad (5)$$

where, the ${}^k a_{ij}$ are the elements of the k^{th} power of $M(G_m)$. These local analogues can also be expressed in matrix form by the expression:

$$q_{kL}(x_m) = [{}^mX]^t M^k_L(G_m) [{}^mX] \quad (6)$$

Note that for any partition of a nucleic acid into Z macromolecular fragments there will be Z local macromolecular-fragment matrices. That is to say, if a nucleic acid is partitioned into Z macromolecular fragments, the matrix $M^k(G_m)$ can be partitioned into Z local matrices $M^k_L(G_m)$, $L = 1, \dots, Z$. The k^{th} power of the matrix $M(G_m)$ is exactly the sum of the k^{th} power of the local Z matrices,

$$M^k(G_m) = \sum_{L=1}^Z M^k_L(G_m) \quad (7)$$

In the same way, $M^k(G_m) = [{}^k a_{ij}]$ where,

$${}^k a_{ij} = \sum_{L=1}^Z {}^k a_{ijL} \quad (8)$$

and the total nucleic acid's quadratic indices are the sum of the macromolecular quadratic indices of the Z molecular fragments (see Table 2),

$$q_k(x_m) = \sum_{L=1}^Z q_{kL}(x_m) \quad (9)$$

Any local nucleic acid's quadratic index has a particular meaning, especially for the first values of k , where the information about the structure of the fragment F_R is contained. Higher values of k relate to the environment information of the fragment F_R considered within the macromolecular graph (G_m). In any case, a complete series of indices performs a specific characterization of the chemical structure. The generalization of the matrices and descriptors to "superior analogues" is necessary for the evaluation of situations where only one descriptor is unable to bring a good structural characterization [25]. The local macromolecular indices can also be used together with total ones as variables for QSAR/QSPR (Quantitative Structure-Activity/Structure Relationship) modeling for properties or activities that depend more on a region or a fragment than on the macromolecule as a whole.

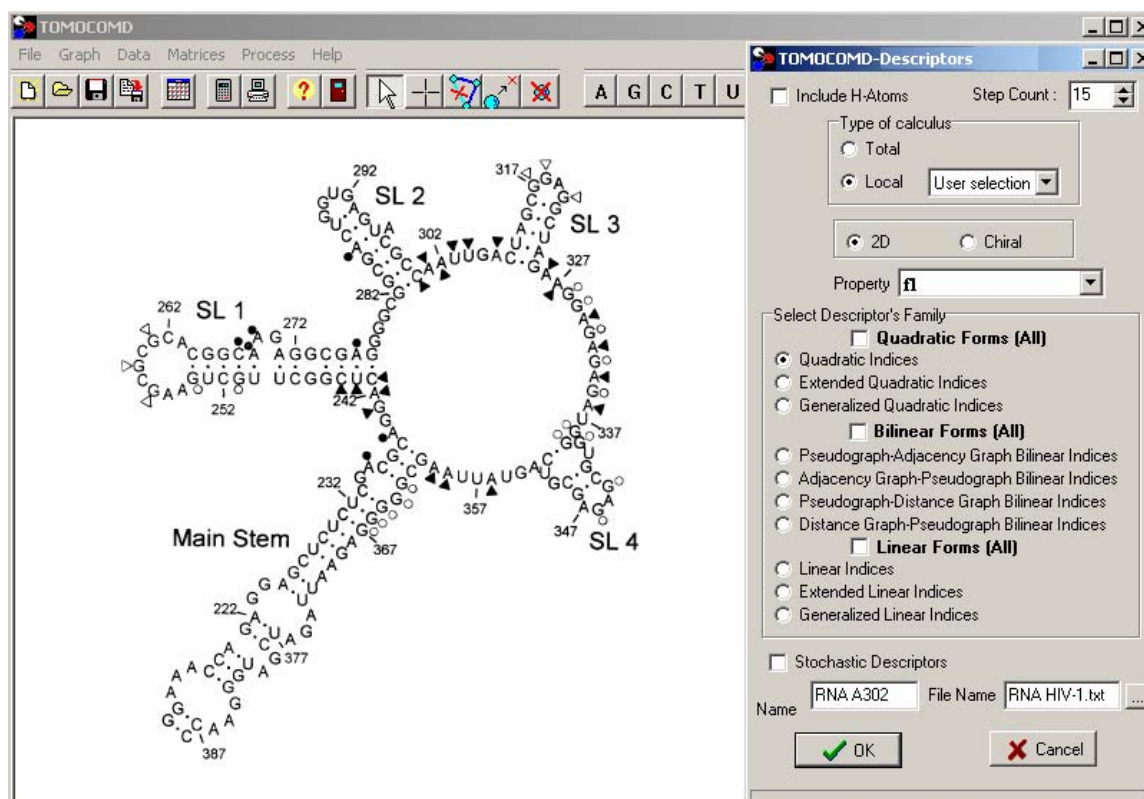
Footprinting Data

The data set of footprinted and binding nucleotides was extracted from the literature [9]. Figure 1 depicts the secondary structure of the HIV-1 Ψ -RNA packaging region as well as the binding sites of Paromomycin. A representation of the Ψ -RNA appears along with a summary of binding/enhancement information for Paromomycin. The RNA consists of the 'main stem', positions 213–238 and 361–388; SL-1, which contains the dimmer initiation site; SL-2, having the 5' splice donor site; SL-3, and SL-4, the latter contains the start codon (AUG) for the *gag* gene.

TOMOCOMD-CANAR Software

TOMOCOMD is an interactive program for molecular design and bioinformatics research [26]. The program is composed by four subprograms, each one of them dealing with drawing structures (drawing mode) and calculating 2D and 3D molecular descriptors (calculation mode). The modules are named CARDD (Computed-Aided 'Rational' Drug Design), CAMPS (Computed-Aided Modeling in Protein Science), CANAR (Computed-Aided Nucleic Acid Research) and CABPD (Computed-Aided Bio-Polymers Docking).

Figure 1. HIV-1 Ψ -RNA packaging region represented on the *TOMOCOMD-CANAR* interface. Nucleotides involved in binding and enhancement (structural changes) for RNase I are shown as filled circles and triangles, respectively (open symbols indicates the use of RNase T1).



In this paper we outline salient features concerning with only one of these subprograms: CANAR. This subprogram bases on a user-friendly philosophy without *prior* knowledge of programming skills. The calculation of total and local macromolecular quadratic indices for any nucleic acids was implemented in the *TOMOCOMD-CANAR* software [26]. The following list briefly resumes the main steps for the application of this method in QSAR/QSPR:

1. Draw the macromolecular graphs (G_m) for each RNA/ADN of the data set, using the software's drawing mode. Selection of the active nucleotide symbol carries out this procedure. Here, we consider only covalent interaction (phosphodiester bond) and hydrogen bond interaction (between complementary bases).

2. Use appropriated purine and pyrimidine bases weights in order to differentiate the residues in each nucleotide. This work uses as nucleotide weights five properties of DNA-RNA bases (see Table 1) [24]. This parametrization is done using the properties of U, T, A, G, and C only, because the only uncommon part of these nucleotides are these bases.

3. Compute the nucleic acid quadratic indices of the "macromolecular graph's nucleotides adjacency matrix". They can be performed in the software calculation mode, which you can select the DNA-RNA bases properties and the family descriptor previously to calculate the macromolecular indices. This software generates a table in which the rows and columns correspond to the compounds and the $q_k(x_m)$, respectively.

4. Find a QSPR/QSAR equation by using statistical techniques, such as multilinear regression analysis (MRA), Neural Networks (NN), Linear Discrimination Analysis (LDA), and so on. That is to say, we can find a quantitative relation between a property P and the $q_k(x_m)$ having, for instance, the following appearance,

$$P = a_0q_0(x_m) + a_1q_1(x_m) + a_2q_2(x_m) + \dots + a_kq_k(x_m) + c \quad (10)$$

Where P is the measurement of the property, $q_k(x_m)$ [or $q_{kL}(x_m)$] is the k^{th} total [or local] macro-molecular quadratic indices, and the a_k 's are the coefficients obtained by the statistical analysis.

5. Test the robustness and predictive power of the QSPR/QSAR equation by using internal and external cross-validation techniques,

6. Develop a structural interpretation of the obtained QSAR/QSPR model using macromolecular quadratic indices as molecular descriptors.

Statistical Analysis

Based on the discussion above, two simple linear models were proposed to either discriminate between footprinted and interacting (binding) nucleotides or to predict drug–nucleotide affinity. Linear Discrimination Analysis (LDA) and Linear Multiple Regression (LMR) were used to obtain quantitative models, respectively. These statistical analyses were carried out with the STATISTICA software package [27]. *TOMOCOMD-CANAR* model used for both statistical procedures the first 10 $q_{kL}(x_m)$ [from $q_{0L}(x_m)$ to $q_{9L}(x_m)$] for each nucleotides in RNA.

Forward stepwise was fixed as the strategy for variable selection. The tolerance parameter (proportion of variance that is unique to the respective variable) used was the default value for minimum acceptable tolerance, which is 0.01.

LDA is used in order to generate the classifier function on the basis of the simplicity of the method [28]. To test the quality of the discriminant functions derived we used the Wilks' λ and the Mahalanobis distance. The Wilks' λ statistic for overall discrimination can take values in the range of 0 (perfect discrimination) to 1 (no discrimination). The Mahalanobis distance indicates the separation of the respective groups. It shows whether the model possesses an appropriate discriminatory power for differentiating between the two respective groups. The classification of cases was performed means of the posterior classification probabilities, which is the probability that the respective case belongs to a particular group, i.e., footprinted or interacting (binding) nucleotides (see Figure 1). In developing this classification function the values of -1 and 1 were assigned to these groups, respectively. The quality of the ADL model also was determined by examining the percentage of good classification and the proportion between the cases and variables in the equation. Validation of the discriminant function was corroborated by means of leave- n -out cross validation procedures.

In addition, external prediction (test) sets assess the robustness and predictive power of the found model. This type of model validation is very important, if we take into consideration that the predictive ability of a QSAR model can only be estimated using an external test set of compounds that was not used for building the model [29,30]. The quality of the LMR model was determined examining the statistic parameters of multivariable comparison of regression and cross-validation procedures. In this sense, the quality of models was determined by examining the regression coefficients (R), determination coefficients (R^2), Fisher ratio's p -level [$p(F)$], standard deviations of the regression (s) and the leave-one-out (LOO) press statistics (q^2 , s_{cv}) [30]. In recent years, the LOO press statistics (e.g.,

q^2) have been used as a means of indicating predictive ability. Many authors consider high q^2 values (for instance, $q^2 > 0.5$) as indicator or even as the ultimate proof of the high predictive power of a QSAR model.

Results and Discussion

Development of the Discrimination Function: Local (Nucleotide) quadratic indices and the probability of footprinting after RNA-Paromomycin interaction.

The best equation found to discriminate between footprinted and binding nucleotides was:

$$\mathbf{Binding} = 1.10836 + 93.6133^{\mathbf{fl}} q_{0L}(x_m) - 5.4682^{\mathbf{fl}} q_{3L}(x_m) + 0.1356^{\mathbf{fl}} q_{5L}(x_m) \quad (11)$$

$$N = 101 \quad \lambda = 0.43 \quad D^2 = 6.0 \quad F(3.97) = 43.342 \quad \rho = 10.1 \quad p < 0.000$$

where N is the number of nucleotides, λ is the Wilks's statistic, D^2 is the squared Mahalanobis distance, F is the Fisher ratio and p is the p-level (probability of error). The coefficient ρ was used to control the ratio of the adjustable parameters in the model with respect to the number of variables [31]. These statistics indicate that model (11) is appropriate for the discrimination of footprinted and non-footprinted nucleotides studied here. It classifies correctly 95.52% (61/64) of footprinted nucleotides and 79.41% (20/27) of binding nucleotides in training set, for a global good classification of 90.10% (91/101). In Table 3 we give the classification of nucleotides in training set together with their posterior probabilities calculated from the Mahalanobis distance.

LOO cross-validation procedure assessed the predictability of the model obtained by LDA. This methodology systematically removed one data point at a time from the data set. A QSAR model was then constructed on the basis of this reduced data set and subsequently used to predict the removed data point. This procedure was repeated until a complete set of predicted was obtained. Using this approach, the model (11) has shown a LOO global good classification of 91.09%.

Secondly, to assess the predictability of the classification model (11), a leave- n -out cross-validation was performed. This model shown an 89.11 and 87.13% of global good classification when n varied from 2 to 3 in the leave- n -out cross validation procedures. The model stabilizes around 88.12% when n was > 3 (see Figure 2).

The most important criterion for the acceptance or not of a discriminant model, such model (11), bases on the statistics for the test set. Equation 11 classifies correctly 81.82% (9/11) of both drug-interacting nucleotides and footprinted ones. In Table 4, we give the classification of nucleotides in test set. If we considered the data set and the test set (*full set*) the percentage of good classification was 88.62% (109/121).

Local (Nucleotide) quadratic indices and modeling of Paromomycin's affinity constant with HIV-1 Ψ -RNA

A model such as equation (11) may prove to be very useful in predicting the probability of the occurrence of an interaction between a drug and a specific site on the RNA chain.

Table 3. Training Set Classification results.

Nucleotide	$\Delta P\%$ ^a	P%-cv ^b	Nucleotide	$\Delta P\%$ ^a	P%-cv ^b	Nucleotide	$\Delta P\%$ ^a	P%-cv ^b
Training Set (Nucleotide non-‘footprinted’)								
RNA-A235	98.44	99.22	RNA-A301	98.40	99.15	RNA-A332	99.61	99.80
RNA-G241	90.65	94.94	RNA-A302	99.41	99.70	RNA-G333	86.70	92.78
RNA-C243	-97.92*	99.49*	RNA-U303	86.59	92.63	RNA-A334	99.62	98.81
RNA-U244	-92.03*	97.05*	RNA-U304	89.23	94.06	RNA-G335	87.77	93.36
RNA-G251	-96.81*	99.17*	RNA-A306	96.57	99.14	RNA-G338	58.59	78.02
RNA-G257	93.56	96.51	RNA-G317	84.47	91.60	RNA-G339	-93.85*	98.55*
RNA-G259	95.11	97.35	RNA-G320	62.44	80.17	RNA-G340	58.67	78.19
RNA-G261	96.06	97.87	RNA-A326	92.93	96.13	RNA-G344	73.39	85.85
RNA-C267	-99.24*	99.86*	RNA-A327	99.35	99.67	RNA-A356	99.60	99.80
RNA-A268	-46.31*	79.05*	RNA-G328	91.63	95.46	RNA-A359	99.46	99.73
RNA-A269	96.94	98.35	RNA-G329	89.54	99.33			
RNA-A276	-96.63*	99.49*	RNA-A330	99.57	97.77			
Training Set (Nucleotides ‘footprinted’)								
RNA-G214	-98.79	99.37	RNA-G265	-44.42	71.41	RNA-G321	-92.24	95.90
RNA-C218	-97.21	98.53	RNA-G266	-92.87	96.14	RNA-C322	-98.44	99.18
RNA-C219	-98.90	99.42	RNA-A271	-84.60	90.81	RNA-U323	-96.61	98.22
RNA-A220	-84.39	90.90	RNA-G272	-98.83	95.00	RNA-A324	-93.41	96.24
RNA-G221	-99.85	99.93	RNA-C274	-96.61	98.20	RNA-G325	-99.62	99.81
RNA-A222	-84.19	90.35	RNA-G275	-98.20	99.04	RNA-G342	-93.34	96.53
RNA-A225	-42.56	56.29	RNA-G277	-98.51	99.21	RNA-C343	-98.23	99.06
RNA-C227	22.41*	66.90	RNA-G282	-92.64	96.01	RNA-C349	-98.05	98.97
RNA-C229	-98.28	99.09	RNA-G283	-96.27	97.85	RNA-C352	-97.26	98.50
RNA-U230	-94.75	97.26	RNA-C284	-98.33	99.10	RNA-G361	-93.71	96.70
RNA-C231	-97.00	98.41	RNA-G285	-95.59	97.58	RNA-C362	-99.20	99.58
RNA-U232	-38.37	68.06	RNA-C287	-99.42	99.70	RNA-A368	-95.09	97.19
RNA-C233	-95.44	97.56	RNA-U288	-88.23	93.75	RNA-A370	-81.08	88.79
RNA-C236	-97.60	98.73	RNA-A293	-79.98	88.22	RNA-U372	-5.37	51.07*
RNA-G237	-94.75	97.14	RNA-G294	-99.35	99.66	RNA-U377	-93.70	96.61
RNA-G246	-90.80	95.03	RNA-U295	-96.61	98.21	RNA-C378	-98.51	99.21
RNA-C248	-97.08	98.45	RNA-C297	-98.11	98.99	RNA-U381	-92.45	96.07
RNA-U249	-94.54	97.11	RNA-G298	-85.42	89.43	RNA-G382	-98.74	99.34
RNA-C252	-97.80	98.83	RNA-C299	-96.23	97.91	RNA-G383	-97.99	98.93
RNA-U253	-53.65	76.07	RNA-C307	-98.35	99.12	RNA-C387	-97.18	98.49
RNA-C258	67.07*	88.75*	RNA-U308	-98.12	99.00	RNA-C388	-84.47	91.70
RNA-C262	59.31*	85.25	RNA-A309	-85.79	91.59			
RNA-C264	-98.03	98.94	RNA-G310	-99.10	99.53			

*Nucleotides that are misclassified by LDA-QSAR model (Eq. 11). ^aNucleotide-Paromomycin interaction predicted by model (11); $\Delta P\% = [P(\text{interaction}) - P(\text{non-interaction})] \times 100$; where P is probability with which the nucleotide is predicted as non-footprinted or footprinted in each group. ^bPercentage of probability with which the nucleotide is predicted as footprinted or non-footprinted in each groups using LOO cross validation procedures.

Figure 2. Behavior of the global or total percentage of good classification in different n -fold cross-validation analysis.

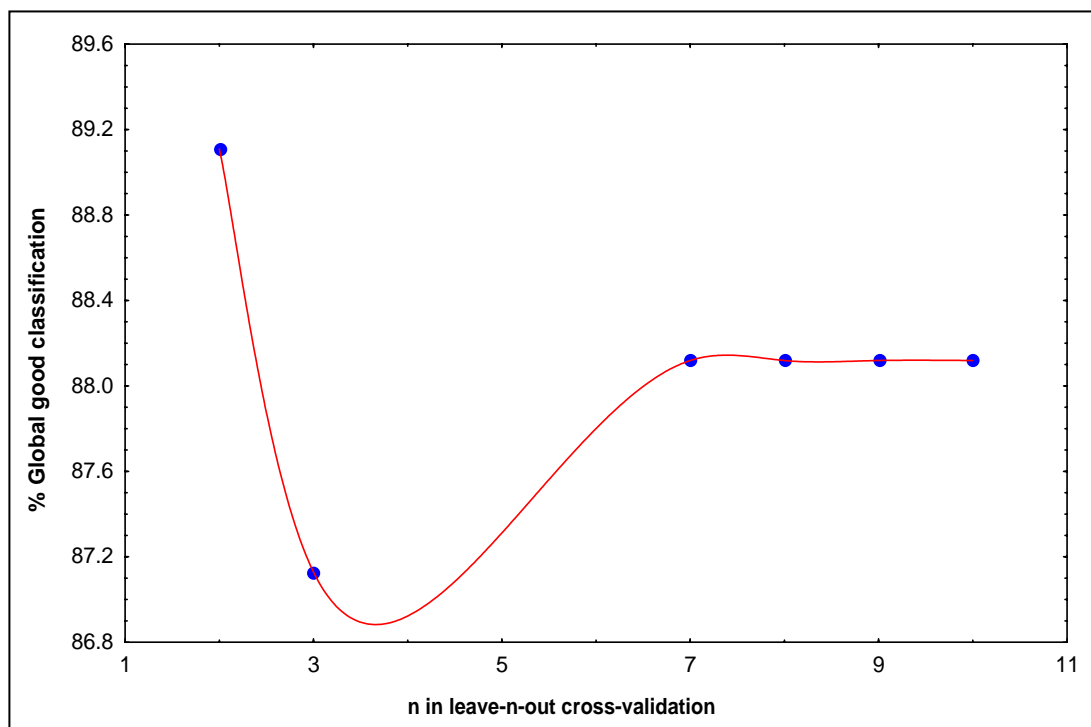


Table 4. Test set classification results.

nucleotide	$\Delta P\%^a$	nucleotide	$\Delta P\%^a$	nucleotide	$\Delta P\%^a$
Test Set (Nucleotides non-'footprinted')					
RNA-A239	98.33	RNA-A286	-80.84*	RNA-A336	99.68
RNA-A242	97.15	RNA-C300	-95.83*	RNA-G346	90.17
RNA-C245	98.23	RNA-G318	90.46	RNA-A360	94.68
RNA-G254	62.44	RNA-G331	87.67		
Test Set (Nucleotides 'footprinted')					
RNA-G213	-85.46	RNA-U250	-97.07	RNA-G348	-97.29
RNA-G226	-21.29	RNA-G273	35.31*	RNA-G369	-99.76
RNA-U228	-87.28	RNA-C311	-97.87	RNA-U373	-92.40
RNA-C238	-98.32	RNA-U341	47.94*		

*Nucleotides that are misclassified by LDA-QSAR model (Eq. 11). ^a Nucleotide-Paromomycin interaction predicted by model (11); $\Delta P\% = [P(\text{interaction}) - P(\text{non-interaction})] \times 100$; where P is probability with which the nucleotide is predicted as non-footprinted or footprinted in each group.

This is very important information for the study of the mechanism of action of potential drugs with RNA as the target.

However, any picture of the drug–RNA interaction is not complete unless the strength of each interaction is also known. With the aim of addressing this issue, a quantitative linear model was developed in order to predict the interaction constants, when they occur. The local affinity constant values [$\log K(10^{-4}\text{M}^{-1})$] were obtained from the same source as the former binding/footprinting data [9].

$$\begin{aligned} \text{Log } K(10^{-4}\text{M}^{-1}) = & -1.3747(\pm 0.3882) + 0.1136(\pm 0.0189)^{\text{AE1}} q_{0\text{L}}(x_{\text{m}}) \\ & -7.5608 \times 10^{-5}(\pm 9.9659 \times 10^{-6})^{\text{e}250} q_{3\text{L}}(x_{\text{m}}) + 0.0393(\pm 0.0069)^{\text{f}2} q_{3\text{L}}(x_{\text{m}}) \\ & -4.6544(\pm 1.63 \times 10^{-9})^{\text{AE1}} q_{10\text{L}}(x_{\text{m}}) \end{aligned} \quad (12)$$

$$N = 23 \quad R = 0.96 \quad R^2 = 0.92 \quad s = 0.07 \quad q^2 = 0.85 \quad s_{\text{cv}} = 0.09 \quad F(4.18) = 54.910 \quad p < 0.0000$$

where N is the number of interactions with a known affinity constant ($\log K$), F is Fisher's statistics, s is the standard error of estimates, R^2 is the squared regression coefficient for training and q^2 the same for the LOO jackknife experiments.

In the development of the quantitative model for the $\log K$ description of the calibration data set, one nucleotide (A276) stands out as a statistical outlier. Outlier detection was performed using the following standard statistical test: residual, standardized residuals, Studentized residual and Cooks distance.

Two of present authors reported a similar equation using MARCH-INSIDE descriptors [32]. They additionally make use of a dummy variable RNase, which has the values RNase = 1 for experiments carried out in the presence of RNase I and RNase = -1 for RNase T1 [32]:

$$\begin{aligned} \text{Log } K(10^{-4}\text{M}^{-1}) = & 0.693(\pm 0.038) + 0.338(\pm 0.068)\text{RNase} - 0.102(\pm 0.025)^1 \mathbf{O}(\Theta_{10}) \\ & + 0.083(\pm 0.035)^4 \mathbf{O}(\Theta_8) \end{aligned} \quad (13)$$

$$N = 24 \quad R = 0.91 \quad R^2 = 0.83 \quad s = 0.115 \quad q^2 = 0.825 \quad F(3.20) = 31.48 \quad p < 0.0000$$

Both equations have very similar statistical parameters. Statistical parameters in Eq. 12 suggest a high quality of the found model. The correlation coefficient R is 0.96 and standard deviation is only $0.07 \times 10^{-4}\text{M}^{-1}$. The squared correlation coefficient (R^2) was 0.92 for Eq. 12, so, this model explained more than 92% of the variance for the experimental Paromomycin affinity constant by HIV-1 RNA.

Predictability and stability of the model (12) to data variation is tested here by means of LOO cross validation. The model shows a cross validation standard error of only 0.09. In Table 5, we depict the observed, predicted and predicted (after LOO cross-validation procedures) values of $\log K$ obtained from Eq. 12 and Eq. 13. One of the main problems concerning the application of TIs to QSPR/QSAR studies is that many descriptors are collinear. Therefore, there will be much redundancy of information. Problems with redundancy of information, and collinearity, have been illustrated with the use of TIs, such as the molecular connectivities [33,34].

For a better statistical interpretation of the QSPR/QSAR models (in order to understand which effects cannot be separated), where inter-related indices are considered (such as topologic or topographic indices based on the same graph-theoretical invariant), the inclusion in the model of strongly interrelated variables should be avoided. It is necessary to consider the above-mentioned criterion because an interrelation among different descriptors produces a highly unstable correlation coefficient and makes it difficult to know the real contribution of each variable included in the model [35]. To solve this problem Randić proposed a procedure of orthogonalization of molecular descriptors that have been applied with much success to QSPR and QSAR studies [36,37].

Table 5. Observed, predicted and predicted (alter LOO cross-validation procedures) values of Log K obtained from Eq. 11 and Eq. 12.

NUC	Obs ^a	Pred ^b	P-cv ^c	Pred ^d	P-cv ^f	NUC	Obs ^a	Pred ^b	P-cv ^c	Pred ^d	P-cv ^f
A235	1.204	1.132	1.111	1.166	0.359	G335	0.845	0.852	0.853	0.862	0.845
A239	1.204	1.173	1.164	1.166	0.359	G338	0.778	0.736	0.732	0.672	0.778
G251	0.447	0.350	0.304	0.518	0.032	G339	0.778	0.647	0.566	0.545	0.778
G254	0.447	0.552	0.578	0.518	0.032	G340	0.778	0.734	0.730	0.672	0.778
C267	0.903	0.893	0.879	0.856	0.058	G344	0.845	0.814	0.811	0.735	0.845
A268	0.903	1.003	1.049	0.856	0.125	G346	0.845	0.855	0.856	0.862	0.845
A269	0.903	0.984	1.026	0.987	0.125	G363	0.415	0.488	0.522	0.399	0.415
A286	0.778	0.704	0.667	1.024	-0.067	G364	0.415	0.477	0.495	0.399	0.415
G328	0.845	0.851	0.852	0.862	0.430	G365	0.415	0.542	0.564	0.399	0.415
G329	0.845	0.852	0.853	0.862	0.430	G366	0.415	0.394	0.386	0.594	0.415
G331	0.845	0.852	0.853	0.862	0.430	G367	0.415	0.378	0.369	0.594	0.415
G333	0.845	0.852	0.853	0.862	0.845						

NUC: Nucleotide. The values are ^aObserved, ^{b y d}Predicted, and ^{c y f}Predicted by LOO procedures for log K (10^{-4}M^{-1}) (affinity constant of Paromomycin for RNA), by Eq. 12 and Eq. 13, respectively.

For the present paper, to alleviate the collinearity between variables in investigated data set, an interrelation study among the nucleic acid quadratic indices was performed, using correlation matrices. The acceptable level of collinearity to avoid is a more subjective issue. In this sense, reports of acceptable correlation coefficients between variables have range from less than 0.4 to 0.9 in the literature. In the view of the Cronin and Schultz [34], the collinearity of the variables should be as low as possible, but must be significantly lower than the statistical fit of the QSPR/QSAR itself. In Table 6, the correlation matrix for this equation shows that there is low collinearity among these variables.

Table 6. The squared correlation matrix showing covariance (r^2) among the macromolecular topological descriptors [local (nucleotide) nucleic acid quadratic indices] used in the regression analysis.

	$r^2_{q_{3L}(X_m)}$	$\Delta E^1_{q_{0L}(X_m)}$	$\Delta E^1_{q_{10L}(X_m)}$	$\epsilon^{250}_{q_{3L}(X_m)}$
$r^2_{q_{3L}(X_m)}$	1	-0.55	-0.68	-0.41
$\Delta E^1_{q_{0L}(X_m)}$		1	0.37	0.17
$\Delta E^1_{q_{10L}(X_m)}$			1	-0.31
$\epsilon^{250}_{q_{3L}(X_m)}$				1

Both LDA- and LMR-QSAR models (Eq. 11 and Eq. 12, respectively) involves short-reaching ($k \leq 3$), middle-reaching ($4 < k < 9$) and far-reaching indices ($k = 10$ or greater). The RNA quadratic indices of order zero ($k = 0$) characterized each kind of RNA bases (nucleotide), but not consider the environmental topology of the nucleotide.

In both model this indices have a positive contribution [${}^f q_{0L}(x_m)$ and ${}^{AE1} q_{0L}(x_m)$ in models (11) and (12), respectively]. This is a logical result, because this indices have a high values for purine nucleotides, which present more probability of drug interaction than pyrimidine ones. This situation means that the probability of binding increased with the consequently increase of electron density of RNA bases, due to this possibility the hydrogen bond and/or electrostatic interaction of amino groups/protonated amine groups with sites on RNA.

Three RNA-quadratic indices of the third order ($k = 3$) of involved in the early stages of Paromomycin-nucleotide interaction. Such a behavior may be explained by taking into consideration the fact that the electronic and/or topologic changes in the nucleotide backbone, which are necessary for the drug-nucleotide interaction, the more marked structural changes in the ± 3 -vicinity of the nucleotide. Consequently, two of these indices had a negative contribution in LDA [${}^f q_{3L}(x_m)$] and LMR [${}^{\epsilon 250} q_{3L}(x_m)$] model. The contribution of the middle-to-high reaching, ± 5 and ± 10 -vicinities of the nucleotide, in both equations show that the interaction between Paromomycin and a nucleotide of RNA depends on the electro-topologic environment of this nucleotide. These results are in relation to the factor that control binding specificity for aminoglycosides' interaction. In general, the Paromomycin prefers to bind bulged or other non-Watson-Crick secondary RNA elements, in consequence this drug is too large to fit into the grooves of regular A-form RNA structure [9].

Concluding Remarks

This study presents a new set of macromolecular descriptors relevant to nucleic acid QSAR/QSPR studies. These descriptors are calculated from the macromolecular graph's nucleotide adjacency matrix. Their derivation is straightforward, and it is easy to interpret the QSARs/QSPRs which include them. The local (nucleotide) quadratic indices, LDA, and LMR have been used to predict the probability and the affinity of Paromomycin binding by the packing HIV-1 region. The resulting quantitative models are significant from a statistical point of view. A LOO cross-validation procedure (internal validation) and an external predicting series (external validation) revealed that the QSAR models had a good predictability.

The models found to describe the interaction profile include nucleotide's quadratic indices accounting for electronic and topologic features of each nucleotide in RNA molecule. These models not only are good enough to predict the interaction parameters, but also permit the interpretation of the driving forces of such interaction processes. In this sense, developed equations involve short-reaching ($k \leq 3$), middle- reaching ($4 < k \leq 9$) and far-reaching ($k = 10$ or greater) nucleotide's quadratic indices. This situation points to that the interaction between Paromomycin and a nucleotide of RNA depends on the electro-topologic environment of the nucleotides.

The approach described here represents a novel and rather promising way to chem & bioinformatics research. We would expect computational nucleic acid science to have a similar effect on the search for new vaccines, receptors, drugs, and so on as molecular modeling and QSAR have had on search for new drugs.

Acknowledgements

Y. Marrero-Ponce would like to express his gratitude to Drs. David Whitley (England), David Livingstone (England), James Devillers (France), Johann Gasteiger (Germany), Klaus L. E. Kaiser (Canada), Lauren Dury (Belgium), Laurence Leherte (Belgium), Ernesto Estrada (Spain), David B. Silverman (USA) and Douglas Klein (USA) for sending him several reprints of their papers on molecular design. Y. M-P is also indebted to the journal's Managing Editor, Dr. Derek J. McPhee and Editor-in-Chief, Dr. Shu-Kun Lin, for their kind attention. F. T. acknowledges financial support from the Spanish MCT (Plan Nacional I+D+I, Project No. BQU2001-2935-C02-01).

References

1. Hua, S.; Sun, Z. Support Vector Machine Approach for Protein Subcellular Localization Prediction. *Bioinformatics*. **2001**, *17*, 721-728.
2. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Rapp, B. A.; Wheeler, D. L. Gen bank. *Nucleic Acid Res.* **2000**, *28*, 15-18.
3. Yuan, Z. Prediction of Proteins Subcellular Location Using Markov Chain Models. *FEBS Lett.* **1999**, *451*, 23-26.
4. Tullius, T. D. Physical Studies of Protein-DNA Complexes by Footprinting. *Ann. Rev. Biophys. Bio.* **1989**, *18*, 213-237.
5. Brenowitz, M.; Senear, D. F.; Shea, M. A.; Ackers G. K. Quantitative Dnase Footprint Titration: a Method for Studying Protein-DNA Interactions. *Methods Enzymol.* **1986**, *130*, 132-181.
6. Henn, A.; Halfon, J.; Kela, I.; Orion, I.; Sagi, I. Nucleic Acid Fragmentation on the Millisecond Timescale Using a Conventional x-Ray Rotating Anode Source: Application to Protein-DNA Footprinting. *Nucleic Acid Res.* **2001**, *29*, e122.
7. Galas, D. J.; Schmitz, A. Dnase Footprinting: a Simple Method for the Detection of Protein-DNA Binding Specificity. *Nucleic Acids Res.* **1978**, *5*, 3157-3170.
8. Ozoline, O. N.; Fujita, N.; Ishihama, A. Mode of DNA-protein Interaction between the C-terminal Domain of Escherichia Coli RNA Polymerase α Subunit and T7D Promoter UP Element. *Nucleic Acids Res.* **2001**, *29*, 4909-4919.
9. McPike, P. M.; Goodisman, J.; Dabrowiak, C. J. Footprinting and Circular Dichroism Studies on Paromomycin Binding to the Packaging Region of the Human Immunodeficiency Virus Type-1. *Bioorg. Med. Chem.* **2002**, *10*, 3663-3672.
10. Sullivan, J. M.; Goodisman, J.; Dabrowiak, C. J. Absorption Studies on Aminoglycosides Binding to the Packaging Region of the Human Immunodeficiency Virus Type-1. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 615-618.
11. Gale, E. F.; Gundliff, E.; Reynolds, P. E.; Richmon, M. H.; Waring, M. J. *The Molecular Basis of Antibiotic Action*; John Wiley & Sons: London. **1981**.
12. Lynch, S. R.; Recht, M. I.; Puglisi, J. D. Biochemical and Nuclear Magnetic Resonance Studies of Aminoglycoside-RNA Complexes. *Meth. Enzymol.* **2000**, *317*, 240-261.
13. Weiss, R.; Teich, N.; Varmus, H.; Coffin, J. (Eds); *RNA Tumor Viruses*; Cold Spring Harbor Laboratory: Cold Spring Harbor (N.Y.), **1984**.
14. Wilson, W. D.; Li, K. Targeting RNA with Small Molecules. *Curr. Med. Chem.* **2000**, *7*, 73-98.

15. Marrero-Ponce, Y. Total and Local Quadratic Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”: Applications to the Prediction of Physical Properties of Organic Compounds. *Molecules*. **2003**, *8*, 687-726. <http://www.mdpi.org>.
16. 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.* In Press: DOI: 10.1021/ci049950k.
17. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; and Montero, L. A. Total and Local Quadratic Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix”. Application to Prediction of Caco-2 Permeability of Drugs. *Int. J. Mol. Sci.* **2003**, *4*, 512-536. www.mdpi.org/ijms/
18. Marrero, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. A New Topological Descriptors Based Model for Predicting Intestinal Epithelial Transport of Drugs in Caco-2 Cell Culture. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186-199.
19. Marrero, Y.; González, H.; Romero, V.; Torrens, F.; Castro, E. A. 3D-Chiral Quadratic Indices of the “Molecular Pseudograph’s Atom Adjacency Matrix” and Their Application to Central Chirality Codification: Classification of ACE Inhibitors and Prediction of σ -Receptor Antagonist Activities. *Bioorg. Med. Chem.* **2004**, *12*, 5331-5342.
20. Stryer, L. *Biochemistry*. W. H. Freeman and Company: New York, **1995**.
21. Mathews, C. K.; van Holde, K. E.; Ahern, K. G. *Biochemistry*; Addison Wesley Longman: San Francisco, **2000**.
22. Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*; Worth Publishers: New York, **1993**.
23. Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*; Garland: New York and London, 1994.
24. Pogliani, L. From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors. *Chem. Rev.* **2000**, *100*, 3827-3858.
25. Randić, M. Generalized Molecular Descriptors. *J. Math. Chem.* **1991**, *7*, 155-168.
26. Marrero-Ponce, Y.; Romero-Zaldivar, V. TOMO-COMD software; Central University of Las Villas, **2002**. *TOMOCOMD* (*TO*pological *MO*lecular *COM*puter *D*esign) for Windows, version 1.0 is a preliminary experimental version; in future a professional version will be available on request from Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es
27. *STATISTICA* version. 5.5, Statsoft, Inc., **1999**.
28. McFarland, J. W.; Gans, D. J. Linear Discriminant Analysis and Cluster Significance Analysis. In *Comprehensive Medicinal Chemistry*; Hansch, C.; Sammes, P. G.; Taylor J. B. Eds.; Pergamon Press: Oxford, **1990**; vol. 4, pp. 667-689.
29. Golbraikh, A.; Tropsha, A. Beware of q^2 !. *J. Mol. Graph. Modell.* **2002**, *20*, 269-276.
30. Wold, S.; Erikson, L. Statistical Validation of QSAR Results. Validation Tools. In *Chemometric Methods in Molecular Design*, van de Waterbeemd, H. Ed.; VCH Publishers: New York, 1995; pp 309-318.
31. García-Domenech, R.; de Julián-Ortíz, J. V. Antimicrobial Activity in a Heterogeneous Group of Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445-449.

32. González, H.; Ramos, R.; Molina, R. Markovian Negentropies in Bioinformatics. 1. A picture of Footprints after the Interaction of the HIV-1 ψ -RNA Packaging Region with Drugs. *Bioinformatics*. **2003**, *16*, 2079-2087.
33. Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. Topological Indices: Their Nature and Mutual Relatedness. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891-898.
34. Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. (Theochem)*. **2003**, *622*, 39-51.
35. Alzina, R. B. *Introduccion Conceptual al Análisis Multivariable. Un Enfoque Informático con los paquetes SPSS-X, BMDP, LISREL Y SPAD*; PPU SA: Barcelona, 1989; Chapter 8, Vol. 1, p 202.
36. Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517-525.
37. Randić, M. Fitting of Nonlinear Regression by Orthogonalized Power Series. *J. Comput. Chem.* **1993**, *14*, 363-370.