Available online at www.sciencedirect.com

SCIENCE DIRECT®

Bioorganic & Medicinal Chemistry 13 (2005) 3397–3404

Bioorganic & Medicinal Chemistry

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 Ψ-RNA packaging region

Yovani Marrero Ponce,[a,*] Juan A. Castillo Garit[a,b] and Delvin Nodarse[a]

[a]*Department of Pharmacy, Faculty of Chemical-Pharmacy and Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba*
[b]*Applied Chemistry Research Center, Central University of Las Villas, Santa Clara 54830, Villa Clara, Cuba*

Received 14 January 2005; revised 1 March 2005; accepted 2 March 2005
Available online 8 April 2005

Abstract—The design of novel anti-HIV compounds has now become a crucial area for scientists around the world. In this paper a new set of macromolecular descriptors (that are calculated from the macromolecular graph's nucleotide adjacency matrix) of relevance to nucleic acid QSAR/QSPR studies, nucleic acids' linear indices. A study of the interaction of the antibiotic Paromomycin with the packaging region of the HIV-1 Ψ-RNA has been performed as example of this approach. A multiple linear regression model predicted the local binding affinity constants [Log K (10^{-4} M^{-1})] between a specific nucleotide and the aforementioned antibiotic. The linear model explains more than 87% of the variance of the experimental Log K ($R = 0.93$ and $s = 0.102 \times 10^{-4}$ M^{-1}) and leave-one-out press statistics evidenced its predictive ability ($q^2 = 0.82$ and $s_{cv} = 0.108 \times 10^{-4}$ M^{-1}). The comparison with other approaches (macromolecular quadratic indices, Markovian Negentropies and 'stochastic' spectral moments) reveals a good behavior of our method.

1. Introduction

The number of new discovered genomes has dramatically increased in recent years and this has once again highlighted the problem of protein and nucleic acid functions.[1,2] The complete sequencing of the genomes of various species will undoubtedly contribute to a better understanding of its evolution. Public databases such as GenBank are growing in size at an exponential rate.[1] A significant proportion of the data corresponds to genomic sequences containing the structures not only of many genes but also of RNA.[3,4]

The study of the interactions of drugs with biomolecules is now the hot topic in modern bioinformatics. This kind of study constitutes a significant step toward rational drug design. In this sense, the use of footprinting techniques has proven to be an important experimental method for the discovery of significant processes in molecular biology and specifically the field of genomics.[5–9] 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]

Recently, a novel scheme to the rational *in silico* molecular design (or selection/identification of chemicals) and to QSAR/QSPR studies has been introduced by our group. The so-called *TO*pological *MO*lecular *COM*puter *D*esign (TOMOCOMD).[13] This method generates molecular fingerprints based on the Discrete Mathematic and Linear Algebra Theory. In this sense, atom, atom type and total quadratic and linear molecular

Keywords: HIV-1 Ψ-RNA packaging region; Paromomycin; Footprinting; TOMOCOMD–CANAR approach; Nucleic acid linear indices.

* Corresponding author. Tel.: +53 42 281192/281473; fax: 53 42 281130/281455; e-mail addresses: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es; ymponce@gmail.com

0968-0896/$ - see front matter © 2005 Published by Elsevier Ltd.
doi:10.1016/j.bmc.2005.03.010

fingerprints have been defined in analogy to the quadratic and linear mathematical maps.[14,15] This approach has been successfully employed in QSPR and QSAR studies,[14–24] including studies related to nucleic acid–drug interactions.[25]

The TOMOCOMD–CARDD (acronym of the *C*omputed-*A*ided '*R*ational' *D*rug *D*esign) strategy is very useful for the selection of novel subsystems of compounds having a desired property/activity,[22–24] which can be further optimized by using some of the many molecular modeling methods available for medicinal chemists. The method has also demonstrated flexibility in relation to many different problems. In this sense, the TOMOCOMD–CARDD approach has been applied to the fast-track experimental discovery of novel anthelmintic compounds.[22,24] The prediction of the physical, chem-physical and chemical properties of organic compounds is a problem that can also be addressed using this approach.[14,19,21] Codification of chirality and other 3D structural features constitutes another advantage of this method.[20] This latter opportunity allows the description of the significance interpretation and the comparison to other molecular descriptors.[15,19] Additionally, promising results have been found in the modeling of the interaction between drugs and HIV packaging-region RNA in the field of bioinformatics using TOMOCOMD-CANAR (*C*omputed-*A*ided *N*ucleic *A*cid *R*esearch) approach.[25] Finally, an alternative formulation of our approach for structural characterization of proteins was carried out recently.[26] This extends methodology [TOMOCOMD-CAMPS (*C*omputed-*A*ided *M*odelling in *P*rotein *S*cience)] which was used to encompass protein stability studies—specifically how alanine scan on Arc repressor wild-type protein affects protein stability—by means of a combination of protein quadratic indices (macromolecular fingerprints) and statistical (linear and nonlinear models) methods.[26]

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 linear indices of the 'macromolecular graph's nucleotides adjacency matrix'. Besides, the present work is focused on developing quantitative structure–property relationships to predict the affinity with which paromomycin binds to the HIV-1 Ψ-RNA packaging region and compare our results with other cheminformatic methods previously reported.

2. Theoretical framework

2.1. Computational methods

A nucleic acid is a long, unbranched 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 a 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.[27–30]

In particular, a typical RNA is the single-stranded polyribonucleotide. This macromolecule has a folded 3D conformation that is held together in part by non-covalent 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 nucleotide 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), that is, 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.[27–30]

On the other hand, the general principles of the molecular linear indices of the 'molecular pseudograph's atom adjacent matrix' for small-to-medium sized organic compounds have been explained in some detail elsewhere.[14–17,20] 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.[31] For instance, the $f_{1(B)}$ property of the DNA–RNA 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.[31] Table 1 depicts nucleotides (bases) descriptors properties for the DNA–RNA bases.

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 $\Re^5, \Re^{10}, \Re^{15}, \ldots, \Re^n$, respectively, where n is the dimension of these real sets ($\Re^n$).

Table 1. Five properties of DNA–RNA bases using as labels to characterized each nucleotides

Purine and pyrimidine bases (RNA/DNA)	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

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.[31]

This approach allows us encoding RNA sequences such as AGUCACGUA throughout 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 $\Re^9$. The use of other DNA–RNA bases properties defines alternative macromolecular vectors.

2.2. Local (nucleotide) nucleic acid's linear indices of the 'macromolecular graph's nucleotide adjacency matrix'

If a nucleic acid consists of n nucleotides (vector of $\Re^n$), then the kth nucleic acid's linear indices, $f_k(x_{mi})$ are calculated as linear map on $\Re^n$ [$f_k(x_{mi})$: $\Re^n \rightarrow \Re^n$; thus $f_k(x_{mi})$: End on $\Re^n$] in canonical basis as shown in Eq. 1.

$$f_k(x_{mi}) = \sum_{j=1}^{n} {}^k a_{ij}\ {}^{m}X_j \tag{1}$$

where ${}^k a_{ij} = {}^k a_{ji}$ (symmetric square matrix), n is the number of nucleotides of the nucleic acid and ${}^m X_j$ are the coordinates of the macromolecular vector (X_m) in a system of basis vectors of $\Re^n$. The coordinates of the same vector will be different according to the basis vectors chosen.[32–35] The values of the coordinates depend thus in an essential way on the choice of the basis. With the so-called canonical ('natural') base, e_j denotes the n-tuple having 1 in the jth position and 0s elsewhere. In the canonical basis, the coordinates of any vector X coincide with the components of this vector.[32–35] For that reason, those coordinates can be considered as weights of the vertices (DNA–RNA bases) of the graph of the nucleic acid's backbone.

The coefficients ${}^k a_{ij}$ are the elements of the kth power of the macromolecular matrix $M(G_m)$ of the nucleic acid's graph (G_m). Here, $M(G_m) = [a_{ij}]$ denotes the matrix of $f_k(x_{mi})$ with respect to the natural basis. In this matrix n is the number of bases (nucleotides) in sugar–phosphate's backbone. The elements a_{ij} are defined as follows:

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

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.

Eq. 1 for $f_k(x_{mi})$ can be written as the single matrix equation:

$$f_k(x_{mi}) = [{}^m X']^k = M^k(G_m)[{}^m X] \tag{3}$$

where $[{}^m X]$ is a column vector (a $n \times 1$ matrix) of the coordinates of X_m in the canonical base of $\Re^n$ and M^k the kth power of the matrix $M(G_m)$ of the macromolecular pseudograph G_m (map's matrix). Table 2 exemplifies the calculation of $f_k(x_m)$ for a secondary structure RNA fragment.

2.3. Total (whole molecule) linear indices of the 'macromolecular graph's nucleotide adjacency matrix'

Total nucleic acid's linear indices are linear functional on $\Re^n$.[15–17,24,25] That is, the kth total nucleic acid's linear indices are linear maps from $\Re^n$ to the scalar $\Re[f_k(x_m) : \Re^n \rightarrow \Re]$. The mathematical definition of these molecular descriptors is the following:

$$f_k(x_m) = \sum_{i=1}^{n} f_k(x_{mi}) \tag{4}$$

where n is the number of nucleotides and $f_k(x_{mi})$ are the nucleic acid's linear indices (linear maps) obtained by Eq. 1. Then, a linear form $f_k(x_m)$ can be written in matrix form,

$$f_k(x_m) = [u]^t[{}^m X']^k \tag{5}$$

or

$$f_k(x_m) = [u]^t M^k [{}^m X] \tag{6}$$

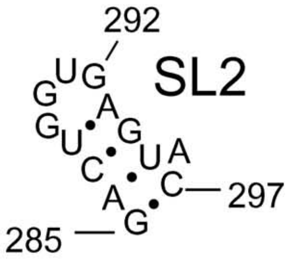
for all macromolecular vector $X_m \in \Re^n$. $[u]^t$ is an n-dimensional unitary row vector. As can be seen, the kth total linear indices are calculated by summing the local (nucleotide) linear indices of all nucleotides in the nucleic acid.

2.4. Local (nucleotide type) nucleic acid's linear indices of the 'macromolecular graph's nucleotide adjacency matrix'

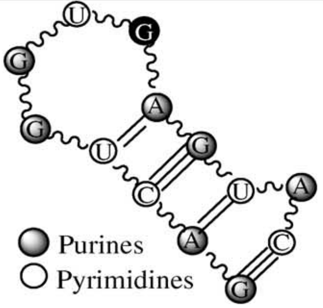
In addition to nucleic acid's linear indices computed for each nucleotide in the nucleic acid, local-fragment (nucleotide-type) formalism can be developed. The kth nucleotide-type linear indices of the 'macromolecular graph's nucleotide adjacency matrix' are calculated by summing the kth nucleic acid's linear indices of all nucleotides of the same nucleotide type in the nucleic acid.

Consequently, if a nucleic acid is partitioned in Z molecular fragment, the total nucleic acid's linear indices can be partitioned in Z local nucleic acid's linear indices $f_{kL}(x_m)$, $L = 1, \ldots, Z$. That is to say, the total nucleic acid's linear indices of order k can be expressed as the

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



Secondary structure of an RNA fragment of the SL 2 motif (see Figure 1)



Macromolecular graph's (*an undirected graph with multiple edges G_m*)

$\mathbf{X_m}$ **= [G A C U G G U G A G U A C];** $\mathbf{X_m} \in \Re^{13}$

In the definition of $\mathbf{X_m}$, as macromolecular vector, the symbol of the bases is used to indicate the corresponding DNA-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 $\Re^{13}$, the coordinates of any macromolecular vector $\mathbf{X_m}$ coincide with the components of that macromolecular vector.

$[^{\mathbf{m}}\mathbf{X}]$ **= [**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**]**

$[^{\mathbf{m}}\mathbf{X}]$: vector of coordinates of $\mathbf{X_m}$ in Canonical base of $\Re^{13}$ (a $nx1$ matrix)

$$f_1(x_{mi}) = \sum_{j=1}^{n} {}^{1}a_{ij}\, {}^{m}X_j = \mathbf{M}^1(\mathrm{G_m})[^{\mathbf{m}}\mathbf{X}] = \left[\begin{array}{cccccccccccccc} & \underline{G} & \underline{A} & \underline{C} & \underline{U} & \underline{G} & \underline{G} & \underline{U} & \underline{G} & \underline{A} & \underline{G} & \underline{U} & \underline{A} & \underline{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 \\ 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 \end{array}\right] \begin{bmatrix} \mathrm{G}_{285} \\ A_{286} \\ C_{287} \\ U_{288} \\ G_{289} \\ G_{290} \\ U_{291} \\ G_{292} \\ A_{293} \\ G_{294} \\ U_{295} \\ A_{296} \\ C_{297} \end{bmatrix} = \begin{bmatrix} \mathrm{A}_{286}+3\mathrm{C}_{297} \\ \mathrm{G}_{285}+\mathrm{C}_{287}+2\mathrm{U}_{295} \\ \mathrm{A}_{286}+\mathrm{U}_{288}+3\mathrm{G}_{294} \\ \mathrm{C}_{287}+\mathrm{G}_{289}+2\mathrm{A}_{293} \\ \mathrm{U}_{288}+\mathrm{G}_{290} \\ \mathrm{G}_{289}+\mathrm{U}_{291} \\ \mathrm{G}_{290}+\mathrm{G}_{292} \\ \mathrm{U}_{291}+\mathrm{A}_{293} \\ 2\mathrm{U}_{288}+\mathrm{G}_{292}+\mathrm{G}_{294} \\ 3\mathrm{C}_{287}+\mathrm{A}_{293}+\mathrm{U}_{295} \\ 2\mathrm{A}_{286}+\mathrm{G}_{294}+\mathrm{A}_{296} \\ \mathrm{U}_{295}+\mathrm{C}_{297} \\ 3\mathrm{G}_{285}+\mathrm{A}_{296} \end{bmatrix}$$

Nucleic acid's linear indices of first order is a *linear map*; $\boldsymbol{f_1}(x_i)$: $\Re^n \to \Re^n$ *such that,*

$\boldsymbol{f_1}(\mathrm{G_{285}, A_{286}, C_{287}, U_{288}, G_{289}, G_{290}, U_{291}, G_{292}, A_{293}, G_{294}, U_{295}, A_{296}, C_{297}}) = (\mathrm{A_{286}+3C_{297}, G_{285}+C_{287}+2U_{295}, A_{286}+U_{288}+3G_{294}, C_{287}+G_{289}+2A_{293}, U_{288}+G_{290}, G_{289}+U_{291}, G_{290}+G_{292}, U_{291}+A_{293}, 2U_{288}+G_{292}+G_{294}, 3C_{287}+A_{293}+U_{295}, 2A_{286}+G_{294}+A_{296}, U_{295}+C_{297}, 3G_{285}+A_{296}}) = (0.67, 0.69, 1.06, 0.89, 0.38, 0.38, 0.40, 0.46, 0.76, 0.85, 1.04, 0.31, 0.88)$ and whole-macromolecule linear indices of first order is a *linear functional*;

$$f_1(x) = \sum_{i=1}^{n} f_1(x_i) = \boldsymbol{f_1}(\mathrm{G_{285}}) + \boldsymbol{f_1}(\mathrm{A_{286}}) + \boldsymbol{f_1}(\mathrm{C_{287}}) + \boldsymbol{f_1}(\mathrm{U_{288}}) + \boldsymbol{f_1}(\mathrm{G_{289}}) + \boldsymbol{f_1}(\mathrm{G_{290}}) + \boldsymbol{f_1}(\mathrm{U_{291}}) + \boldsymbol{f_1}(\mathrm{G_{292}}) + \boldsymbol{f_1}(\mathrm{A_{293}})$$

$$\boldsymbol{f_1}(\mathrm{G_{294}}) + \boldsymbol{f_1}(\mathrm{U_{295}}) + \boldsymbol{f_1}(\mathrm{A_{296}}) + \boldsymbol{f_1}(\mathrm{C_{297}}) = 8.77$$

Nucleotide (N)	$f_{0L}(\mathbf{x_m}, \mathbf{N})$	$f_{1L}(\mathbf{x_m}, \mathbf{N})$	$f_{2L}(\mathbf{x_m}, \mathbf{N})$	$f_{3L}(\mathbf{x_m}, \mathbf{N})$	$f_{4L}(\mathbf{x_m}, \mathbf{N})$
G285	0.20	0.67	3.33	10.77	48.55
A286	0.28	0.69	3.81	12.82	63.72
C287	0.13	1.06	4.41	23.55	95.01
U288	0.18	0.89	2.96	11.58	49.57
G289	0.20	0.38	1.55	5.04	22.81
G290	0.20	0.38	0.78	2.39	7.16
U291	0.18	0.40	0.84	2.12	8.05
G292	0.20	0.46	1.34	5.66	21.4
A293	0.28	0.76	3.09	12.06	45.11
G294	0.20	0.85	5.16	20.59	104.63
U295	0.18	1.04	2.54	14.7	51.09
A296	0.28	0.31	1.92	4.86	26.61
C297	0.13	0.88	2.32	11.91	37.17
ARN fragment	2.64	8.77	34.05	138.05	580.88

sum of the local nucleic acid's linear indices of the Z fragments of the same order:

$$f_k(x_\mathrm{m}) = \sum_{L=1}^{Z} f_{kL}(x_\mathrm{m}) \qquad (7)$$

Any local nucleic acid's linear index has a particular meaning, especially for the first values of k, where the information about the structure of the fragment is contained. Higher values of k relate to the environment information of the fragment 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.[36] The local macromolecular indices can also be used together with total ones as variables for QSAR/QSPR modeling for properties or activities that depend more on a region or a fragment than on the macromolecule as a whole.

3. Results and discussion

The data set of footprinted and binding nucleotides was extracted from the literature.[37] Figure 1 depicts the secondary structure of the HIV-1 Ψ-RNA packaging region as well as the binding sites of Paromomycin. The local affinity constant values [$\mathrm{Log}\,K(10^{-4}\ \mathrm{M}^{-1})$] were also obtained from the literature.[37] Is very important to know the strength of each interaction in the studies of the drug–RNA interaction. In order to prove the applicability of this new approach, a quantitative linear model was developed for predicting the magnitude of such interactions. The obtained model, using these local nucleic acid's linear indices as molecular descriptors, together with its statistical parameters are given below:

$$\begin{aligned}\mathrm{Log}\,K\ (10^{-4}\ \mathrm{M}^{-1}) &= -10.498(\pm 1.359) + 4.705(\pm 0.567)^{\Delta E1}f_{0L}(x_\mathrm{m})\\ &\quad - 2.6\times 10^{-5}(\pm 3.35\times 10^{-6})^{\in 260}f_{5L}(x_\mathrm{m})\\ &\quad - 0.099(\pm 0.020)^{\in 260}f_{0L}(x_\mathrm{m})\\ &\quad - 1.915(\pm 0.450)^{\Delta E2}f_{0L}(x_\mathrm{m})\end{aligned}$$

$$N = 24 \quad R = 0.93 \quad R^2 = 0.87 \quad s = 0.102 \quad q^2 = 0.82$$
$$s_\mathrm{cv} = 0.108 \quad F(4.19) = 31.61 \quad p < 0.0001 \qquad (8)$$

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 estimation, R^2 and q^2 are both the squared regression coefficient for training set and Leave-One-Out (LOO) jackknife experiments, respectively. These statistics indicate that the model is appropriate for the description of the magnitude of the interactions between the aminoglycosides and the packaging region of type-1 HIV. The correlation coefficient R is 0.93 and standard deviation is only $0.102\times 10^{-4}\ \mathrm{M}^{-1}$. The squared correlation coefficient (R^2) was 0.87 for Eq. 8, so, this model explained about the 87% of the experimental variance on Paromomycin affinity constant for HIV-1 RNA.

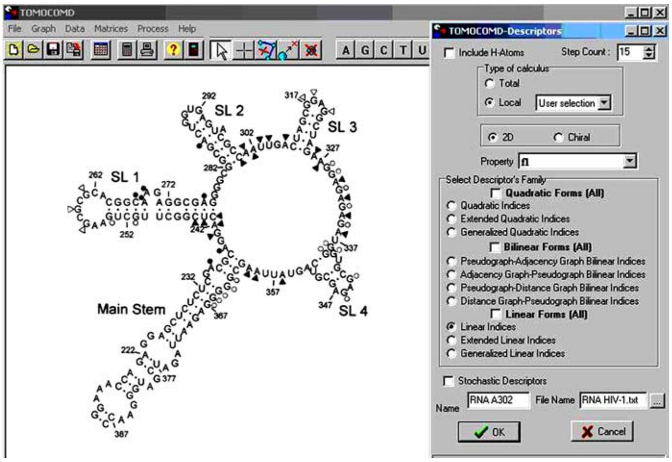
Predictability and stability of the model (8) to data variation is carried out here by means of LOO cross-validation. The model shows a cross-validation standard error of only 0.108. In Table 3, we depict the observed, predicted and predicted after the LOO cross-validation procedure values of Log K obtained from Eq. 8.

Two of the present authors reported a similar equation (see Eq. 9) using local (nucleotide) quadratic indices.[25] In the development of the quantitative model for the Log K description they detect one nucleotide (A276) as statistical outlier. This equation is given below with their statistical parameters:

$$\begin{aligned}\mathrm{Log}\,K\ (10^{-4}\ \mathrm{M}^{-1}) &= -1.3747(\pm 0.3882) + 0.1136(\pm 0.0189)^{\Delta E1}q_{0L}(x_\mathrm{m})\\ &\quad - 7.5608\times 10^{-5}(\pm 9.9659\times 10^{-6})^{\in 250}q_{3L}(x_\mathrm{m})\\ &\quad + 0.0393(\pm 0.0069)^{f2}q_{3L}(x_\mathrm{m})\\ &\quad - 4.6544(\pm 1.63\times 10^{-9})^{\Delta E1}q_{10L}(x_\mathrm{m})\end{aligned}$$

$$N = 23 \quad R = 0.96 \quad R^2 = 0.92 \quad s = 0.07 \quad q^2 = 0.85$$
$$s_\mathrm{cv} = 0.09 \quad F(4.18) = 54.910 \quad p < 0.0000 \qquad (9)$$

In addition, Gonzalez et al. reported similar equations (see Eqs. 10 and 11) using MARCH-INSIDE descriptors.[38,39] 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:[38]



□ Include H-Atoms Step Count: 15

Type of calculus
○ Total
● Local User selection

● 2D ○ Chiral

Property fl

Select Descriptor's Family

□ Quadratic Forms (All)
○ Quadratic Indices
○ Extended Quadratic Indices
○ Generalized Quadratic Indices

□ Bilinear Forms (All)
○ Pseudograph-Adjacency Graph Bilinear Indices
○ Adjacency Graph-Pseudograph Bilinear Indices
○ Pseudograph-Distance Graph Bilinear Indices
○ Distance Graph-Pseudograph Bilinear Indices

□ Linear Forms (All)
● Linear Indices
○ Extended Linear Indices
○ Generalized Linear Indices

□ Stochastic Descriptors

Name RNA A302 File Name RNA HIV-1.txt

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).

Table 3. Observed, predicted and predicted (after LOO cross-validation procedure) values of Log*K* obtained from Eq. 8

NUC	Obs[a]	Pred[b]	Pred-CV[c]
A235	1.204	1.085	1.053
A239	1.204	1.104	1.076
G251	0.447	0.317	0.274
G254	0.447	0.524	0.532
C267	0.903	0.903	0.903
A268	0.903	0.918	0.922
A269	0.903	1.099	1.153
A276	1.230	1.230	1.230
A286	0.778	0.786	0.789
G328	0.845	0.840	0.839
G329	0.845	0.843	0.843
G331	0.845	0.843	0.843
G333	0.845	0.843	0.843
G335	0.845	0.843	0.843
G338	0.778	0.767	0.766
G339	0.778	0.616	0.605
G340	0.778	0.748	0.746
G344	0.845	0.752	0.745
G346	0.845	0.820	0.818
G363	0.415	0.405	0.403
G364	0.415	0.614	0.628
G365	0.415	0.513	0.523
G366	0.415	0.569	0.580
G367	0.415	0.361	0.347

NUC: Nucleotide. The values are [a]Observed, [b]Predicted, and [c]Predicted by LOO cross-validation experiment procedure for Log*K* (10^{-4} M^{-1}) (affinity constant of Paromomycin for RNA), by Eq. 8.

$$\begin{aligned}\text{Log}K\ (10^{-4}\ \text{M}^{-1}) &= 0.693(\pm 0.038) + 0.338(\pm 0.068)\text{RNAse} \\ &\quad - 0.102(\pm 0.025)\,{}^{1}\text{O}(\Theta_{10}) \\ &\quad + 0.083(\pm 0.035)\,{}^{4}\text{O}(\Theta_{8}) \\ N = 24 \quad R = 0.91 \quad R^2 = 0.83 \quad s = 0.115 \\ q^2 = 0.825 \quad F(3.20) = 31.48 \quad p < 0.0000 \qquad (10)\end{aligned}$$

$$\begin{aligned}\text{Log}K\ (10^{-4}\ \text{M}^{-1}) &= 1.0230 + 0.52(\pm 0.04)\text{RNAse} \\ &\quad - 0.098(\pm 0.01)\,{}^{\text{SR}}\Gamma_0 + 3.606(\pm 1.444)\,{}^{\text{SR}}\Gamma_2 \\ &\quad - 3.654(\pm 1.606)\,{}^{\text{SR}}\Gamma_3 + 1.023 \\ N = 24 \quad R = 0.956 \quad R^2 = 0.914 \quad s = 0.083 \\ q^2 = 0.863 \quad F(4.19) = 50.443 \quad p < 0.0000 \qquad (11)\end{aligned}$$

All these equations have similar statistical parameters and explain between the 83% and the 92% of the variance of the experimental Log*K* values. Our model explain the 87% of the variance, it is remarkable that in the development of Eqs. 10 and 11 a dummy variable was used. On the other hand, Eq. 9 had a statistical outlier. However, these four models are useful tools to predict the probability of the occurrence of an interaction between a drug and a specific site on the RNA chain. Table 4 shows a comparison with these approaches previously described.

4. Conclusions

Although there have been many discoveries in the last years in the field of bioinformatics, it is necessary the definition of novel macromolecular descriptors that could explain different bio-macromolecular properties by means of a QSAR approach. In this sense, the approach described here represents a novel and very promising method for bioinformatics research. It presents a new set of macromolecular descriptors that are calculated from the macromolecular graph's nucleotide adjacency matrix. We have shown here that the use of the local (nucleotide) nucleic acid linear indices is able to depict the affinity with which paromomycin binds to the HIV-1 Ψ-RNA packaging region. The resulting model is significant of the statistical point of view. A LOO cross-validation experiment revealed that the QSAR model had a good predictability. The satisfactory comparative result showed that nucleic acid linear indices used here will be a novel chem and bioinformatics tool for further research.

5. Experimental section

5.1. Footprinting data

The data set of footprinted and binding nucleotides was extracted from the literature.[37] 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.

5.2. TOMOCOMD–CANAR software

TOMOCOMD is an interactive program for molecular design and bioinformatics research.[13] The program is

Table 4. Statistical parameters of the QSAR models obtained, using different molecular descriptors, to describe the magnitude of the interactions between the aminoglycosides and the packaging region of type-1 HIV

Molecular descriptors	R^2	s	q^2	s_{cv}	F
Nucleotide linear indices (Eq. 8)	0.87	0.102	0.82	0.108	31.61
Nucleotide quadratic indices (Eq. 9)	0.92	0.07	0.85	0.09	54.91
Markovian negentropies (Eq. 10)	0.83	0.115	0.825	[a]	31.48
'Stochastic' spectral moments (Eq. 11)	0.914	0.083	0.863	[a]	50.44

[a] Values are not reported in the literature.

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).

In this paper we outline salient features concerning 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 (nucleotide) macromolecular linear indices for any nucleic acids was implemented in the TOMOCOMD–CANAR software.[13] 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/DNA 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).[31] This parameterization 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 linear 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 $f_k(x_m)$, 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 $f_k(x_m)$ having, for instance, the following appearance:

$$P = a_0 f_0(x_m) + a_1 f_1(x_m) + a_2 f_2(x_m) + \cdots + a_k f_k(x_m) + c \qquad (12)$$

   where P is the measurement of the property, $f_k(x_m)$ [or $f_{kL}(x_m)$] is the kth total [or local] macromolecular linear 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 linear indices as molecular descriptors.

5.3. Statistical analysis

Based on the discussion above, a simple linear model was proposed to predict drug–nucleotide affinity. Multiple Linear Regression (MLR) was used to obtain a quantitative model. This statistical analysis was carried out with the STATISTICA software package.[40] TOMOCOMD–CANAR model used for the statistical procedure the first 10 $f_{kL}x_m$ [from $f_{0L}(x_m)$ to $f_{9L}(x_m)$] 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.

The quality of the MLR model was determined examining the statistic parameters of multivariable comparison of regression and cross-validation procedures. In this sense, the quality of the model 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}).[41]

Acknowledgements

The authors would like to offer their sincere thanks to the referees for their critical opinions about the manuscript, which have significantly contributed to improve its presentation and quality.

References and notes

1. Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Rapp, B. A.; Wheeler, D. L. *Nucleic Acids Res.* **2000**, *28*, 15.
2. Yuan, Z. *FEBS Lett.* **1999**, *451*, 23.
3. Saxonov, S.; Daizadeh, I.; Fedorov, A.; Gilbert, W. *Nucleic Acids Res.* **2000**, *28*, 185.
4. Schisler, N. J.; Palmer, J. D. *Nucleic Acids Res.* **2000**, *28*, 181.
5. Tullius, T. D. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 213.
6. Brenowitz, M.; Senear, D. F.; Shea, M. A.; Ackers, G. K. *Methods Enzymol.* **1986**, *130*, 132.
7. Henn, A.; Halfon, J.; Kela, I.; Orion, I.; Sagi, I. *Nucleic Acids Res.* **2001**, *29*, 122.
8. Galas, D. J.; Schmithz, A. *Nucleic Acid Res.* **1978**, *5*, 3157.
9. Ozoline, O. N.; Fujita, N.; Ishihama, A. *Nucleic Acids Res.* **2001**, *29*, 4909.
10. Sullivan, J. M.; Goodisman, J.; Dabrowiak, C. J. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 615.
11. Gale, E. F.; Gundliff, E.; Reynolds, P. E.; Richmon, M. H.; Waring, M. J. *The Molecular Basis of Antibiotic Action*; Wiley: London, 1981.
12. Lynch, S. R.; Recht, M. I.; Puglisi, J. D. *Methods Enzymol.* **2000**, *317*, 240.
13. Marrero-Ponce, Y.; Romero, V. TOMOCOMD software. Central University of Las Villas. 2002. TOMOCOMD

(TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in the future a professional version will be available upon request from Y. Marrero: yovanimp@qf.uclv.edu.cu; ymarrero77@yahoo.es or ymponce@gmail.com.

14. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687.
15. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
16. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; Ofori, E.; Montero, L. A. *Int. J. Mol. Sci.* **2003**, *4*, 512.
17. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186.
18. Marrero-Ponce, Y.; Cabrera, M. A.; Romero-Zaldivar, V.; Bermejo, M.; Siverio, D.; Torrens, F. *Internet Electronic J. Mol. Des.* **2005**, *4*, 124.
19. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
20. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
21. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. *Molecules* **2004**, *9*, 1100.
22. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Jorge, E.; del Valle, A.; Torrens, F.; Castro, E. A. *J. Comput. Aided Mol. Des.* **2004**, *18*, 615.
23. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, I.; Mayón Peréz, M.; García-Sánchez, R. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
24. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castañedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sánchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005.
25. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. *Int. J. Mol. Sci.* **2004**, *5*, 276.
26. Marrero-Ponce, Y.; Medina, R.; Castro, E. A.; de Armas, R.; González, H.; Romero, V.; Torrens, F. *Molecules* **2004**, *9*, 1124.
27. Stryer, L. *Biochemistry*; W.H. Freeman and Company: New York, 1995.
28. Mathews, C. K.; van Holde, K. E.; Ahern, K. G. *Biochemistry*; Addison Wesley Longman: San Francisco, 2000.
29. Lehninger, A. L.; Nelson, D. L.; Cox, M. M. *Principles of Biochemistry*; Worth: New York, 1993.
30. Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*; Garland: New York and London, 1994.
31. Pogliani, L. *Chem. Rev.* **2000**, *100*, 3827.
32. Browder, A. *Mathematical Analysis. An Introduction*; Springer: New York, 1996, pp 176–296.
33. Axler, S. *Linear Algebra Done Right*; Springer: New York, 1996, pp 37–70.
34. Ross, K. A.; Wright, C. R. B. *Matemàticas discretas*; Prentice Hall Hispanoamericana: Mẹxico DF, 1990.
35. Maltsev, A. I. *Fundamentos del Álgebra Lineal*; Mir: Moscuw, 1976, pp 68–262.
36. Randić, M. *J. Math. Chem.* **1991**, *7*, 155.
37. McPike, P. M.; Goodisman, J.; Dabrowiak, C. J. *Bioorg. Med. Chem.* **2002**, *10*, 3663.
38. González, H.; Ramos, R.; Molina, R. *Bioinformatics* **2003**, *16*, 2079.
39. González, H.; Ramos, R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
40. STATISTICA (1999) *version*. 5.5, Statsoft, Inc.
41. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269.