12th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-12) 1-30 November 2008
http://www.usc.es/congresos/ecsoc/12/ECSOC12.htm & http://www.mdpi.org/ecsoc-12

[G0010]

Nucleotide's Bilinear Indices: Novel Bio-Macromolecular Descriptors for Bioinformatics Studies of Nucleic Acids. I. Prediction of Paromomycin's Affinity Constant with HIV-1 Ψ-RNA Packaging Region

Yovani Marrero-Ponce,[1,2,3*] Sadiel E. Ortega-Broche[1,4] Yunaimy Echevería Díaz,[1] Ysaias J. Alvardo,[5] Nestor Cubillan,[5] Ricardo Grau,[6] Francisco Torrens,[2] and Facundo Pérez-Giménez.[3]

[1]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Faculty of Chemistry-Pharmacy. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*

[2]*Institut Universitari de Ciència Molecular, Universitat de València, Edifici d'Instituts de Paterna, Poligon la Coma s/n, E-46071 Valencia, Spain.*

[3]*Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Departamento de Química Física, Facultad de Farmacia, Universitat de València, Spain.*

[4]*Department of Physiology, Medical School "Faustino Pérez Hernández", Km # 3 Circumvallation, Sancti-Spíritus, Cuba.*

[5]*Laboratorio de Electrónica Molecular, Departamento de Química, Modulo II, grano de Oro, Facultad Experimental de Ciencias, La Universidad del Zulia (LUZ), Venezuela.*

[6]*Bioinformatics Group, Informatics Research Center (CEI), Faculty of Mathematics, Physics and Computer Science. Central University of Las Villas, Santa Clara, 54830, Villa Clara, Cuba.*

[*]***Corresponding author:***

Fax: 963543156

Phone: 963543156

Cell: 610028990

e-**mail**: ymarrero77@yahoo.es; ymponce@gmail.com or yovanimp@uclv.edu.cu

URL: http://www.uv.es/yoma/
http://ymponce.googlepages.com/home

Abstract

A new set of nucleotide-based biomacromolecular descriptors are presented. This novel approach to biomacromolecular design from a linear algebra point of view is relevant to nucleic acids QSAR (Quantitative Structure-Activity Relationship) studies. These bio-macromolecular indices are based on the calculus of bilinear maps on $\Re^n$ [$b_{mk}(\bar{x}_m, \bar{y}_m)$: $\Re^n x \Re^n \rightarrow \Re$] in canonical basis. Nucleic acid's bilinear indices are calculated from k^{th} power of non-stochastic and stochastic nucleotide's graph–theoretic electronic-contact matrices, M_m^k and ${}^sM_m^k$, respectively. That is to say, the k^{th} non-stochastic and stochastic nucleic acid's bilinear indices are calculated using M_m^k and ${}^sM_m^k$ as matrix operators of bilinear transformations. Moreover, biochemical information is codified by using different pair combinations of nucleotide-base properties as weightings (experimental molar absorption coefficient $\in_{260}$ 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. As example of this approach, an interaction study of the antibiotic Paromomycin with the packaging region of the HIV-1 Ψ-RNA have been performed and it have been obtained several linear models in order to predict the interaction strength. The best linear model obtained by using non-stochastic bilinear indices explains about 91% of the variance of the experimental Log *K* (R = 0.95 and s = 0.08x10^{-4}M^{-1}) as long as the best stochastic bilinear indices-based equation account for 89% of the Log *K* variance (R = 0.94 and s = 0.10 x10^{-4}M^{-1}). The Leave-One-Out (LOO) press statistics, evidenced high predictive ability of both models (q^2 = 0.86 and s_{cv} = 0.09×10^{-4}M^{-1} for non-stochastic and q^2 = 0.79 and s_{cv} = 0.11 x10^{-4}M^{-1} for stochastic bilinear indices). The nucleic acid's bilinear indices based models compared favourably with other nucleic acid's indices based approaches reported nowadays. 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 bilinear 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 theoretical-biology studies.

Keywords: ***TOMOCOMD-CANAR*** software, Nucleic Acid and Nucleotide Bilinear Indices, HIV-1 Ψ-RNA Packaging Region, Paromomycin, Footprinting, QSPR, Linear Multiple Regression.

Running head: *Nucleotide's Bilinear Indices: Novel Bio-Macromolecular Descriptors for Bioinformatics Studies…*

INTRODUCTION

The knowledge about the functions of an huge amounts of nucleotide and amino-acid sequences, generated from the sequencing projects in recent years, highlights among the challenges to modern biology (Benson et al., 2000; Sakharkar et al., 2000a; Sakharkar et al., 2000b; Saxonov et al., 2000; Schisler and Palmer, 2000 ; Yuan, 1999). This data expects for capable methods to translate the information into biological significance (Hua and Sun, 2001).

At the present time, the study of the interactions of drugs with biomolecules is a field of lively research (González-Díaz et al., 2003b). Specifically, design of molecules that bind RNA fragment is currently an interesting and important issue in drug discovery (Hamasaki and Akihiko, 2001). In this respect, the combination of experimental techniques with the modern *Bioinformatics* has arise as a promising alternative (González-Díaz et al., 2003b). In this sense, the foot-printing techniques have proven to be an important experimental method for the discovery of significant processes in molecular biology and specifically the field of genomics (Brenowitz et al., 1986; Galas and Schmithz, 1978; Henn et al., 2001; Ozoline et al., 2001; Tullius, 1989).

The interactions of antibiotics (aminoglycosides) with the packaging region of HIV Type-1 seems to be a promising route for antiviral discovery (Sullivan et al., 2002). Aminoglycoside drugs are cationic natural products that interact with RNA (Gale et al., 1981). Some structurally related aminoglycoside antibiotics bind RNA specifically and disturb their activity (Hamasaki and Akihiko, 2001). For example, the bactericidal effects inherent in these compounds stem from their ability to block protein synthesis by binding to the A-site on ribosomal RNA (Lynch et al., 2000). Moreover, aminoglycoside analogues

can be used to treat certain diseases. For instance, the genetic information in HIV and various tumour viruses is in the form of RNA (Weiss et al., 1984). 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 (Wilson and Li, 2000).

Increasingly, modern bioinformatics approaches have been used to provide structural information about bio-molecules and its interaction with drugs (Österberg et al., 2002). Several computational drugs design methods have been developed to research drug-biomolecules interactions. For instance, ***MARCH-INSIDE*** methodology has been generalized to protein structure/property relationships studies (Gonzalez-Diaz and Uriarte, 2005; Gonzalez-Diaz et al., 2005; Ramos de Armas et al., 2004) and the research in nucleic acid-drug interactions, respectively (González-Díaz et al., 2003a; González-Díaz et al., 2003b).

On the other hand, a novel scheme to the rational *in silico* molecular design (or selection/identification of drugs-like compounds) and to QSAR/QSPR (Quantitative Activity/Structure–Property Relationships) studies has been introduced by our group, the so-called ***TO***pological ***MO***lecular ***COM***puter ***D***esign (***TOMOCOMD***) (Marrero-Ponce and Romero, 2002). This method generates molecular descriptors (MDs) based on the Discrete Mathematic and Linear Algebra Theory. In this sense, atom, atom-type and total quadratic and linear molecular indices have been defined in analogy to the quadratic and linear mathematical maps (Marrero-Ponce, 2003; Marrero Ponce, 2004). This approach has been successfully employed in QSPR and QSAR studies (Marrero-Ponce, 2003; Marrero-Ponce, 2004b; Marrero-Ponce et al., 2003; Marrero-Ponce et al., 2004a; Marrero-Ponce et al., 2004b; Marrero-Ponce et al., 2005d; Marrero-Ponce et al., 2005e; Marrero-Ponce et al.,

2005g; Marrero-Ponce et al., 2004e; Marrero Ponce, 2004; Marrero Ponce et al., 2004), including studies related to nucleic acid–drug interactions (Marrero-Ponce et al., 2004d).

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 (Marrero-Ponce et al., 2005d; Marrero-Ponce et al., 2005g; Marrero-Ponce et al., 2004e), which can be further optimized by using some of the many molecular modelling 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 antihelmintic compounds (Marrero-Ponce et al., 2005d; Marrero-Ponce et al., 2005g; Marrero-Ponce et al., 2004e). The prediction of the physical, chem-physical and chemical properties of organic compounds is a problem that can also be addressed using this approach (Marrero-Ponce, 2003; Marrero-Ponce, 2004b; Marrero-Ponce et al., 2004a). Codification of chirality and other 3D structural features constitutes another advantage of this method (Marrero-Ponce et al., 2004b). This latter opportunity allows the description of the significance interpretation and the comparison to other molecular descriptors (Marrero-Ponce, 2004b; Marrero Ponce, 2004). Additionally, promising results have been found in the modeling of the interaction between drugs and HIV packaging-region RNA in the field of bioinformatics by using TOMOCOMD-***CANAR*** (***C***omputed-***A***ided **N**ucleic **A**cid **R**esearch) approach (Marrero-Ponce et al., 2004d; Marrero Ponce et al., 2005). Finally, an alternative formulation of our approach for structural characterization of proteins was carried out (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004c). These extends methodologies [TOMOCOMD-***CAMPS*** (***C***omputed-***A***ided ***M***odelling in ***P***rotein ***S***cience)] which were 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 quadratic and protein linear indices, correspondingly, (bio-macromolecular descriptors) and statistical (linear and nonlinear models) methods (Marrero-Ponce et al., 2005b; Marrero-Ponce et al., 2004c).

More recently, some of present authors also proposed new MDs in analogy to the bilinear mathematical forms in $\Re^n$ in canonical basis sets (Marrero-Ponce et al., 2008b), *namely atom-based non-stochastic and stochastic bilinear indices* (Castillo-Garit et al., 2007; Marrero-Ponce et al., 2008a; Marrero-Ponce et al., 2008b; Marrero-Ponce et al., 2007; Marrero-Ponce et al., 2006b). The calculation of these novel sets of atom-level MDs can also be carried out employing our *in house TOMOCOMD-CARDD* program (Marrero-Ponce and Romero, 2002). The computation of the non-stochastic and stochastic bilinear indices is develop by using the k^{th} "nonstochastic and stochastic atom(atomic nuclei)-based graph–theoretical electronic-density matrices" $\mathbf{M}^k$ and $\mathbf{S}^k$, correspondingly, as matrices of the mathematical forms (Castillo-Garit et al., 2007; Marrero-Ponce, 2004a; Marrero-Ponce, 2004b; Marrero-Ponce et al., 2005a; Marrero-Ponce et al., 2008a; Marrero-Ponce et al., 2008b; Marrero-Ponce et al., 2007; Marrero-Ponce et al., 2005h; Marrero-Ponce et al., 2006b; Montero-Torres et al., 2006).These matricial operators are graph-theoretical electronic-structure models, like the ''extended Hückel MO model.'' The $\mathbf{M}^1$ matrix considers all valence-bond electrons (σ- and π-networks) in one step, and their power k (k = 0, 1, 2, 3,...) can be considered as an interacting-electronic chemical-network in step k. The present approach is based on a simple model for the intramolecular (stochastic) movement of all outer-shell electrons. The theoretical scaffold of these atom-based bilinear maps and their use to represent small-to-medium size organic chemicals as well as QSAR and drug design studies has been explained in some detail elsewhere (Castillo-Garit et al., 2007;

Marrero-Ponce et al., 2008a; Marrero-Ponce et al., 2008b; Marrero-Ponce et al., 2007; Marrero-Ponce et al., 2006b). In this connection, these new MDs have also been useful for the selection of novel molecular *subsystems* having a desired property/activity. For instance, they were successfully applied to the virtual screening (computational discovery) of novel trichomonacidals(Marrero-Ponce et al., 2006b) and tyrosinase inhibitors (Marrero-Ponce et al., 2007). Thus it is desirable to also to extend the already defined atom-based (atom-level) bilinear indices to bilinear index for nucleotide, and nucleotide-type as well as for whole nucleic acid.

Therefore, describing an extended ***TOMOCOMD-CANAR*** approach to account for RNA structure, by mean of bilinear forms, constitutes the main aim of this paper. In the present study, we propose a nucleotide, nucleotide-type and total definition of non-stochastic and stochastic nucleic acid bilinear indices in analogy to the bilinear mathematical maps. Besides, the present work is focused on developing QSPRs to predict the affinity with which paromomycin binds to the HIV-1 Ψ-RNA packaging region and compare our results with other bio-chem-informatic methods previously reported.

2. MATHEMATICAL DEFINITION

In previous publications, one of the present authors (M-P,Y) of this work describes remarkable features concerned with the theory of 2D atom-based ***TOMOCOMD-CARDD*** MDs (Castillo-Garit et al., 2007; Marrero-Ponce, 2004a; Marrero-Ponce, 2004b; Marrero-Ponce et al., 2005a; Marrero-Ponce et al., 2008a; Marrero-Ponce et al., 2008b; Marrero-Ponce et al., 2007; Marrero-Ponce et al., 2005h; Marrero-Ponce et al., 2006b; Montero-Torres et al., 2006). This method codifies the molecular structure by means of mathematical quadratic, linear and bilinear transformations. In order to calculate these algebraic maps for

a molecule, the atom-based molecular vector, $\bar{x}$ (vector representation) and k^{th} "non-stochastic and stochastic graph–theoretic electronic-density matrices", $\mathbf{M}^k$ and $\mathbf{S}^k$ correspondingly (matrix representations), are constructed (Casañola-Martin et al., 2006; Marrero-Ponce, 2003; Marrero-Ponce, 2004b; Marrero-Ponce et al., 2005a; Marrero-Ponce et al., 2003; Marrero-Ponce et al., 2004a; Marrero-Ponce et al., 2005c; Marrero-Ponce et al., 2005d; Marrero-Ponce et al., 2005e; Marrero-Ponce et al., 2005f; Marrero-Ponce et al., 2006a; Marrero-Ponce et al., 2005g; Marrero-Ponce et al., 2004e; Marrero-Ponce et al., 2005h; Marrero Ponce, 2004; Marrero Ponce et al., 2004; Meneses-Marcel et al., 2005a; Meneses-Marcel et al., 2005b; Montero-Torres et al., 2005; Montero-Torres et al., 2006). In connection with, atom-based quadratic and linear indices were recently extended to structural codification and biological properties prediction of biopolymers (Marrero-Ponce et al., 2004c; Marrero Ponce et al., 2005) by using amino-acid or nucleotide-adjacency relationships and chemical-information codification as it corresponds. Here, we will extend this mathematical approach but by using bilinear maps. Therefore, the structure of this section will be as follows: 1) a background in nucleotide-based macromolecular vector and non-stochastic and stochastic nucleotides's graph–theoretic electronic-contact matrices will be described in the next subsections (2.1 and 2.2, respectively), and 2) an outline of the mathematical definition of bilinear maps and a definition of our procedures will be develop in subsections 2.3 and 2.4, correspondingly.

2.1. Chemical Information and Nucleotide-based Macromolecular Vector

In analogy to the molecular vector $\bar{x}$ used to represent organic molecules (Marrero-Ponce et al., 2004d; Marrero Ponce et al., 2004) we introduce here the nucleotide based macromolecular vector ($\bar{x}_m$). The components of this vector are numeric values, which

represent a certain nitrogenous base property. These properties characterize each kind of nucleotide (nitrogenous base) within a nucleic acid. Such properties can be experimental molar absorption coefficient $\in_{260}$ at 260 nm and PH = 7.0, first **(** ΔE_1**)** and second **(** ΔE_2 **)** single excitation energies in eV, and first **(***f1***)** and second **(***f2***)** oscillator strength values (of the first singlet excitation energies) of the nucleotide DNA-RNA bases, and so on (Pogliani, 2000). For instance, the $f_{1\,(B)}$ property of the DNA-RNA bases B takes the values f_1 = 0.28 for adenine, $f_{1\,(G)}$ = 0.20 for guanine, $f_{1\,(U)}$ = 0.18 for uracile, and so on (Pogliani, 2000). Table 1 depicts nucleotides (bases) descriptors properties for DNA-RNA bases.

Table 1 comes about here (see end of the document)

Thus, a RNA (or DNA) having 5, 10, 15,..., *n* nucleotides can be represented by means of vectors, with 5, 10, 15,..., *n* components, belonging to the spaces $\Re^5$, $\Re^{10}$, $\Re^{15}$,..., $\Re^n$, respectively. Where *n* is the dimension of the real sets ($\Re^n$).

This approach allows us encoding RNA sequences such as **5'-AGCGCCU- 3'** through out the macromolecular $\bar{x}_m$ = [0.28 0.20 0.13 0.20 0.13 0.13 0.18], in the f_1-scale (see Table 1 for more details). This vector belongs to the product space $\Re^7$. The use of other scales defines alternative macromolecular vectors.

Now, if we are interested to codify the chemical information by means of two different macromolecular vectors, for instance, $\bar{x}_m = [x_{m1},\ldots, x_{mn}]$ and $\bar{y}_m = [y_{m1},\ldots, y_{mn}]$; then different combinations of macromolecular vectors ($\bar{x}_m \neq \bar{y}_m$) are possible when a weighting scheme is used. In the present report, we characterized each nucleotide with the chem-physical parameters shown in Table 1. From this weighting scheme, ten (or twenty if $\bar{x}_{m\,w}$-$\bar{y}_{m\,z} \neq \bar{x}_{m\,z}$-$\bar{y}_{m\,w}$) combinations (pairs) of macromolecular vectors ($\bar{x}_m, \bar{y}_m; \bar{x}_m \neq \bar{y}_m$) can be

computed, $\bar{x}_{m\,f1}\text{-}\bar{y}_{m\,f2}$, $\bar{x}_{m\,f1}\text{-}\bar{y}_{m\,\in 260}$, $\bar{x}_{m\,f1}\text{-}\bar{y}_{m\,\mathrm{E1}}$, $\bar{x}_{m\,f1}\text{-}\bar{y}_{m\,\mathrm{E2}}$, $\bar{x}_{m\,f2}\text{-}\bar{y}_{m\,\in 260}$, $\bar{x}_{m\,f2}\text{-}\bar{y}_{m\,\mathrm{E1}}$, $\bar{x}_{m\,f2}\text{-}\bar{y}_{m\,\mathrm{E2}}$, $\bar{x}_{m\,\in 260}\text{-}\bar{y}_{m\,\mathrm{E1}}$, $\bar{x}_{m\,\in 260}\text{-}\bar{y}_{m\,\mathrm{E2}}$, $\bar{x}_{m\,\mathrm{E1}}\text{-}\bar{y}_{m\,\mathrm{E2}}$. Here, we used the symbols $\bar{x}_{m\,w}\text{-}\bar{y}_{m\,z}$, where the subscripts *w* and *z* mean two nitrogenous-base properties from our weighting scheme and a hyphen (-) expresses the combination (pair) of two selected nucleotide-label physic-chemical properties.

In order to illustrate this, let us consider the same RNA sequence mentioned previously and the following weighting scheme: f_1 and f_2 ($\bar{x}_{m\,f1}\text{-}\bar{y}_{m\,f2} = \bar{x}_{m\,f2}\text{-}\bar{y}_{m\,f1}$). The next macromolecular vectors $\bar{x}_m$ = [0.28 0.20 0.13 0.20 0.13 0.13 0.18] and $\bar{y}_m$ = [0.54 0.27 0.72 0.27 0.72 0.72 0.37] are obtained when we use f_1 and f_2 as chem-physical weights for codifying each nucleotide in the example RNA fragment in $\bar{x}_m$ and $\bar{y}_m$ vectors, respectively. (See Table 2 for more details).

Table 2 comes about here (see end of the document)

2.2. Background in non-stochastic and stochastic nucleotide's graph–theoretic electronic-contact matrices.

In molecular topology, molecular structure is expressed, generally, by the hydrogen-suppressed graph. That is, a molecule is represented by a graph. Informally a graph G is a collection of vertices (points) and edges (lines or bonds) connecting these vertices (I. Gutman, 1986; Rouvray, 1976; Trinajstić, 1983). In more formal terms, a simple graph G is defined as an ordered pair [V(G), E(G)] which consists of a nonempty set of vertices V(G) and a set E(G) of unordered pairs of elements of V(G), called edges (I. Gutman, 1986; Rouvray, 1976; Trinajstić, 1983).

On the other hand, the nucleic acids are polymeric biomolecules which use the nucleotides like structural basic units. The nucleotides are compound by three characteristic

components: 1) a pentose, 2) a nitrogenous base and 3) a phosphate. The nitrogenous bases are derivatives of pyrimidine and purine. The base of a nucleotide is linked covalently in an *N*-*β*-glycosil bond to the 1'carbon of the pentose, and the phosphate is esterified to the 5'carbon (Lehninger et al., 1993).

Both DNA and RNA contain two major purine bases, adenine (A) and guanine (G), and two major pyrimidines. In both DNA and RNA one of the pyrimidines is cytosine (C), but the second major pyrimidine is not the same in both: it is thymine (T) in DNA and uracil (U) in RNA (Lehninger et al., 1993). Nucleic acids have two kinds of pentoses. The recurring deoxyribonucleotide units of DNA contain 2'-deoxy-D-ribose, and the ribonucleotide units of RNA contain D-ribose (Lehninger et al., 1993). The successive nucleotides of both DNA and RNA are covalently linked through phosphate-group "bridges", in which the 5'-phosphate group of one nucleotide unit is joined to the 3'-hydroxyl group of the next nucleotide, creating a phosphodiester linkage. Thus the covalent backbones of nucleic acids consist of alternating phosphate and pentose residues, and the nitrogenous bases may be regarded as side groups joined to the backbone at regular intervals (Lehninger et al., 1993).

The purines and pyrimidines common in DNA and RNA are highly conjugated molecules, a property with important consequences for the structure, electron distribution, and light absorption of nucleic acids. The most important functional groups of pyrimidines and purines are ring nitrogens, carbonyl groups, and exocyclic amino groups. Hydrogen bonds involving the amino and carbonyl groups are the second important mode of interaction between bases in nucleic acid molecules (Lehninger et al., 1993).

Most of the weak interactions (hydrogen bonds) form between Watson–Crick complementary bases (between pairs of non-consecutive bases), that is, between A and T

(or A and U in RNA) 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 (Alberts et al., 1994; Lehninger et al., 1993; Mathews et al., 2000; Stryer, 1995). Therefore, a RNA (or DNA) molecule can be depicted by means a graph. Graph's vertices are nucleotides into polynucleotide chain and edges are both covalent interactions between nucleotides (phosphodiester bonds) and non-covalent interactions between nitrogenous bases (hydrogen bonds) from different nucleotides into polynucleotide sequence. Table 2 displays an example of how to depict a RNA sequence through a macromolecular graph.

The *n*x*n* k^{th} non-stochastic nucleotide's graph–theoretic electronic-contact matrix, M_m^k, is a square and symmetric matrix, where *n* is the number of nucleotides in the RNA (or DNA) sequence. The coefficients ${}^k m_{ij}$ are the elements of the k^{th} power of M_m and are defined as follows:

$m_{ij} = 1$ if $i \neq j$ and $\exists\, e_k \in E(G_m)$ **(1)**

$= 0$ otherwise

where $E(G_m)$ represents the set of edges of G_m.

The matrix M_m^k provides the numbers of walks of length *k* that links every pair of vertices v_i and v_j. For this reason, each edge in M_m^1 represents a phosphodiester bond (covalent bond) or hydrogen-bonds (non-covalent bond) between nucleotides *i* and *j*.

On the other hand, the k^{th} stochastic nucleotide's graph–theoretic electronic-contact matrix of G_m, ${}^s M_m^k$, can be directly obtained from M_m^k. Here, ${}^s M_m^k = [{}^k\mathbf{sm}_{ij}]$, is a square matrix of order *n* (*n* = number of nucleotides) and the elements ${}^k\mathbf{sm}_{ij}$ are defined as follows (Marrero-Ponce and F., 2005; Y. Marrero-Ponce, 2005a; Y. Marrero-Ponce, 2005b):

$$ {}^{k}sm_{ij} = \frac{{}^{k}m_{ij}}{{}^{k}SUM_{i}} = \frac{{}^{k}m_{ij}}{{}^{k}\delta_{i}} \qquad (2) $$

where, ${}^{k}m_{ij}$ are the elements of the k^{th} power of M_m^k and the SUM of the i^{th} row of M_m^k are named the *k*-order vertex degree of nucleotide *i*, ${}^{k}\delta_{i}$. It should be remarked that the matrix ${}^{s}M_m^k$ has the property that *the sum of the elements in each row is* 1. A *n*x*n* matrix with nonnegative entries having this property is called a "stochastic matrix" (Edwards and Penney, 1988). For an example of this matrices see Tables 3 and 4.

Tables 3 and 4 come about here (see end of the document)

2.3. A Theoretical Scaffold of Mathematical Bilinear Forms.

In mathematics, a bilinear form in a real vector space is a mapping $b : VxV \to \Re$, which is linear in both arguments (Burgos-Román, 1994; Burgos-Román, 2000; Hernández, 1987; Jacobson, 1985; K. F. Riley, 1998; Werner, 1981). That is, this function satisfies the following axioms for any scalar α and any choice of vectors $\overline{v}, \overline{w}, \overline{v}_1, \overline{v}_2, \overline{w}_1$ and $\overline{w}_2$.

i. $b(\alpha\overline{v}, \overline{w}) = b(\overline{v}, \alpha\overline{w}) = \alpha b(\overline{v}, \overline{w})$

ii. $b(\overline{v}_1 + \overline{v}_2, \overline{w}) = b(\overline{v}_1, \overline{w}) + b(\overline{v}_2, \overline{w})$

iii. $b(\overline{v}, \overline{w}_1 + \overline{w}_2) = b(\overline{v}, \overline{w}_1) + b(\overline{v}, \overline{w}_2)$

That is, *b* is *bilinear* if it is linear in each parameter, taken separately.

Let *V* be a real vector space in $\Re^n$ ($V \in \Re^n$) and consider that the following vector set, $\{\overline{e}_1, \overline{e}_2, ..., \overline{e}_n\}$ is a basis set of $\Re^n$. This basis set permits us to write in unambiguous form any vectors $\overline{x}$ and $\overline{y}$ of *V*, where $(x^1, x^2, ..., x^n) \in \Re^n$ and $(y^1, y^2, ..., y^n) \in \Re^n$ are the coordinates of the vectors $\overline{x}$ and $\overline{y}$, respectively. That is to say,

$$\bar{x} = \sum_{i=1}^{n} x^i \bar{e}_i \qquad (3)$$

and,

$$\bar{y} = \sum_{j=1}^{n} y^j \bar{e}_j \qquad (4)$$

Subsequently,

$$b(\bar{x}, \bar{y}) = b(x^i \bar{e}_i, y^j \bar{e}_j) = x^i y^j b(\bar{e}_i, \bar{e}_j) \qquad (5)$$

if we take the a_{ij} as the nxn scalars $b(\bar{e}_i, \bar{e}_j)$. That is,

$$a_{ij} = b(\bar{e}_i, \bar{e}_j), \text{ to } i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,n \qquad (6)$$

Then,

$$b(\bar{x}, \bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = [X]^T A[Y] = \begin{bmatrix} x^1 & \ldots & x^n \end{bmatrix} \begin{bmatrix} a_{11} & \ldots & a_{jn} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix} \begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} \qquad (7)$$

As it can be seen, the defined equation for b may be written as the single matrix equation (see Eq. **7**), where [Y] is a column vector (an nx1 matrix) of the coordinates of $\bar{y}$ in a basis set of $\Re^n$, and $[X]^T$ (a 1xn matrix) is the transpose of [X], where [X] is a column vector (an nx1 matrix) of the coordinates of $\bar{x}$ in the same basis of $\Re^n$.

Finally, we introduce the formal definition of symmetric bilinear form. Let V be a real vector space and b be a bilinear function in VxV. The bilinear function b is called symmetric if $b(\bar{x}, \bar{y}) = b(\bar{y}, \bar{x}), \forall \bar{x}, \bar{y} \in V$ (Burgos-Román, 1994; Burgos-Román, 2000; Hernández, 1987; Jacobson, 1985; K. F. Riley, 1998; Werner, 1981). Then,

$$b(\bar{x}, \bar{y}) = \sum_{i,j}^{n} a_{ij} x^i y^j = \sum_{i,j}^{n} a_{ji} x^j y^i = b(\bar{y}, \bar{x}) \quad (8)$$

2.4. Non-Stochastic and Stochastic Nucleotide-Based Bilinear Indices: Total (Global) Definition.

The k^{th} non-stochastic and stochastic bilinear indices for a nucleic acid, $b_{mk}(\bar{x}_m, \bar{y}_m)$ and ${}^s b_{mk}(\bar{x}_m, \bar{y}_m)$, are computed from these k^{th} non-stochastic and stochastic graph–theoretic electronic-contact matrix, M_m^k and ${}^s M_m^k$ as shown in Eq. **9** and **10**, respectively:

$$b_{mk}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k m_{ij} x_m{}^i y_m{}^j \quad (9)$$

$${}^s b_{mk}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^k sm_{ij} x_m{}^i y_m{}^j \quad (10)$$

where n is the number of nucleotides in the nucleic acid, and $x_m{}^1, \ldots, x_m{}^n$ and $y_m{}^1, \ldots, y_m{}^n$ are the coordinates or components of the macromolecular vectors $\bar{x}_m$ and $\bar{y}_m$ in a canonical basis set of $\Re^n$.

The defined equations (**9**) and (**10**) for $b_{mk}(\bar{x}_m, \bar{y}_m)$ and ${}^s b_{mk}(\bar{x}_m, \bar{y}_m)$ may be also written as the single matrix equations **11** and **12**, correspondingly:

$$b_{mk}(\bar{x}_m, \bar{y}_m) = [X_m]^T M_m^{\boldsymbol{k}} [Y_m] \quad (11)$$

$${}^s b_{mk}(\bar{x}_m, \bar{y}_m) = [X_m]^T {}^s M_m^{\boldsymbol{k}} [Y_m] \quad (12)$$

where $[Y_m]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{y}_m$ in the canonical basis set of $\Re^n$, and $[X_m]^T$ is the transpose of $[X_m]$, where $[X_m]$ is a column vector (an $nx1$ matrix) of the coordinates of $\bar{x}_m$ in the canonical basis of $\Re^n$. Therefore, if we use the

canonical basis set, the coordinates [($x_m^1,...,x_m^n$) and ($y_m^1,...,y_m^n$)] of any macromolecular vectors ($\bar{x}_m$ and $\bar{y}_m$) coincide with the components of those vectors [($x_{m1},...,x_{mn}$) and ($y_{m1},...,y_{mn}$)]. For that reason, those coordinates can be considered as weights (nitrogenous bases, that is to say "nucleotide labels") of the vertices of G_m, due to the fact that components of the macromolecular vectors are values of some nucleotide property that characterizes each kind of nitrogenous base in the nucleic acid.

It should be remarked that non-stochastic and stochastic bilinear indices are symmetric and non-symmetric bilinear forms, respectively. Therefore, if in the following weighting scheme, *w* and *z* are used as nucleotide weights to compute these nucleic acid's bilinear indices, two different sets of stochastic bilinear indices, ${}^{w\text{-}z\,s}b_{mk}(\bar{x}_m,\bar{y}_m)$ and ${}^{z\text{-}w\,s}b_{mk}(\bar{x}_m,\bar{y}_m)$ [because $\bar{x}_{m\,w}\text{-}\bar{y}_{m\,z} \neq \bar{x}_{m\,z}\text{-}\bar{y}_{m\,w}$] can be obtained and only one group of non-stochastic bilinear indices ${}^{w\text{-}z}b_{mk}(\bar{x}_m,\bar{y}_m) = {}^{z\text{-}w}b_{mk}(\bar{x}_m,\bar{y}_m)$ because in this case $\bar{x}_{m\,w}\text{-}\bar{y}_{m\,z} = \bar{x}_{m\,z}\text{-}\bar{y}_{m\,w}$ can be calculated. Tables 3 and 4 show how determine the non-stochastic and stochastic total bilinear indices of several orders for the RNA sequence of Table 2.

2.5. Non-Stochastic and Stochastic Local Bilinear Indices: Nucleotide, Nucleotide-type and Nucleic Acid Fragment Bilinear Indices Definition.

In the last decade, Randić (Randić, 1991) proposed a list of desirable attributes for a MDs. Therefore, this list can be considered as a methodological guide for the development of new topological indices. One of the most important criteria is the possibility of defining the descriptors locally. This attribute refers to the fact that the index could be calculated for the molecule (for us nucleic acids) as a whole but also over certain fragments of the structure itself. Sometimes, the properties of a group of biomolecules (nucleic acid or

protein) are related more to a certain zone or fragment than to the bio-macromolecule as a whole. Thereinafter, the global definition never satisfies the structural requirements needed to obtain a good correlation in QSAR and QSPR studies.

Therefore, in addition to *total bilinear indices* computed for the whole nucleic acid, a local-fragment (polynucleotidic fragment) formalism can be developed. These descriptors are termed *local non-stochastic and stochastic bilinear indices*, $b_{mkL}(\bar{x}_m, \bar{y}_m)$ and ${}^{s}b_{mkL}(\bar{x}_m, \bar{y}_m)$, respectively. The definition of these descriptors is as follows:

$$b_{mkL}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}m_{ijL}\, x_m^{\,i}\, y_m^{\,j} \qquad (\mathbf{13})$$

$${}^{s}b_{mkL}(\bar{x}_m, \bar{y}_m) = \sum_{i=1}^{n} \sum_{j=1}^{n} {}^{k}sm_{ijL}\, x_m^{\,i}\, y_m^{\,j} \qquad (\mathbf{14})$$

where ${}^{k}m_{ijL}$ [${}^{k}sm_{ijL}$] is the k^{th} element of the row "*i*" and column "*j*" of the local matrix M^{k}_{mL} [${}^{s}M^{k}_{mL}$]. This matrix is extracted from the M^{k}_{m} [${}^{s}M^{k}_{m}$] matrix and contains information referred to the vertices of the specific nucleic acid fragments (F_r) and also of the molecular environment in *k* step. The matrix M^{k}_{mL} [${}^{s}M^{k}_{mL}$] with elements ${}^{k}m_{ijL}$ [${}^{k}sm_{ijL}$] is defined as follows (see Table 5 and 6 for the performance of M^{k}_{mL} and ${}^{s}M^{k}_{mL}$ practical examples):

${}^{k}m_{ijL}$ [${}^{k}sm_{ijL}$] = ${}^{k}m_{ij}$ [${}^{k}sm_{ijL}$] *if* both v_i and v_j are vertices (amino-acid) contained within the F_r

= $^1/_2\, {}^{k}m_{ij}$ [${}^{k}sm_{ijL}$] *if* v_i or v_j are vertices contained within F_r but not both

= 0 otherwise. (**15**)

Tables 5 and 6 comes about here (see end of the document)

These local analogues can also be expressed in matrix form by the expressions:

$$b_{mkL}(\bar{x}_m, \bar{y}_m) = [X_m]^T M_m^{\boldsymbol{k}_L} [Y_m] \quad (\mathbf{16})$$

$$^s b_{mk}(\bar{x}_m, \bar{y}_m) = [X_m]^T {}^s M_m^{\boldsymbol{k}_L} [Y_m] \quad (\mathbf{17})$$

It should be remarked that the scheme above follows the spirit of a Mulliken population analysis (D.Walker, 1993). It should be also pointed out that for every partitioning of a nucleic acid into Z macromolecular fragments there will be Z local macromolecular fragment matrices. In this case, if a nucleic acid is partitioned into Z molecular fragments, the matrix M_m^k [$^s M_m^k$] can be correspondingly partitioned into Z local matrices M_{mL}^k [$^s M_{mL}^k$], L = 1,..., Z, and the k^{th} power of matrix M_m^k [$^s M_m^k$] is exactly the sum of the k^{th} power of the local Z matrices. In this way, the total non-stochastic and stochastic bilinear indices are the sum of the non-stochastic and stochastic bilinear indices, respectively, of the Z macromolecular fragments (see Table 7 for a realistic example):

$$b_m(\bar{x}_m, \bar{y}_m) = \sum_{L=1}^{Z} b_{mkL}(\bar{x}_m, \bar{y}_m) \quad (\mathbf{18})$$

$$^s b_m(\bar{x}_m, \bar{y}_m) = \sum_{L=1}^{Z} {}^s b_{mkL}(\bar{x}_m, \bar{y}_m) \quad (\mathbf{19})$$

In addition, the nucleotide-type bilinear indices can also be calculated. Nucleotide and nucleotide-type bilinear indices are specific cases of local nucleic acid bilinear indices. In this sense, the k^{th} nucleotide bilinear indices are calculated by summing the k^{th} nucleotide bilinear indices of all nucleotide of the same nucleotide type in the nucleic acid. Any local nucleic acid's bilinear 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 bio-macromolecular graph.

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 (Randić, 1991; Todeschini and Consonni, 2000). The local bio-macromolecular indices can also be used together with total ones as variables for QSAR/QSPR modelling of properties or activities that depend more on a region or a fragment than on the macromolecule as a whole.

Table 7 comes about here (see end of the document)

3. MATERIAL AND METHODS

3.1. Computational Strategies

TOMOCOMD is an interactive program for molecular design and bioinformatics research (Marrero-Ponce and Romero, 2002). 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). 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 (nucleotide) macromolecular bilinear indices for any nucleic acids was implemented in the *TOMOCOMD-CANAR* software (Marrero-Ponce and Romero, 2002). The following list briefly resumes the main steps for the application of this method in QSAR/QSPR:

1. Draw the bio-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**) (Marrero-Ponce and Romero, 2002). 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 bilinear indices of the k^{th} non-stochastic and stochastic nucleotide's graph–theoretic electronic-contact matrix of G_m, M_m^k and ${}^sM_m^k$, respectively. 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 bio-macromolecular indices. This software generates a table in which the rows and columns correspond to the compounds and the $b_{mk}(\bar{x}_m, \bar{y}_m)$, correspondingly.
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 $b_{mk}(\bar{x}_m, \bar{y}_m)$ having, for instance, the following appearance,

$$\boldsymbol{P} = a_0\, b_{m0}(\bar{x}_m, \bar{y}_m) + a_1 b_{m1}(\bar{x}_m, \bar{y}_m) + a_2 b_{m2}(\bar{x}_m, \bar{y}_m) + \ldots + a_k b_{mk}(\bar{x}_m, \bar{y}_m) + \text{c} \quad (\mathbf{20})$$

where ***P*** is the measurement of the property, $b_{mk}(\bar{x}_m, \bar{y}_m)$ [or $b_{mkL}(\bar{x}_m, \bar{y}_m)$] is the k^{th} total [or local] bio-macromolecular bilinear indices, an the $\boldsymbol{a_k}$*'s* are the coefficients obtained by the statistical analysis.

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

3.2. Data Sets

The data set of footprinted and binding nucleotides was extracted from the literature (McPike et al., 2002). 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.

Figure 1 comes about here (see end of the document)

3.3. Chemometric Analysis: Regression-Based QSAR Model.

Based on the discussion above, a simple linear model was proposed to predict drug–nucleotide affinity. Multiple Linear Regression (MLR) statistical technique was used to obtain a quantitative model. This statistical analysis was carried out with the STATISTICA

software package (Statsoft, 1999). *TOMOCOMD-CANAR* model used for the statistical procedure the first 16 $b_{mkL}(\bar{x}_m, \bar{y}_m)$ [from $b_{m0L}(\bar{x}_m, \bar{y}_m)$ to $b_{m15L}(\bar{x}_m, \bar{y}_m)$] 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}) (Golbraikh and Tropsha, 2002).

4. RESULTS AND DISCUSSION

In order to prove the applicability of this new approach, quantitative linear models based on local (nucleotide) non-stochastic and stochastic bilinear indices were obtained by using LMR, with the aim to predict the interaction strength between Paromomycin and its binding ribonucleotides within HIV packaging region. The found equations show the relatedness of this method. These were selected taking into account several statistical parameters listed below. Next it is showed the best two non-stochastic and stochastic equations, respectively (others two can be seen in Table 8):

$$\textbf{Log } \boldsymbol{K} = 0.689\ (\pm 0.044) + 0.016\ (\pm 0.001)\ {}^{f_2 - \in_{260}} b_{0L}(\bar{x}_m, \bar{y}_m)$$

$$- 1.5 \times 10^{-5}\ (\pm 2.0 \times 10^{-6})\ {}^{\in_{260} - E_1} b_{5L}(\bar{x}_m, \bar{y}_m)$$

$$+ 2.04 \times 10^{-4} (\pm 0.34 \times 10^{-4}) \; {}^{F_2 - E_2} b_{5L}(\bar{x}_m, \bar{y}_m) \qquad \textbf{(21)}$$

N = 24 R = 0.95 R^2 = 0.91 s = 0.08 q^2 = 0.86 s_{cv} = 0.09 F (3, 19) = 60.71 p< 0.0001

$$\textbf{Log } \boldsymbol{K} = 0.307 (\pm 0.137) + 0.016 (\pm 0.001) \; {}^{f_1 - f_2} s b_{0L}(\bar{x}_m, \bar{y}_m)$$

$$- 1.295 \times 10^{-15} (\pm 1.788 \times 10^{-16}) \; {}^{f_1 - \in_{260}} s b_{15L}(\bar{x}_m, \bar{y}_m)$$

$$+ 0.051 (\pm 0.013) \; {}^{\in_{260} - E_1} s b_{2L}(\bar{x}_m, \bar{y}_m)$$

$$- 0.050 (\pm 0.012) \; {}^{\in_{260} - E_2} s b_{2L}(\bar{x}_m, \bar{y}_m) \qquad \textbf{(22)}$$

N = 24 R = 0.94 R^2 = 0.89 s = 0.10 q^2 = 0.79 s_{cv} = 0.11 F (4, 18) = 36.88 p< 0.0001

where N is the number of interactions with 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 cross-validation experiments.

Table 8 comes about here (see end of the document)

In the development of the quantitative models for the Log*K* description of the calibration data set, one nucleotide (**A276**) highlights as a statistical outlier. Outlier detection was performed using the following standard statistical test: residual, standardized residuals, Studentized residual and Cook's distance.

Equations **21** and **22** successfully explained about 91% and 89%, correspondingly, of the variability in the data for the interaction magnitudes between the aminoglycoside and HIV. LOO cross-validation procedure was chosen to test predictability and stability of these models. The squared cross-validation regression coefficients showed that models **21** and **22** accounted for 86% and 79%, respectively, of the data variability in cross-validation

study, what could be an indicator of both stability and predictability (Golbraikh and Tropsha, 2002). The results for the residual analysis are depicted in Table 9.

Table 9 comes about here (see end of the document)

Therefore, taking into account statistical parameters in both non-stochastic and stochastic equations (Eqs. **21** and **22**, respectively) it can be said that they are appropriated for description of interaction magnitude between the antibiotic and HIV packaging region.

Statistical parameters in non-stochastic equation (**21**) suggest a high quality of found model. Consequently, non-stochastic model must be preferred instead stochastic what suggest that non-stochastic local bilinear indices are better than stochastic in quantitative description of bio-macromolecular structure.

Some authors have reported similar equations at the introduced here. For instance Marrero-Ponce applied quadratic (Marrero-Ponce et al., 2004d) and linear indices (Marrero Ponce et al., 2005) with the same purpose. In the development of the quadratic indices based model for the Log *K* description, it was too detected the nucleotide (A276) as statistical outlier.

Likewise in González-Díaz *et al.* work's (2003) it was developed similar equations in order to predict antibiotic-nucleotide interaction magnitudes using ***MARCH-INSIDE*** (González-Díaz et al., 2003a) based descriptors. They additionally make use of a ***dummy*** variable RNAse, which has the values RNAse = 1 for experiments performed in the presence of RNAse I and RNAse = -1 for RNAse T1**.** Table 8 shows a comparison between ours models with approaches previously described (González-Díaz et al., 2003a; González-Díaz et al., 2003b; Marrero-Ponce et al., 2004d; Marrero Ponce et al., 2005)**.**

As can be observed in Table 8, the present results are similar-to-better to previously report (González-Díaz et al., 2003a; González-Díaz et al., 2003b; Marrero-Ponce et al.,

2004d; Marrero Ponce et al., 2005), showing the best LOO press statistic parameters. It is rather important to remarkable that our models not use dummy variables like Gonzalez-Diaz equations (González-Díaz et al., 2003a; González-Díaz et al., 2003b; Marrero-Ponce et al., 2004d; Marrero Ponce et al., 2005)), which used experimental information (RNAse *dummy* variable) in addition to structural (nucleotide) descriptors (MARCH-INSIDE Method).

On the other hand, the LMR-QSAR models (Eqs. **21-24**, see also Table 8) involve short-reaching [$k \leq 3$, i.e., ${}^{f_2-\epsilon_{260}}b_{0L}(\bar{x}_m, \bar{y}_m)$, ${}^{f_1-f_2}{}^{s}b_{0L}(\bar{x}_m, \bar{y}_m)$, ${}^{f1\text{-}f2}{}^{s}\boldsymbol{b}_{1L(\bar{x}_m, \bar{y}_m)}$, ${}^{\epsilon_{260}-E_1}{}^{s}b_{2L}(\bar{x}_m, \bar{y}_m)$, ${}^{\epsilon_{260}-E_2}{}^{s}b_{2L}(\bar{x}_m, \bar{y}_m)$], middle-reaching [$4 < k \leq 9$, i.e., ${}^{\epsilon_{260}-E_1}b_{5L}(\bar{x}_m, \bar{y}_m)$, ${}^{F_2-E_2}b_{5L}(\bar{x}_m, \bar{y}_m)$, ${}^{f2\text{-}\epsilon260}\boldsymbol{b}_{7L(\bar{x}_m, \bar{y}_m)}$] and far-reaching [$k = 10$ or greater i.e., ${}^{f_1-\epsilon_{260}}{}^{s}b_{15L}(\bar{x}_m, \bar{y}_m)$]. The RNA (nucleotide) bilinear indices of cero order ($k = 0$) characterized each kind of RNA bases (nucleotide), but not consider the environmental topology of the nucleotide. In all models these indices have a positive contribution. 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. Others RNA-bilinear indices of short-reaching 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 is necessary for the drug-nucleotide interaction, the more marked structural changes in the ±1 and ±2-vicinity of the nucleotides. The contribution of the middle-to-high reaching, ±5, ±7 and ±15-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 (middle-to-long-range interactions). 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 (McPike et al., 2002).

5. CONCLUDING REMARKS

Although there have been many discoveries in the last years in the field of bioinformatics, it is necessary the definition of novel bio-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 theoretical-biology studies. It presents a new set of bio-macromolecular descriptors that are calculated by using bilinear forms, which are relevant to nucleic acid QSAR/QSPR studies. Their derivation is straightforward, and it is easy to interpret the QSARs/QSPRs which include them.

We have shown here that the use of the local (nucleotide) nucleic acid bilinear indices is able to depict the affinity with which paromomycin binds to the HIV-1 W-RNA packaging region. The resulting models are significant of the statistical point of view. The models found to describe the interaction profile include nucleotide's bilinear 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 bilinear indices. This situation points to that the interaction between Paromomycin and a nucleotide of RNA depends on the electro-topologic environment of the nucleotides. Finally, the satisfactory comparative results showed that nucleic acid bilinear indices used here will be a novel chem- and bio-informatics tool for further research.

Acknowledgement: Sadiel Ortega-Broche (O-B. S) acknowledges to Bioinformatics Research Center of Central University 'Marta Abreu'of Las Villas for kind hospitality during the 2006-2007. Yovani Marrero-Ponce (M-P. Y) acknowledges the Valencia University for kind hospitality during the first semester of 2008. M-P. Y thanks are given to the Valencia University, (Spain) for partial financial support as well as the program 'Estades Temporals per an Investigadors Convidats' for a fellowship to work at Pharmacy Faculty (2008). M-P. Y also thanks support from Spanish MEC (Project Reference: SAF2006-04698). Finally, but not less, CAMD-BIR Unit thanks are given to the research project called: "*Strengthening Postgraduate Education and Research in Pharmaceutical Sciences*". This project is funded by the Flemish Interuniversity Council (VLIR) of Belgium.

6. REFERENCES

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. Eds.), 1994. Molecular Biology of the Cell. Garland, New York and London.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., and Wheeler, D.L., 2000. GenBank. Nucleic Acid Res 28, 15-18.

Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K., 1986. Methods Enzymol. 130, 132.

Burgos-Román, J.d. (Ed.), 1994. Curso de Álgebra y Geometría, Madrid, Spain.

Burgos-Román, J.d. (Ed.), 2000. Álgebra y Geometría Cartesíana. .

Casañola-Martin, G.M., Khan, M.T.H., Marrero-Ponce, Y., Ather, A., Sultankhodzhaev, M.N., and Torrens, F., 2006. New Tyrosinase Inhibitors Selected by Atomic Linear Indices-Based Classification Models. Bioorg. Med. Chem. Letter. 16, 324-330.

Castillo-Garit, J.A., Marrero-Ponce, Y., Torrens, F., and Rotondo, R., 2007. Atom-Based Stochastic and Non-Stochastic 3D-Chiral Bilinear Indices and Their Applications to Central Chirality Codification. J. Mol. Graph. Modell. 26, 32-47.

D.Walker, P.G.M., 1993. J. Am. Chem. Soc. 115, 12423.

Edwards, C.H., and Penney, D.E., 1988. Elementary Linear Algebra. Prentice-Hall, Englewood Cliffs, New Jersey, USA

Galas, D.J., and Schmithz, A., 1978. Nucleic Acid Res. 5, 3157.

Gale, E.F., Gundliff, E., Reynolds, P.E., Richmon, M.H., and Waring, M.J. Eds.), 1981. The Molecular Basis of Antibiotic Action, London.

Golbraikh, A., and Tropsha, A., 2002. Beware of q2! J Mol Graph Model 20, 269-76.

Gonzalez-Diaz, H., and Uriarte, E., 2005. Proteins QSAR with Markov average electrostatic potentials. Bioorg Med Chem Lett 15, 5088-94.

Gonzalez-Diaz, H., Uriarte, E., and Ramos de Armas, R., 2005. Predicting stability of Arc repressor mutants with protein stochastic moments. Bioorg Med Chem 13, 323-31.

González-Díaz, H., Ramos de Armas, R., and Molina, R., 2003a. Vibrational Markovian Modelling of Footprints after the Interaction of Antibiotics with the Packaging Region of HIV Type 1. Bull. Math. Biol. 65, 991-1002.

González-Díaz, H., Ramos de Armas, R., and Molina, R., 2003b. Markovian Negentropies in Bioinformatics. 1. A Picture of Footprints after the Interaction of the HIV-1 RNA Packaging Region with Drugs. Bioinformatics 19, 2079-2087.

Hamasaki, K., and Akihiko, U., 2001. Aminoglycoside Antibiotics, neamine and Its derivatives as Potent Inhibitors for the RNA-Proteins interactions Derived from HIV-1 Activators Biorganic & Medicinal Chemistry Letters 11, 591-594.

Henn, A., Halfon, J., Kela, I., Orion, I., and Sagi, I., 2001. Nucleic Acids Res. 29, 122.

Hernández, E. (Ed.), 1987. Álgebra y Geometría, Madrid, Spain.

Hua, S., and Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17, 721-8.

I. Gutman, a.O.E.P. (Ed.), 1986. Mathematical Concepts in Organic Chemistry, Berlin.

Jacobson, N. (Ed.), 1985. Basic Algebra I New York.

K. F. Riley, M.P.H., and S. J. Vence (Ed.), 1998. Mathematical Methods for Physics and Engineering.

Lehninger, A.L., Nelson, D.L., and Cox, M.M. Eds.), 1993. Principles of Biochemistry, New York.

Lynch, S.R., Recht, M.I., and Puglisi, J.D., 2000. Biochemical and Nuclear Magnetic Resonance Studies of Aminoglycoside-RNA Complexes. . Meth. Enzymol. 317, 240-261.

Marrero-Ponce, Y., 2003. Total and Local Quadratic Indices of the Molecular Pseudograph´s Atom Adjacency Matrix: Applications to the Prediction of Physical Properties of Organic Compounds. Molecules 8, 687-726.

Marrero-Ponce, Y., 2004a. 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 44, 2010-26.

Marrero-Ponce, Y., 2004b. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. Bioorg. Med. Chem. 12, 6351-6369.

Marrero-Ponce, Y., and Romero, V., TOMOCOMD-CARDD software. TOMOCOMD (TOpological MOlecular COMputer Design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version can be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es Central University of Las Villas, Santa Clara, Villa Clara 2002.

Marrero-Ponce, Y., Huesca-Guillen, A., and Ibarra-Velarde, F., 2005a. Quadratic indices of the ¨molecular pseudograph´s atom adjacency matrix¨ and their stochastic forms: a novel approach for virtual screening and in silico discovery of new lead paramphistomicide drugs-like compounds. J. Mol. Struct. (Theochem) 717, 67-79.

Marrero-Ponce, Y., Cabrera, M.A., Romero, V., Ofori, E., and Montero, L.A., 2003. 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. 4, 512-36.

Marrero-Ponce, Y., Castillo-Garit, J.A., Torrens, F., Romero-Zaldivar, V., and Castro, E., 2004a. Atom, Atom-Type, and Total Linear Indices of the Molecular Pseudograph's Atom Adjacency Matrix: Application to QSPR/QSAR Studies of Organic Compounds. Molecules 9, 1100-1123.

Marrero-Ponce, Y., Díaz, H.G., Romero, V., Torrens, F., and Castro, E.A., 2004b. 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 r-receptor antagonist activities. Bioorg. Med. Chem. 12, 5331-5342.

Marrero-Ponce, Y., Castillo-Garit, J.A., Castro, E.A., Torrens, F., and Rotondo, R., 2008a. 3D-Chiral (2.5) Atom-Based TOMOCOMD-CARDD Descriptors: Theory and QSAR Applications to Central Chirality Codification. J. Math. Chem. DOI 10.1007/s10910-008-9386-3.

Marrero-Ponce, Y., Torrens, F., García-Domenech, R., Ortega-Broche, S.E., and Romero Zaldivar, V., 2008b. Novel 2D TOMOCOMD-CARDD Descriptors: Atom-Based Stochastic and Non-Stochastic Bilinear Indices and Their QSPR Applications. J. Math. Chem., DOI 10.1007/s10910-008-9389-0.

Marrero-Ponce, Y., Medina-Marrero, R., Castillo-Garit, J.A., Romero-Zaldivar, V., Torrens, F., and Castro, E.A., 2005b. Protein linear indices of the 'macromolecular pseudograph alpha-carbon atom adjacency matrix' in bioinformatics. Part 1: prediction

of protein stability effects of a complete set of alanine substitutions in Arc repressor. Bioorg Med Chem 13, 3003-15.

Marrero-Ponce, Y., Medina-Marrero, R., Torrens, F., Martinez, Y., Romero-Zaldivar, V., and Castro, E.A., 2005c. Atom, atom-type, and total nonstochastic and stochastic quadratic fingerprints: a promising approach for modeling of antibacterial activity. Bioorg Med Chem 13, 2881-99.

Marrero-Ponce, Y., Montero-Torres, A., Zaldivar, C.R., Veitia, M.I., Perez, M.M., and Sanchez, R.N., 2005d. Non-stochastic and stochastic linear indices of the 'molecular pseudograph's atom adjacency matrix': application to 'in silico' studies for the rational discovery of new antimalarial compounds. Bioorg Med Chem 13, 1293-304.

Marrero-Ponce, Y., Cabrera, M.A., Romero-Zaldivar, V., Bermejo, M., Siverio, D., and Torrens, F., 2005e. Prediction of Intestinal Epithelial Transport of Drug in (Caco-2) Cell Culture from Molecular Structure using in silico Approaches During Early Drug Discovery. Internet Electron. J. Mol. Des. 4 124-150.

Marrero-Ponce, Y., Medina, R., Castro, E.A., de Armas, R., González, H., Romero, V., and Torrens, F., 2004c. Protein Quadratic Indices of the "Macromolecular Pseudograph's α-Carbon Atom Adjacency Matrix". 1. Prediction of Arc Repressor Alanine-mutant's Stability. Molecules 9 1124-1147.

Marrero-Ponce, Y., Nodarse, D., González, H.D., Ramos de Armas, R., Romero-Zaldivar, V., Torrens, F., and Castro, E., 2004d. 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. Int. J. Mol. Sci. 5, 276-293.

Marrero-Ponce, Y., Khan, M.T.H., Casañola-Martín, G.M., Ather, A., Sultankhodzhaev, M.N., Torrens, F., and Rotondo, R., 2007. Prediction of Tyrosinase Inhibition Spectra for Chemicals Using Novel Atom-Based Bilinear Indices. CheMedChem 2, 449–478.

Marrero-Ponce, Y., Iyarreta-Veitia, M., Montero-Torres, A., Romero-Zaldivar, C., Brandt, C.A., Avila, P.E., Kirchgatter, K., and Machado, Y., 2005f. Ligand-based virtual screening and in silico design of new antimalarial compounds using nonstochastic and stochastic total and atom-type quadratic maps. J Chem Inf Model 45, 1082-100.

Marrero-Ponce, Y., Marrero, R.M., Torrens, F., Martinez, Y., Bernal, M.G., Zaldivar, V.R., Castro, E.A., and Abalo, R.G., 2006a. Non-stochastic and stochastic linear indices of the molecular pseudograph's atom-adjacency matrix: a novel approach for computational in silico screening and "rational" selection of new lead antibacterial agents. J. Mol. Mod. 12, 255–271.

Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castanedo, N., Ibarra-Velarde, F., Huesca-Guillen, A., Sanchez, A.M., Torrens, F., and Castro, E.A., 2005g. Atom, atom-type and total molecular linear indices as a promising approach for bioorganic and medicinal chemistry: theoretical and experimental assessment of a novel method for virtual screening and rational design of new lead anthelmintic. Bioorg Med Chem 13, 1005-20.

Marrero-Ponce, Y., Castillo-Garit, J.A., Olazabal, E., Serrano, H.S., Morales, A., Castanedo, N., Ibarra-Velarde, F., Huesca-Guillen, A., Jorge, E., del Valle, A., Torrens, F., and Castro, E.A., 2004e. TOMOCOMD-CARDD, a novel approach for computer-aided 'rational' drug design: I. Theoretical and experimental assessment of a promising

method for computational screening and in silico design of new anthelmintic compounds. J. Comput.-Aided Mol. Des. 18, 615-634.

Marrero-Ponce, Y., Meneses-Marcel, A., Machado-Tugores, Y., Montero Pereira, D., Escario, J.A., Nogal-Ruiz, J.J., Ochoa, C., Arán, V.J., Martínez-Fernández, A.R., García Sánchez, R.N., Montero-Torres, A., and Torrens., F., 2005h. A Computer-Based Approach to the Rational Discovery of New Antitrichomonas Drugs by Atom-Type Linear Indices. Curr. Drug Disc. Tech. 2, 245-265.

Marrero-Ponce, Y., Meneses-Marcel, A., Catillo-Garit, J.A., Machado-Tugores, Y., Escario, J.A., Gómez-Barrio, A., Montero Pereira, D., Nogal-Ruiz, J.J., Arán, V.J., Martínez-Fernández, A.R., Torrens, F., and Rotondo, R., 2006b. Predicting Antitrichomonal Activity: A Computational Screening Using Atom-Based Bilinear Indices and Experimental Proofs. Bioorg. Med. Chem. 14, 6502–6524.

Marrero-Ponce, Y.H.-G., A.; Ibarra-Velarde,, and F., 2005. J. Mol. Struct.(THEOCHEM) 717, 67-79.

Marrero Ponce, Y., 2004. Linear Indices of the ¨Molecular Pseudograph´s Atom Adjacency Matrix¨: Definition Significance-Interpretation, and Application to QSAR Analysis to Flavone Derivatives as HIV-1 Integrase Inhibitors. J Chem Inf Comput Sci 44, 2010-26.

Marrero Ponce, Y., Castillo Garit, J.A., and Nodarse, D., 2005. 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 psi-RNA packaging region. Bioorg Med Chem 13, 3397-404.

Marrero Ponce, Y., Cabrera Perez, M.A., Romero Zaldivar, V., Gonzalez Diaz, H., and Torrens, F., 2004. A new topological descriptors based model for predicting intestinal epithelial transport of drugs in Caco-2 cell culture. J Pharm Pharm Sci 7, 186-99.

Mathews, C.K., van Holde, K.E., and Ahern, K.G. Eds.), 2000. Biochemistry, San Francisco

McPike, P.M., Goodisman, J., and Dabrowiak, C.J., 2002. Footprinting and Circular Dichroims Studies on Paromomycin Binding to the Packaging Region of the Human Immunodeficiency Virus Type-1. Bioorg. Med. Chem. 10, 3663–3672.

Meneses-Marcel, A., Marrero-Ponce, Y., Machado-Tugores, Y., Montero-Torres, A., Pereira, D.M., Escario, J.A., Nogal-Ruiz, J.J., Ochoa, C., Aran, V.J., Martinez-Fernandez, A.R., and Garcia Sanchez, R.N., 2005a. A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: outcomes of in silico studies supported by experimental results. Bioorg Med Chem Lett 15, 3838-43.

Meneses-Marcel, A., Marrero-Ponce, Y., Machado-Tugores, Y., Montero-Torres, A., Montero Pereira, D., Escario, J.A., Nogal-Ruiz, J.J., Ochoa, C., Arán, V.J., Martínez-Fernández, A.R., and García Sánchez, R.N., 2005b. A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: Outcomes of in silico studies supported by experimental results. Bioorg. Med. Chem Lett. 17, 3838-3843.

Montero-Torres, A., Celeste Vega, M., Marrero-Ponce, Y., Rolón, M., Gómez-Barrio, A., Escario, J.A., Arán, V.J., Martínez-Fernández, A.R., and Meneses-Marcel, A., 2005. A novel non-stochastic quadratic fingerprints-based approach for the ‘in silico’ discovery of new antitrypanosomal compounds Bioorg. Med. Chem. 13, 6264–6275.

Montero-Torres, A., García-Sánchez, R.N., Marrero-Ponce, Y., Machado-Tugores, Y., Nogal-Ruiz, J.J., Martínez-Fernández, A.R., Arán, V.J., Ochoa, C., Meneses-Marcel, A., and Torrens, F., 2006. Non-stochastic quadratic fingerprints and LDA-based QSAR models in hit and lead generation through virtual screening: theoretical and experimental assessment of a promising method for the discovery of new antimalarial compounds. Eur. J. Med. Chem. 41, 483–493.

Österberg, F., Garrett, M.M., Sanner, M.F., Olson, A.J., and Goodsell, S.D., 2002. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock. Proteins Struc. Funct. Genet 46, 34.

Ozoline, O.N., Fujita, N., and Ishihama, A., 2001. Nucleic Acids Res. 29, 4909.

Pogliani, L., 2000. From Molecular Connectivity Indices to Semiempirical Connectivity Terms: Recent Trends in Graph Theoretical Descriptors. Chem. Rev. 100, 3827-3858.

Ramos de Armas, R., Gonzalez Diaz, H., Molina, R., and Uriarte, E., 2004. Markovian Backbone Negentropies: Molecular descriptors for protein research. I. Predicting protein stability in Arc repressor mutants. Proteins 56, 715-23.

Randić, M., 1991. Generalized Molecular Descriptors. J. Math. Chem. 7, 155-168.

Rouvray, D.H. (Ed.), 1976. In Chemical Applications of Graph Theory, London.

Sakharkar, M.K., Long, M., Tin, W.T., and Souza, S.J., 2000a. ExInt: an exon/intron database. Nucleic Acid Res 28, 191-192.

Sakharkar, M.K., Kangueane, P., Woon, T.W., Tan, T.W., Kolatkar, P.R., Long, M., and de Souza, S.J., 2000b. IE-KB:intron exon knowledge base. Bioinformatics 16, 1151-1152.

Saxonov, S., Daizadeh, I., Fedorov, A., and Gilbert, W., 2000. EID: the Exon-Intron Database-an Exhaustible Database of Protein Coding Intron-Containing Genes. Nucleic Acid Res 28, 185-190.

Schisler, N.J., and Palmer, J.D., 2000 The IDB and IEDB: intron sequence and evolution databases. Nucleic Acid Res 28, 181-184.

Statsoft, I., STATISTICA, 1999.

Stryer, L. (Ed.), 1995. Biochemistry, New York

Sullivan, J.M., Goodisman, J., and Dabrowiak, C.J., 2002. Absorption studies on aminoglycosides binding to the packaging region of the human immunodeficiency virus type-1. Bioorg. Med. Chem.Lett. 12 615-618.

Todeschini, R., and Consonni, V., 2000. Handbook of Molecular Descriptors. Wiley-VCH, Germany.

Trinajstić, N., 1983. Chemical Graph Theory. CRC Press, Boca Raton, FL.

Tullius, T.D., 1989. Annu. Rev. Biophys. Biophys. Chem. 18, 213.

Weiss, R., Teich, N., Varmus, H., and Coffin, J. Eds.), 1984. RNA Tumor Viruses, Cold Spring Harbor (N.Y.).

Werner, G. (Ed.), 1981. Linear Algebra, New York.

Wilson, W.D., and Li, K., 2000. Targeting RNA with Small Molecules. . Curr. Med. Chem. 7, 73-98.

Y. Marrero-Ponce, a.J.A.C.-G., 2005a. J. Comput.-Aided Mol. Des

19, 369.

Y. Marrero-Ponce, M.I.-V., A. Montero-Torres, C. Romero-Zaldivar, C. A. Brandt, P. E. Avila, K. Kirchgatter, and Y. Machado, 2005b. J. Chem. Inf. Comput. Sci 45, 1082.

Yuan, Z., 1999. Prediction of Protein Subcellular Location Using Markov Chain Models. FEBS Lett. 451, 23-26.

ANEXES
(Tables and Figures should be inserted in the main text)

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

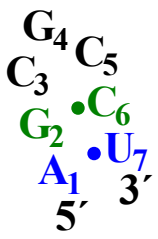
Purine and pyrimidine bases (RNA/ADN)	f_1	f_2	$\epsilon_{260}/1000$	ΔE_1	ΔE_2
Adenine (*A*)	0.28	0.54	15.4	4.75	5.99
Guanine (*G*)	0.20	0.27	11.7	4.49	5.03
Uracil (*U*)	0.18	0.3	9.9	4.81	6.11
Thymine (*T*)	0.18	0.37	9.2	4.67	5.94
Cytosine (*C*)	0.13	0.72	7.5	4.61	6.26

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 (Pogliani, 2000).

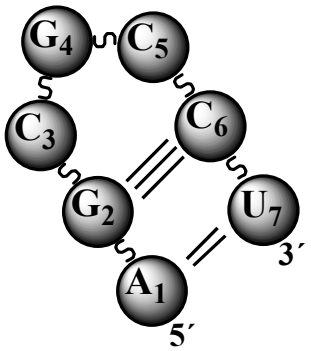
Table 2. Representation of the primary and secondary structures of a RNA sequence and its graph and bio-macromolecular vectors associated.

a 5′-AGCGCCU-3′

b



[a]Primary and [b]secondary structures of a RNA sequence. A dot between two base pairs means hydrogen-bond interactions.



Bio-Macromolecular graph´s (G_m) representation of the left RNA sequence. Each sphere represents one nucleotide. Nucleotides are covalently linked through phosphodiester linkage, represented as **S**. Hydrogen-bonds between bases are drawn as continued lines.

Macromolecular vector:

$\bar{x}_m$ = [A G C G C C U]; $\bar{x}_m \in \Re^7$

In the definition of $\bar{x}_m$, as macromolecular vector, the symbol of the bases is used to indicate the corresponding DNA or RNA base properties, for instance, f_1. That is: if we write A it means $f_{1(A)}$, adenine first oscillator strength value or some other base property, which characterizes each nucleotide in the nucleic acid molecule. So, if we use the canonical bases of $\Re^7$, the coordinates of any macromolecular vector $\bar{x}_m$ coincide with the components of that macromolecular vector.

$[X_m]^T$ = [A G C G C C U]

$[X_m]^T$ = transposed of $[X_m]$ and it means the vector of the coordinates of $\bar{x}_m$ in the canonical basis of R^7 (an 1x7 matrix)

$[X_m]$: vector of coordinates of $\bar{x}_m$ in Canonical base of $\Re^7$(a 7x1 matrix)

$\bar{x}_m, \bar{y}_m$ components are first (f_1) and second (f_2) oscillator strength values, respectively.

$\bar{x}_m$ =[0.28 0.20 0.13 0.20 0.13 0.13 0.18]

$\bar{y}_m$ =[0.54 0.27 0.72 0.27 0.72 0.72 0.37]

Table 3. Values of the total non-stochastic bilinear indices of *zero*, *first* and *second* orders for RNA fragment used as example above (see also Table 2).

Non-stochastic Total Bilinear Indices
$^{f_1-f_2}b_{m0}(\bar{x}_m,\bar{y}_m) = [X_m]^T M_m^0 [Y_m] = [0.28 \;\; 0.20 \;\; 0.13 \;\; 0.20 \;\; 0.13 \;\; 0.13 \;\; 0.18] \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.54 \\ 0.27 \\ 0.72 \\ 0.27 \\ 0.72 \\ 0.72 \\ 0.30 \end{bmatrix} = 0.594$
$^{f_1-f_2}b_{m1}(\bar{x}_m,\bar{y}_m) = [X_m]^T M_m^1 [Y_m] = [0.28 \;\; 0.20 \;\; 0.13 \;\; 0.20 \;\; 0.13 \;\; 0.13 \;\; 0.18] \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 3 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 3 & 0 & 0 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0.54 \\ 0.27 \\ 0.72 \\ 0.27 \\ 0.72 \\ 0.72 \\ 0.30 \end{bmatrix} = 1.976$
$^{f_1-f_2}b_{m2}(\bar{x}_m,\bar{y}_m) = [X_m]^T M_m^2 [Y_m] = [0.28 \;\; 0.20 \;\; 0.13 \;\; 0.20 \;\; 0.13 \;\; 0.13 \;\; 0.18] \begin{bmatrix} 5 & 0 & 1 & 0 & 0 & 5 & 0 \\ 0 & 11 & 0 & 1 & 3 & 0 & 5 \\ 1 & 0 & 2 & 0 & 1 & 3 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 3 & 1 & 0 & 2 & 0 & 1 \\ 5 & 0 & 3 & 1 & 0 & 11 & 0 \\ 0 & 5 & 0 & 0 & 1 & 0 & 5 \end{bmatrix} \begin{bmatrix} 0.54 \\ 0.27 \\ 0.72 \\ 0.27 \\ 0.72 \\ 0.72 \\ 0.30 \end{bmatrix} = 7.047$

Table 4. Values of the total stochastic bilinear indices of *zero*, *first* and *second* orders for RNA fragment used as example above (see also Table 2).

Stochastic Total Bilinear Indices
$${}^{f_1-f_2^s}b_{m0}(\bar{x}_m,\bar{y}_m)=[X_m]^{T\,s}M_m^0[Y_m]=\begin{bmatrix}0.28 & 0.20 & 0.13 & 0.20 & 0.13 & 0.13 & 0.18\end{bmatrix}\begin{bmatrix}1&0&0&0&0&0&0\\0&1&0&0&0&0&0\\0&0&1&0&0&0&0\\0&0&0&1&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&1\end{bmatrix}\begin{bmatrix}0.54\\0.27\\0.72\\0.27\\0.72\\0.72\\0.30\end{bmatrix}=0.594$$
$${}^{f_1-f_2^s}b_{m1}(\bar{x}_m,\bar{y}_m)=[X_m]^{T\,s}M_m^1[Y_m]=\begin{bmatrix}0.28 & 0.20 & 0.13 & 0.20 & 0.13 & 0.13 & 0.18\end{bmatrix}\begin{bmatrix}0&1/3&0&0&0&0&2/3\\1/5&0&1/5&0&0&3/5&0\\0&1/2&0&1/2&0&0&0\\0&0&1/2&0&1/2&0&0\\0&0&0&1/2&0&1/2&0\\0&3/5&0&0&1/5&0&1/5\\2/3&0&0&0&0&1/3&0\end{bmatrix}\begin{bmatrix}0.54\\0.27\\0.72\\0.27\\0.72\\0.72\\0.30\end{bmatrix}=0.617$$
$${}^{f_1-f_2^s}b_{m2}(\bar{x}_m,\bar{y}_m)=[X_m]^{T\,s}M_m^2[Y_m]=\begin{bmatrix}0.28 & 0.20 & 0.13 & 0.20 & 0.13 & 0.13 & 0.18\end{bmatrix}\begin{bmatrix}5/11&0&1/11&0&0&5/11&0\\0&1/20&0&1/20&3/20&0&5/20\\1/7&0&2/7&0&1/7&3/7&0\\0&1/4&0&2/4&0&1/4&0\\0&3/7&1/7&0&2/7&0&1/7\\5/20&0&3/20&1/20&0&11/20&0\\0&5/11&0&0&1/11&0&5/11\end{bmatrix}\begin{bmatrix}0.54\\0.27\\0.72\\0.27\\0.72\\0.72\\0.30\end{bmatrix}=0.618$$

Table 5. The zero (M^0), first (M^1) and second (M^2) powers of the local non-stochastic nucleotide graph–theoretic electronic-contact matrices of G_m (see also Table 2).

The zero, first and second powers of the local non-stochastic (nucleotide) matrices

$$M^0(G_m, A_1) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^1(G_m, A_1) = \begin{bmatrix} 0 & ½ & 0 & 0 & 0 & 0 & 1 \\ ½ & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^2(G_m, A_1) = \begin{bmatrix} 5 & 0 & ½ & 0 & 0 & 5/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ ½ & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M^0(G_m, A_1) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^1(G_m, A_1) = \begin{bmatrix} 0 & ½ & 0 & 0 & 0 & 0 & 1 \\ ½ & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^2(G_m, A_1) = \begin{bmatrix} 5 & 0 & ½ & 0 & 0 & 5/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ ½ & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M^0(G_m, G_2) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^1(G_m, G_2) = \begin{bmatrix} 0 & ½ & 0 & 0 & 0 & 0 & 0 \\ ½ & 0 & ½ & 0 & 0 & 3/2 & 0 \\ 0 & ½ & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^2(G_m, G_2) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 11 & 0 & ½ & 3/2 & 0 & 5/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & ½ & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5/2 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M^0(G_m, C_3) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^1(G_m, C_3) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & ½ & 0 & 0 & 0 & 0 \\ 0 & ½ & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & ½ & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^2(G_m, C_3) = \begin{bmatrix} 0 & 0 & ½ & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ ½ & 0 & 2 & 0 & ½ & 3/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & ½ & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M^0(G_m, G_4) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^1(G_m, G_4) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & ½ & 0 & ½ & 0 & 0 \\ 0 & 0 & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M^2(G_m, G_4) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & ½ & 0 & 2 & 0 & ½ & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 5. *Cont.*

$$M^0(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad M^1(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&\frac{1}{2}&0&0\\0&0&0&\frac{1}{2}&0&\frac{1}{2}&0\\0&0&0&0&\frac{1}{2}&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad M^2(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&\frac{3}{2}&0&0\\0&0&0&0&\frac{1}{2}&0&0\\0&0&0&0&0&0&0\\0&\frac{3}{2}&\frac{1}{2}&0&2&0&\frac{1}{2}\\0&0&0&0&0&0&0\\0&0&0&0&\frac{1}{2}&0&0\end{bmatrix}$$

$$M^0(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&0\end{bmatrix}\quad M^1(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&\frac{3}{2}&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{2}&0\\0&\frac{3}{2}&0&0&\frac{1}{2}&0&\frac{1}{2}\\0&0&0&0&0&\frac{1}{2}&0\end{bmatrix}\quad M^2(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&\frac{5}{2}&0\\0&0&0&0&0&0&0\\0&0&0&0&0&\frac{3}{2}&0\\0&0&0&0&0&\frac{1}{2}&0\\0&0&0&0&0&0&0\\\frac{5}{2}&0&\frac{3}{2}&\frac{1}{2}&0&11&0\\0&0&0&0&0&0&0\end{bmatrix}$$

$$M^0(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&1\end{bmatrix}\quad M^1(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&1\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{2}\\1&0&0&0&0&\frac{1}{2}&0\end{bmatrix}\quad M^2(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{5}{2}\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{2}\\0&0&0&0&0&0&0\\0&\frac{5}{2}&0&0&\frac{1}{2}&0&5\end{bmatrix}$$

Table 6. The zero ($^sM^0$), first ($^sM^1$) and second ($^sM^2$) powers of the local (nucleotide) stochastic graph–theoretic electronic-contact matrices of G_m (see also Table 2).

The zero, first and second powers of the local stochastic (nucleotide) matrices

$$^sM^0(G_m,A_1)=\begin{bmatrix}1&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^1(G_m,A_1)=\begin{bmatrix}0&1/6&0&0&0&0&2/6\\1/10&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\1/3&0&0&0&0&0&0\end{bmatrix}\quad ^sM^2(G_m,A_1)=\begin{bmatrix}5/11&0&5/110&0&0&5/22&0\\0&0&0&0&0&0&0\\1/14&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\1/8&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}$$

$$^sM^0(G_m,G_2)=\begin{bmatrix}0&0&0&0&0&0&0\\0&1&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^1(G_m,G_2)=\begin{bmatrix}0&1/6&0&0&0&0&0\\1/10&0&1/10&0&0&3/10&0\\0&1/4&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&3/10&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^2(G_m,G_2)=\begin{bmatrix}0&0&0&0&0&0&0\\0&11/20&0&1/40&3/40&0&1/8\\0&0&0&0&0&0&0\\0&1/8&0&0&0&0&0\\0&3/14&0&0&0&0&0\\0&0&0&0&0&0&0\\0&5/22&0&0&0&0&0\end{bmatrix}$$

$$^sM^0(G_m,C_3)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&1&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^1(G_m,C_3)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&1/10&0&0&0&0\\0&1/4&0&1/4&0&0&0\\0&0&1/4&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^2(G_m,C_3)=\begin{bmatrix}0&0&5/110&0&0&0&0\\0&0&0&0&0&0&0\\1/14&0&2/7&0&1/14&3/14&0\\0&0&0&0&0&0&0\\0&0&1/14&0&0&0&0\\0&0&3/40&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}$$

$$^sM^0(G_m,G_4)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&1&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^1(G_m,G_4)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&1/4&0&0&0\\0&0&1/4&0&1/4&0&0\\0&0&0&1/4&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}\quad ^sM^2(G_m,G_4)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&1/40&0&0&0\\0&0&0&0&0&0&0\\0&1/8&0&1/2&0&1/8&0\\0&0&0&0&0&0&0\\0&0&0&1/40&0&0&0\\0&0&0&0&0&0&0\end{bmatrix}$$

Table 6. *Cont.*

$$ {}^{s}M^{0}(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&1&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\end{bmatrix} \quad {}^{s}M^{1}(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&\frac{1}{4}&0&0\\0&0&0&\frac{1}{4}&0&\frac{1}{4}&0\\0&0&0&0&\frac{1}{10}&0&0\\0&0&0&0&0&0&0\end{bmatrix} \quad {}^{s}M^{2}(G_m,C_5)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&\frac{3}{40}&0&0\\0&0&0&0&\frac{1}{14}&0&0\\0&0&0&0&0&0&0\\0&\frac{3}{14}&\frac{1}{14}&0&\frac{2}{7}&0&\frac{1}{14}\\0&0&0&0&0&0&0\\0&0&0&0&\frac{5}{110}&0&0\end{bmatrix} $$

$$ {}^{s}M^{0}(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&1&0\\0&0&0&0&0&0&0\end{bmatrix} \quad {}^{s}M^{1}(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&\frac{3}{10}&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&\frac{1}{4}&0\\0&\frac{3}{10}&0&0&\frac{1}{10}&0&\frac{1}{10}\\0&0&0&0&0&\frac{1}{6}&0\end{bmatrix} \quad M^{2}(G_m,C_6)=\begin{bmatrix}0&0&0&0&0&\frac{5}{22}&0\\0&0&0&0&0&0&0\\0&0&0&0&0&\frac{3}{14}&0\\0&0&0&0&0&\frac{1}{8}&0\\0&0&0&0&0&0&0\\\frac{1}{8}&0&\frac{3}{40}&\frac{1}{8}&0&\frac{11}{20}&0\\0&0&0&0&0&0&0\end{bmatrix} $$

$$ {}^{s}M^{0}(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&1\end{bmatrix} \quad {}^{s}M^{1}(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&\frac{1}{3}\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{10}\\\frac{1}{3}&0&0&0&0&\frac{1}{6}&0\end{bmatrix} \quad {}^{s}M^{2}(G_m,U_7)=\begin{bmatrix}0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{8}\\0&0&0&0&0&0&0\\0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{14}\\0&0&0&0&0&0&0\\0&\frac{5}{22}&0&0&\frac{5}{110}&0&\frac{5}{11}\end{bmatrix} $$

Table 7. Values of the local non-stochastic and stochastic bilinear indices of zero, first and second orders, respectively, for the RNA fragment used as example above (see also Table 2).

Local Non-Stochastic Bilinear Indices			
Nucleotide	${}^{f_1-f_2}b_{0L}(\bar{x},\bar{y})$	${}^{f_1-f_2}b_{1L}(\bar{x},\bar{y})$	${}^{f_1-f_2}b_{2L}(\bar{x},\bar{y})$
Adenylate (A_1)	0.151	0.273	1.571
Guanylate (G_2)	0.054	0.450	1.188
Cytidylate (C_3)	0.094	0.179	0.698
Guanylate (G_4)	0.054	0.179	0.252
Cytidylate (C_5)	0.094	0.183	0.634
Cytidylate (C_6)	0.094	0.447	2.079
Uridylate (U_7)	0.054	0.266	0.626
RNA fragment	**0.594**	**1.976**	**7.048**
Local Stochastic Bilinear Indices			
Nucleotide	${}^{f_1-f_2}{}^{s}b_{0L}(\bar{x},\bar{y})$	${}^{f_1-f_2}{}^{s}b_{1L}(\bar{x},\bar{y})$	${}^{f_1-f_2}{}^{s}b_{2L}(\bar{x},\bar{y})$
Adenylate (A_1)	0.151	0.084	0.137
Guanylate (G_2)	0.054	0.100	0.075
Cytidylate (C_3)	0.094	0.068	0.081
Guanylate (G_4)	0.054	0.090	0.054
Cytidylate (C_5)	0.094	0.078	0.067
Cytidylate (C_6)	0.094	0.111	0.152
Uridylate (U_7)	0.054	0.086	0.058
RNA fragment	**0.594**	**0.617**	**0.618**

Table 8. Statistical parameters of the QSAR models obtained, by using different bio-macromolecular descriptors, to describe the magnitude of the interactions between the aminoglycosides and the packaging region of type-1 HIV.

Methods	R^2	s	q^2	s_{cv}	F	Equations	[b]Ref.
	0.91	0.08	0.86	0.09	60.71	**See Eq. 21**	This Report
Nucleotide Non-Stochastic Bilinear Indices	0.92	0.08	0.83	0.10	49.95	**(Eq. 23)** $\mathbf{Log}K$ = 0.450 (±0.098) + 0.008 (±0.001) $^{€260-\Delta E2}b_{0L}(\bar{x}_m, \bar{y}_m)$ - 6.98 x 10^{-4} (±7.0 x 10^{-5}) $^{\Delta E1-\Delta E2}b_{3L}(\bar{x}_m, \bar{y}_m)$ + 8.76 x 10^{-4} (± 1.04 x 10^{-4}) $^{f2-\Delta E2}b_{5L}(\bar{x}_m, \bar{y}_m)$ - 1.50 x 10^{-5} (±2.0 x 10^{-6}) $^{f2-€260}b_{7L}(\bar{x}_m, \bar{y}_m)$	This Report
Nucleotide Stochastic Bilinear Indices	0.84	0.11	0.74	0.12	33.33	**(Eq. 23)** $\mathbf{Log}K$ = 2.648 (±0.690) + 0.065 (±0.013) $^{f2-€260\ s}b_{0L}(\bar{x}_m, \bar{y}_m)$ + 1.34 x 10^{-15} (±2.08 x 10^{-16}) $^{f1-€260\ s}b_{15L}(\bar{x}_m, \bar{y}_m)$ - 0.017 (± 0.005) $^{f1-f2\ s}b_{1L}(\bar{x}_m, \bar{y}_m)$	This Report
	0.89	0.10	0.79	0.11	36.88	**See Eq. 22**	This Report
Nucleotide Linear Indices	0.87	0.10	0.82	0.108	31.61	$\mathbf{Log}K$ = -10.5 (±1.36) + 4.71(±0.57)$^{\Delta E1}f_{0L}(x_m)$ -2.6 x 10^{-5}(±3.35 x 10^{-6})$^{€260}f_{5L}(x_m)$ -0.099(±0.02)$^{€260}f_{0L}(x_m)$ -1.915(±0.450)$^{\Delta E2}f_{0L}(x_m)$	(Marrero Ponce et al., 2005)
Nucleotide Quadratic Indices	0.92	0.07	0.85	0.09	54.91	$\mathbf{Log}K$ = -1.3747(±0.3882) +0.1136(±0.0189)$^{\Delta E1}q_{0L}(x_m)$ -7.5608 x 10^{-5}(±9.9659 x 10^{-6}) $^{€260}q_{3L}(x_m)$ +0.0393(±0.0069)$^{f2}q_{3L}(x_m)$ -4.6544(±1.63 x 10^{-9})$^{\Delta E1}q_{10L}(x_m)$	(Marrero-Ponce et al., 2004d)
Markovian Negentropies	0.83	0.12	0.83	a	31.48	$\mathbf{Log}K$ = 0.693(±0.038) +0.338(±0.068)**RNA**se -0.102(±0.025) $^{1}\mathbf{O}(\Theta_{10})$ +0.083(±0.035) $^{4}\mathbf{O}(\Theta_{8})$	(González-Díaz et al., 2003b)
Stochastic Spectral Moments	0.91	0.08	0.86	a	50.44	$\mathbf{Log}K$ = 1.023+ 0.52(±0.04)**RNAse** -0.098(±0.01)$^{SR}\Gamma_0$ +3.606(±1.444) $^{SR}\Gamma_2$ -3.654(±1.606) $^{SR}\Gamma_3$	(González-Díaz et al., 2003a)

[a]Values are not reported.
[b]References.

Table 9. Observed, predicted and predicted (after LOO cross-validation procedure) values of Log *K* obtained from Eqs. **21** and **22**.

NUC	Obs[a]	Pre[b]	Pre-CV[c]	Pre[d]	Pre-CV[e]
A235	1.204	1.111	0.118	1.215	-0.015
A239	1.204	1.147	0.073	1.217	-0.019
G251	0.447	0.349	0.127	0.323	0.171
G254	0.447	0.497	-0.056	0.427	0.028
C267	0.903	0.902	0.002	0.918	-0.021
A268	0.903	0.984	-0.103	0.892	0.014
A269	0.903	1.029	-0.173	1.011	-0.127
A276	0.778	0.721	0.084	1.107	0.215
A286	0.845	0.854	-0.011	0.765	0.017
G328	0.845	0.862	-0.019	0.842	0.004
G329	0.845	0.863	-0.020	0.870	-0.029
G331	0.845	0.863	-0.020	0.791	0.061
G333	0.845	0.862	-0.019	0.932	-0.105
G335	0.778	0.743	0.038	0.757	0.095
G339	0.778	0.599	0.191	0.753	0.043
G340	0.778	0.730	0.052	0.588	0.228
G344	0.845	0.793	0.057	0.767	0.093
G346	0.845	0.848	-0.003	0.839	0.007
G363	0.415	0.530	-0.145	0.469	-0.065
G364	0.415	0.496	-0.097	0.519	-0.146
G365	0.415	0.526	-0.121	0.607	-0.211
G366	0.415	0.412	0.004	0.505	-0.117
G367	0.415	0.391	0.029	0.453	-0.043

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 using Eq. **21**; [d]Predicted and [e]Predicted by LOO cross-validation by using Eq. **22**.

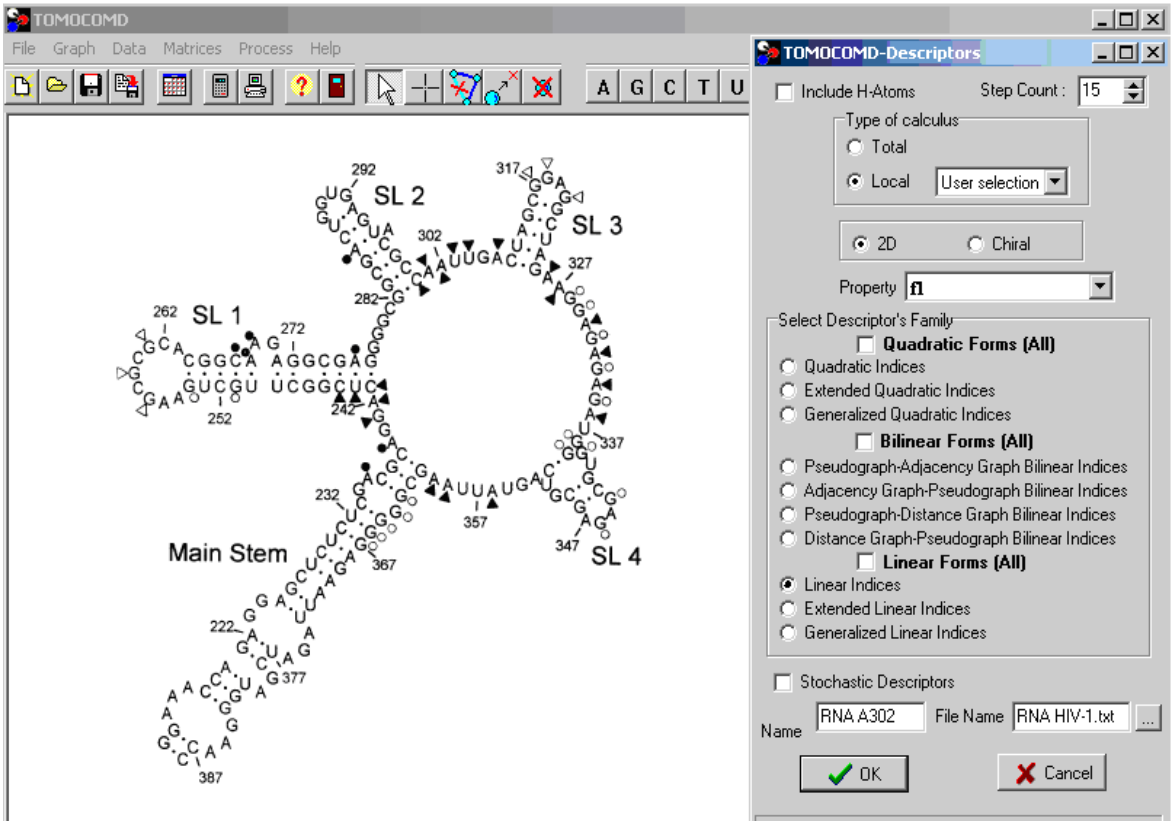


Figure 1. HIV-1 Ψ-RNA packaging region represented on the *TOMOCOMD-CANAR* interface. Nucleotides involved in binding and enhancement (structural changes) for RNAse I are shown as filled circles and triangles, respectively (open symbols indicates the use of RNAse T1).