Revista Cubana de Química
ISSN: 0258-5995
revcubanaquimica@cnt.uo.edu.cu
Universidad de Oriente
Cuba

González-Díaz, H.; Ramos de Armas, R.; Molina-Ruiz, R.; Cruz-Monteagudo, M.
MARKOVIAN NEGENTROPIES IN BIOINFORMATICS. A PICTURE OF FOOTPRINTS AFTER THE INTERACTION OF THE HIV-1 -RNA PACKAGING REGION WITH DRUGS.
Revista Cubana de Química, vol. XVII, núm. 2, 2005, pp. 216-226
Universidad de Oriente
Santiago de Cuba, Cuba

REVISTA CUBANA DE Química

MARKOVIAN NEGENTROPIES IN BIOINFORMATICS. A PICTURE OF FOOTPRINTS AFTER THE INTERACTION OF THE HIV-1 Ψ-RNA PACKAGING REGION WITH DRUGS.

H. González-Díaz [1, 2, a], R. Ramos de Armas[1, 3, b], R. Molina-Ruiz [1, 4, c], M. Cruz-Monteagudo [2, 5, d]

[1] Chemical Bioactives Center, Central University of "Las Villas" 54830, Cuba.

[2]Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela 15706, Spain.

[3] Department of Chemistry, Central University of "Las Villas" 54830, Cuba.

[4] Universität Rostock, FB Chemie, Albert-Einstein-Str. 3a, D 18059 Rostock, Germany.

[1]Applied Chemistry Research Center, Central University of "Las Villas" 54830, Cuba.

[a] humbertogd@uclv.edu.cu , [b] ronalr@uclv.edu.cu , [c] reymolina@uclv.edu.cu , [d] maikelcm@uclv.edu.cu

Abstract

Motivation: Many experts worldwide have highlighted the potential of RNA molecules as drug targets for the chemotherapeutic treatment of a range of diseases. In particular, the molecular pockets of RNA in the HIV-1 packaging region have been postulated as promising sites for antiviral action. The discovery of simpler methods to accurately represent drug–RNA interactions could therefore become an interesting and rapid way to generate models that are complementary to docking-based systems.

Results: The entropies of a vibrational Markov chain have been introduced here as physically meaningful descriptors for the local drug–nucleic acid complexes. A study of the interaction of the antibiotic Paromomycin with the packaging region of the RNA present in type-1 HIV has been carried out as an illustrative example of this approach. A linear discriminant function gave rise to excellent discrimination among 80.13% of interacting/non-interacting sites. More specifically, the model classified 36/45 nucleotides (80.0%) that interacted with paromomycin and, in addition, 85/106 (80.2%) footprinted (non-interacting) sites from the RNA viral sequence were recognized. The model showed a high Matthews' regression coefficient *(C* = 0.64*)*. The Jackknife method was also used to assess the stability and predictability of the model by leaving out adenines, C, G, or U. Matthews' coefficients and overall accuracies for these approaches were between 0.55 and 0.68 and 75.8 and 82.7, respectively. On the other hand, a linear regression model predicted the local binding affinity constants between a specific nucleotide and the aforementioned antibiotic (R^2 = 0.83, Q^2 = 0.825). These kinds of models may play an important role either in the discovery of new anti-HIV compounds or in the elucidation of their mode of action.

Keywords: *markovian negentropies, linear discriminant analysis, multiple linear regression, HIV-1 Ψ-RNA packaging, drug–RNA interactions*

Introduction

The number of new genomes has dramatically increased in recent years and this has once again highlighted the problem of protein and nucleic acid functions [1-6]. In this respect, the use of footprint techniques has proven to be an important experimental method for the discovery of significant processes in molecular biology and the field of genomics [7-12].

The study of drug interactions with biomolecules is a very important area of research and constitutes a significant step towards rational drug design. In this respect, modern Bioinformatics has emerged as a promising alternative and/or complement to existing experimental techniques. For example, the interactions of antibiotics (aminoglycosides) with the packaging region of HIV Type-1 appear to represent a possible new mechanism for antiviral discovery [13]. Aminoglycoside drugs are

cationic natural products that interact with RNA [14]. These compounds act by blocking protein synthesis [15, 16] and their structures can be chemically modified. In fact, aminoglycoside analogues can be used to treat certain diseases. For example, the genetic information in human immunodeficiency virus and various tumour viruses is in the form of RNA [17]. 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 [18].

Docking molecular mechanics approaches have been amongst the most widely used techniques to study host–guest molecular recognition. However, increasingly modern Bioinformatics methods (that do not involve docking techniques) have been utilized to provide structural information on Biomolecules as well as their integration with drugs [19]. In particular, Markov models are well-known tools for analysing biological sequence data and they have been used to find new genes from the open reading frames [20, 21], data based searching, multiple sequence alignment of protein families and protein domains [22], protein α-turn types [23], subcellular locations [6], for predicting secondary structures [24] and protein folding recognition [25].

Our research group has recently introduced the Markovian Chemicals Drug Design (MARCH-INSIDE) approach, which has been successfully applied to fluckicidal drugs design, chirality codification and protein structure/property relationships [26-28]. The main objectives of the work described here were focused on deriving quantitative structure/property relationships to predict the probability (and the affinity) of drugs binding with the packaging HIV Type-1 region. MARCH-INSIDE molecular descriptors were redefined as nucleic acid backbone Markovian Vibrational Negentropies (MVN) in order to account for the structures of RNA–drug complexes.

Methods

Modelling vibrational delocalization throughout the nucleic acid backbone

In this work the MARCH-INSIDE methodology was generalized for the study of nucleic acid–drug complexes. This approach used Markov's chain (MCH) [26-31] to codify information about molecular structure. The model considers a free drug molecule that it is in the vicinity of the RNA molecule at an arbitrary initial time (t_0) and interacts with some specific nucleotides of the RNA at a subsequent time (t_1). In order to characterize this model a number of important approximations must be taken into account. Firstly, we must accept the hypothesis that an elastic interaction governs drug–RNA binding [32]. Secondly, the energy of this interaction will dissipate with time throughout the RNA backbone with probabilities that obey Chapman–Kolgomorov equations; consequently, ${}^{k}\mathbf{\Pi} = ({}^{1}\mathbf{\Pi})^{k}$, where ${}^{k}\mathbf{\Pi}$ are stochastic matrices that describe the probabilities with which energy dissipates after a time t_k and ${}^{1}\mathbf{\Pi}$ is the same kind of matrix for the specific case $t_k = 1$, when interaction begins [33]. Accordingly, a stochastic matrix (${}^{1}\mathbf{\Pi}$, with elements p_{ij}) could be defined. This matrix was called the 1-step vibrational-transition stochastic matrix. ${}^{1}\mathbf{\Pi}$ was built as a squared table of order n, where n represented the number of nucleotides in the RNA molecule. The elements of ${}^{1}\mathbf{\Pi}$ (${}^{1}p_{ij}$) were defined to codify information about the strength (energy) of the drug–nucleotide interaction. Thus, when the drug molecule interacts with a specific nucleotide j of the RNA chain, the energy or 'strength' of the interaction may be described by a harmonic oscillator [32]. Also, this energy may be carried by a wave that 'vanishes' along the polynucleotide chain with absolute probabilities ${}^{A}p_k(j)$ at discrete time periods $t_k = 0, 1, 2 \ldots, \infty$. Subsequently, both the strength of the local interaction and the topology of the RNA (secondary structure) were codified in the present scheme [see equation (1) and Fig. 1].

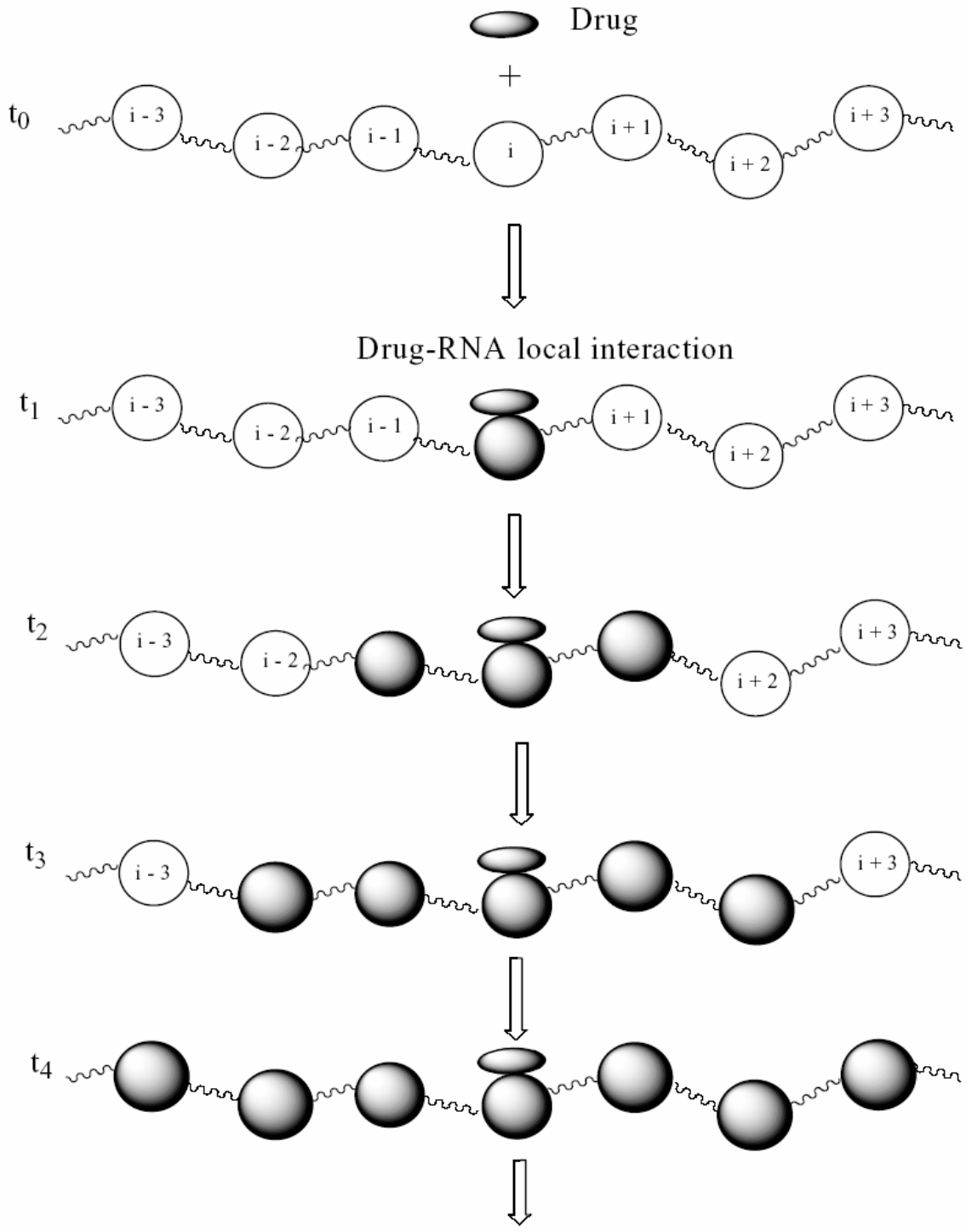


Fig. 1. Wave decay with time after the initial local Drug–RNA interaction (only the first five steps are represented); the change from a white to a black circle indicates that the vibration reaches this nucleotide (circle) at the period of time indicated to the left of the RNA chain (t_0–t_4). Black coloured nucleotides at t_4 are those in the $i \pm 3$ vicinity.

The elements ($^1p_{ij}$) of $^1\mathbf{\Pi}$ are the transition probabilities if the nucleotides *i* and *j* are adjacent and are 0 otherwise:

$$^1p_{ij} = \frac{\upsilon_j}{\sum_{k=1}^{\delta+1}\upsilon_k} = \frac{\sqrt{k_{j-\mathrm{drug}}/M_{j-\mathrm{drug}}}}{\sum_{k=1}^{\delta+1}\sqrt{k_{k-\mathrm{drug}}/M_{k-\mathrm{drug}}}} = \frac{\sqrt{(1+Hb_j)/M_{j-\mathrm{drug}}}}{\sum_{k=1}^{\delta+1}\sqrt{(1+Hb_k)/M_{k-\mathrm{drug}}}} \qquad (1)$$

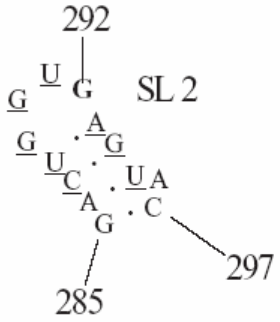
For the sake of simplicity, the interaction (nucleotide–drug) was considered harmonic with energy $h \times \upsilon j$ (*h* is Planck's constant), where υj is the harmonic frequency of vibration of the local complex between the nucleotide *j* and the drug. Thence, $k_{j-\mathrm{drug}}$ is the force constant of the link between the drug and the nucleotide *j* and $M_{k-\mathrm{drug}}$ is the reduced mass of this complex ($m_j \times m_{\mathrm{drug}}/m_j + m_{\mathrm{drug}}$). Another simplification was to consider $k_{j-\mathrm{drug}}$ as being proportional to the number of hydrogen bonding interactions Hb_j and the number of electrostatic interactions (1 account for this specific interaction) between the drug and the nucleotide. The parameter Hb_j equals the number of possible hydrogen bonds of each nucleotide, for nucleotides not hydrogen bonded to other nucleotides of the RNA chain. Conversely, $Hb_j = 0$ otherwise. To see more in detail the definition and calculation of the elements of the $^1\Pi$ matrix the study of a real example (nucleotide **G292**) is depicted in Table 1.

Table 1. A close up to the mathematical definition of $^1\Pi$ exemplifying with **G292**

Sub-matrix of $^1\Pi$ for the SL2 region of the HIV-1 RNA packaging region[a]

	...G	A	C	U	G	G	U	G	A	G	U	A	C...
				$j \pm 3$ vicinity of **G292**[b]									
G285	$^1p_{GG}$	$^1p_{GG}$	0	0	0	0	0	0	0	0	0	0	$^1p_{GC}$
A286	$^1p_{AG}$	$^1p_{AA}$	$^1p_{AC}$	0	0	0	0	0	0	0	$^1p_{AU}$	0	0
C287	0	$^1p_{CA}$	$^1p_{CC}$	$^1p_{CU}$	0	0	0	0	0	$^1p_{CG}$	0	0	0
U288	0	0	$^1p_{UC}$	$^1p_{UU}$	$^1p_{UG}$	0	0	0	$^1p_{UA}$	0	0	0	0
G289	0	0	0	$^1p_{GU}$	$^1p_{GG}$	$^1p_{GG}$	0	0	0	0	0	0	0
G290	0	0	0	0	$^1p_{GG}$	$^1p_{GG}$	$^1p_{GU}$	0	0	0	0	0	0
U291	0	0	0	0	0	$^1p_{UG}$	$^1p_{UU}$	$^1p_{UG}$	0	0	0	0	0
G292	0	0	0	0	0	0	$^1p_{GU}$	$^1p_{GG}$	$^1p_{GA}$	0	0	0	0
A293	0	0	0	$^1p_{AU}$	0	0	0	$^1p_{AG}$	$^1p_{AA}$	$^1p_{AG}$	0	0	0
G294	0	0	$^1p_{GC}$	0	0	0	0	0	$^1p_{GA}$	$^1p_{GG}$	$^1p_{GU}$	0	0
U295	0	$^1p_{UA}$	0	0	0	0	0	0	0	$^1p_{UG}$	$^1p_{UU}$	$^1p_{UA}$	0
A296	0	0	0	0	0	0	0	0	0	0	$^1p_{AU}$	$^1p_{AA}$	$^1p_{AC}$
C297	$^1p_{CG}$	0	0	0	0	0	0	0	0	0	0	$^1p_{CC}$	$^1p_{CA}$

Secondary structure of an RNA fragment of the SL 2 motif (see Fig. 2).[c]



Mathematical expression of the $^1p_{ij}$ element of $^1\Pi$ When i = **G292** and j = A293.

$$^1p_{\mathbf{G292}\underline{\mathrm{A293}}} = \frac{\upsilon_{\underline{\mathrm{A293}}}}{\upsilon_{\underline{\mathrm{U291}}} + \upsilon_{\mathbf{G292}} + \upsilon_{\underline{\mathrm{A293}}}}$$

$$\upsilon_{\underline{\mathrm{A293}}} = \sqrt{\frac{(1 + Hb_{\underline{\mathrm{A293}}}) \cdot (m_{\mathrm{Paromycin}} + m_A)}{m_{\mathrm{Paromycin}} \cdot m_A}}$$

$$\upsilon_{\mathbf{G292}} = \sqrt{\frac{(1 + Hb_{\mathbf{G292}}) \cdot (m_{\mathrm{Paromycin}} + m_G)}{m_{\mathrm{Paromycin}} \cdot m_G}}$$

$$\upsilon_{\underline{\mathrm{U291}}} = \sqrt{\frac{(1 + Hb_{\underline{\mathrm{U291}}}) \cdot (m_{\mathrm{Paromycin}} + m_U)}{m_{\mathrm{Paromycin}} \cdot m_U}}$$

[a] The $^1\Pi$ matrix of the HIV-1 RNA packaging region as a hole contains too many (more than 300) nucleotides to be represented here (see Fig. 2).
[b,c] The nucleotides in the $i \pm 3$ vicinity of **G292** are underlined, note that this set includes nucleotides that the vibration reach by first time at $t_k = 3$, after an initial interaction of the drug with **G292**, including both covalent (U295, G294,A293, and U291, G290, G289) or hydrogen bonds (U288, C289) for wave transmission.

The Markovian vibrational entropies

The extended MARCH-INSIDE RNA–drug complex molecular descriptors: Θ (j −drug ±3) were defined as the entropies with which the energy of the drug–nucleotide interaction vanished in the ±3-vicinity of the nucleotide along the time (t_k):

$$\Theta_k(j \pm 3) = -\sum_{j-3}^{j+3} {}^A p_k(j) \log({}^A p_k(j)) \qquad (2)$$

The ±3-vicinity of the nucleotide j is a nucleotide sequence that contains the nucleotides at positions $j - 3$, $j - 2$, $j - 1$, j, $j +1$, $j +2$, $j +3$ (see Fig. 1). The number 3 was selected as a reasonable but a priori topological distance at which the energy of the local interaction dissipates. Thereafter, the symbol $\Theta_k(j \pm 3)$ was reduced to Θ_k for the sake of simplicity. These molecular descriptors for RNA–drug complexes can be interpreted as the entropy involved in the movement of the wave that carries the vibrational energy due to the interaction of the drug with the nucleotide j in the ±3-vicinity of this nucleotide after time t_k.

The calculation of the absolute probabilities was straightforward on the basis of classical results from Markov chains theory [30].

$$ {}^{A}\Pi_k = {}^{A}\Pi_0 \times ({}^{k}\Pi) \qquad (3) $$

where ${}^{A}\mathbf{\Pi}_k$ represent $1 \times n$ vectors whose elements ${}^{A}p_k(j)$ are the aforementioned absolute probabilities, ${}^{A}\mathbf{\Pi}_0$ represents a $1 \times n$ vector whose elements are the ${}^{A}p_0(j)$ probabilities for the n atoms in the molecule and ${}^{k}\mathbf{\Pi}$ were the kth natural powers of the ${}^{1}\mathbf{\Pi}$ matrix. The ${}^{A}p_k(j)$ values were defined in a similar way to the ${}^{k}p_{ij}$ probabilities [equation (1)] but, in this case, in the sum of all the nucleotides in the RNA molecule; see González *et al*. for an exhaustive explanation [26-28]. All of the calculations were performed using our experimental software MARCH-INSIDE [34]. The data set of footprinted and binding nucleotides was extracted from the literature [35]. The secondary structure of the HIV Type-1 Ψ-RNA packaging region as well as the binding sites of Paromomycin is depicted in Figure 2.

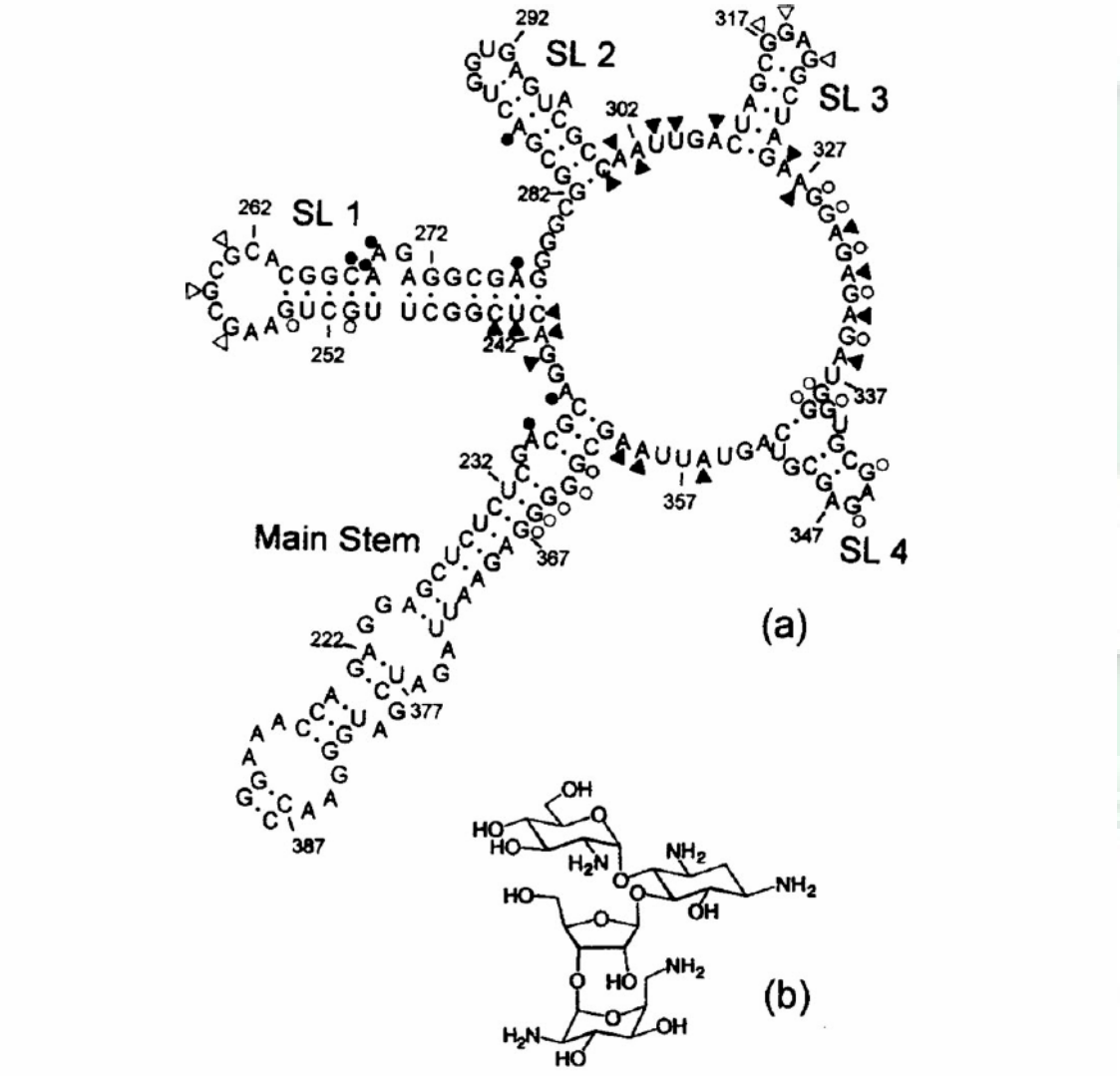


Fig. 2. HIV Type-1 RNA packaging region (**a**), nucleotides involved in binding and enhancement (structural changes) for RNAse I are shown as filled circles and triangles, respectively (open symbols indicates the use of RNAse T1). Structure of paromomycin (**b**).

Results

MVN and the probability of footprinting after RNA–paromomycin interaction

Linear Discriminant Analysis (LDA) [36, 37] was used to classify nucleotides as footprinted or not footprinted after exposure of RNA to the drug. In the development of the LDA the output was a dummy variable: *Binding*, which codify the presence (*Binding*=1) or absence (*Binding* = 0) of structural changes for each specific nucleotide on the presence of RNAses (without specifying the RNAse type). In this problem the input was the Θ_k, with k in the range [0, 10]. The best equation found to discriminate between footprinted and non-footprinted nucleotides was:

$$\text{Binding} = -2.468\ {}^{1}\mathbf{O}(\Theta_{10}) - 0.897\ {}^{3}\mathbf{O}(\Theta_{5}) + 0.658\ {}^{2}\mathbf{O}(\Theta_{9}) - 0.068 \qquad (4)$$

$$N = 151 \quad \lambda = 0.58 \quad F = 35.46 \quad C = 0.64 \quad p < 0.05 \quad \rho = 37.7$$

where λ is Wilk's statistic, N is the number of interactions, F is Fisher's statistics and p is the p-level (probability of error). This linear discriminant function classified correctly 80% (36/45) of the drug-interacting nucleotides and 80.2% (85/106) of the footprinted ones. The model gave rise to a general level of accuracy of 80.13% (121/151) in training series. Matthews' regression coefficient (C) quantified the strength of the linear relation between the molecular descriptors and the drug–nucleotide interaction classifications [38, 39]. The coefficient ρ was used to control the ratio of the adjustable parameters in the model with respect to the number of variables [40].

A high degree of collinearity was detected between Θ_k and so the dependent variables were orthogonalized according to Randic's orthogonalization procedure [41], mean-centered (with the arithmetic mean) and scaled with $1/\text{sd}(\mathbf{x}i)^2$, the inverse of the squared standard deviation [42-44]. Here, we used the symbols ${}^{I}\mathbf{O}(\Theta_k)$, where the superscript I expresses the order of importance of the variable (Θ_k) after a preliminary forward stepwise analysis and **O** means orthogonal.

Jackknife experiments were performed by leaving out all the nucleotides with a common base and deriving the model under the same conditions. All of the statistical parameters of the new models obtained after removing the nucleotides were checked in order to assess the stability of the model. Matthews' coefficients and the overall accuracies for those models were: $C_A = 0.64$ (82.7%), $C_C = 0.68$ (80%), $C_G = 0.55$ (75.8%) and $C_U = 0.61$ (78.8%).

$$\log \text{K}(10^{-4}\,\text{M}^{-1}) = 0.338(\pm 0.068)\text{RNAse} - 0.102(\pm 0.025)\ {}^{1}\mathbf{O}(\Theta_{10}) + 0.083(\pm 0.035)\ {}^{4}\mathbf{O}(\Theta_{8}) + 0.693(\pm 0.038) \qquad (5)$$

$$N = 24 \quad R = 0.91 \quad R^2 = 0.83 \quad F(3,20) = 31.48$$

$$p < 0.00 \quad \text{Std.Err.} = 0.115 \quad Q_2 = 0.825$$

where N is the number of interactions with a known affinity constant *(K)*, log$K$ was the output of the model, F is Fisher's statistics, Std.Err. is the standard error of estimates, R^2 is the squared regression coefficient for training and Q^2 the same for the leave-one-out procedure. One special input for this model was the 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. Additionally, the Θ_k, with k in the range [0, 10], were used as well. The results for the residuals analysis are depicted in Table 2.

Table 2. Observed, predicted and predicted (after cross-validation procedure) values obtained from the quantitative model of log *K*

NUC	Obs[a]	Pred[b]	Predcv[c]	NUC	Obs[a]	Pred[b]	Predcv[c]	NUC	Obs[a]	Pred[b]	Predcv[c]
A235	1.204	1.166	0.359	**G254**	0.447	0.518	0.032	**G340**	0.778	0.672	0.778
A239	1.204	1.166	0.359	**G328**	0.845	0.862	0.43	**G344**	0.845	0.735	0.845
A268	0.903	0.856	0.125	**G329**	0.845	0.862	0.43	**G346**	0.845	0.862	0.845
A269	0.903	0.987	0.125	**G331**	0.845	0.862	0.43	**G363**	0.415	0.399	0.415
A276	1.23	1.071	0.452	**G333**	0.845	0.862	0.845	**G364**	0.415	0.399	0.415
A286	0.778	1.024	−0.067	**G335**	0.845	0.862	0.845	**G365**	0.415	0.399	0.415
C267	0.903	0.856	0.058	**G338**	0.778	0.672	0.778	**G366**	0.415	0.594	0.415
G251	0.447	0.518	0.032	**G339**	0.778	0.545	0.778	**G367**	0.415	0.594	0.415

NUC: Nucleotide. The values are [a]observed, [b]predicted, and [c]predicted by leave-one-out procedure for log K (10^{-4} M^{-1}) (affinity constant of Paromomycin for RNA).

Discussion

In the first instance, the statistical significance of the results must be discussed before arriving at any conclusions concerning the biology involved. As shown in the results section, the *p*-level of Fisher's test for the LDA was <0.05. This means that the hypothesis of groups overlapping with a 5% error can be rejected. Additionally, a high *C* value (0.64) indicates a strong linear relation between molecular descriptors and the classification of nucleotides. This value, as well as the overall accuracy, remained stable in Jackknife procedures, a fact that indicates an acceptable level of predictability [42, 43]. In the LDA, calculating the ρ coefficient (37.7) controls the number of variables in the equation with respect to the number of cases. This coefficient must be higher than 4 for linear models [40].

The MVNs (Θ_k), as defined here, make use of the Shannon concept of entropy as reported by Richard and Kier [45]. However, these parameters were defined over a time-dependent Markov chain and such a definition attempts to model the biological data using a realistic physical model.

It must be pointed out that equation (4) involves shortrange (k =5), middle-range ($5<k<10$) and far-reaching vibrations (k =10 or greater). This situation means that the probability of binding decreases with theMVNinvolved in the early stages of drug–nucleotide interaction (−0.897). Such behaviour may be explained by taking into consideration the fact that the higher the vibrational entropy changes in the nucleotide backbone, which is necessary for the drug–nucleotide interaction, the more marked structural changes in the ±3-vicinity of the nucleotide. Consequently, the probability of the drug binding to the RNA decreases. Conversely, middle range vibrations seem to stabilize the nucleotide–drug complex with the probability of binding increasing by a factor of 0.658 for each kJK^{-1} of entropy [see equation (4)]. In any case, interactions that generated far-reaching waves with vanishing periods higher than 10 also decreased the binding probability. These seem to be very unstable interactions between the drug and the nucleotide. The time scale in the present model (t_k) is the same as that used for the specific experiment described in the literature [35]. Footprinting techniques can now act as time-resolved methods in the millisecond scale [9].

Equation (5) was found to be statistically significant ($p < 0.05$). The high R^2 value meant that the model accounted for 83% of the data variability. The relatively few cases reported for log $K(10^{-4}M^{-1})$ led us to perform a Jackknife procedure to assess the stability and predictability of the model. The squared cross-validation regression coefficient showed that the model is capable of explaining the same variance as in training (83/82.5%). Equation (5) involves the use of either middle-range or long-range vibrational waves. It is remarkable that the parameters that appear in equation (4) are virtually identical to those used in equation (5) and also have proportional contributions.

Conclusions

In general, it was shown here that the generalized vibrational MARCH-INSIDE model can codify useful chemical information that accounts for drug–RNA interactions. In this way, the model could be of great help in studying the mechanism of action of antiviral drugs that are designed to act against HIV using their packaging region as a target. The models described here may be generalized to other RNA–drug systems by changing both the RNA structure and the reduced mass of the nucleotide–drug pairs. The approach described here represents a novel and very promising way to design antiviral drugs [46]. However, a more in-depth discussion is beyond the scope of the present work.

References

1. D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, et al "GenBank" Nucleic Acids Res. 28:15–18 (2000).
2. M.K. Sakharkar, M. Long, W.T. Tin, et al "ExInt: an exon/intron database" Nucleic Acids Res. 28:191–192 (2000).
3. M.K. Sakharkar, P. Kangueane, T.W. Woon, et al "IE-KB: intron exon knowledge base" Bioinformatics 16:1151–1152 (2000).
4. S. Saxonov, I. Daizadeh, A. Fedorov, et al "EID: the exon–intron database—an exhaustible database of proteincoding intron-containing genes" Nucleic Acids Res. 28:185–190 (2000).
5. N.J. Schisler, J.D. Palmer, "The IDB and IEDB: intron sequence and evolution databases" Nucleic Acids Res. 28:181–184 (2000).
6. Z. Yuan "Prediction of protein subcellular location using Markov chain models" FEBS Lett. 451:23–26 (1999).
7. T.D. Tullius "Physical studies of protein–DNA complexes by footprinting" Annu. Rev. Biophys. Biophys. Chem., 18:213–237 (1989).
8. M. Brenowitz, D.F. Senear, M.A. Shea, et al "Quantitative Dnase footprint titration: a method for studying protein–DNA interactions" Methods Enzymol., 130:132–181(1986).
9. A., Henn, J. Halfon, I. Kela, et al "Nucleic acid fragmentation on the millisecond timesacale using a conventional X-ray rotating anode source: application to protein–DNA footprinting" Nucleic Acids Res. 29:122 (2001).
10. D.J. Galas, A. Schmithz "Dnase footprinting: a simple method for the detection of protein–DNA binding specificity" Nucleic Acid Res. 5:3157–3170 (1978).
11. O.N. Ozoline, N. Fujita, A. Ishihama "Mode of DNA–protein interaction between the C-terminal domain of Escherichia coli RNA polymerase α subunit and T7D promoter UP element" Nucleic Acids Res. 29:4909–4919 (2001).
12. O.N. Ozoline, M.A. Tsyaganov "Structure of open promoter complexes with Escherichia coli RNA polymerase as revealed by the Dnase I footprinting technique: compilation analysis" Nucleic Acids Res. 23:2134–2147 (1995).
13. J.M. Sullivan, J. Goodisman, C.J. Dabrowiak "Absorption studies on aminoglycosides binding to the packaging region of the human immunodeficiency virus type-1" Bioorg. Med. Chem. Lett. 12:615–618 (2002).
14. E.F. Gale, E. Gundliff, P.E. Reynolds, et al "The Molecular Basis of Antibiotic Action". Wiley, London (1981).
15. J.M. Coffin "RNA Tumor Virases" Cold Springs Harbor (1984).

16. S.R. Lynch, M.I. Recht, J.D. Puglisi "Biochemical and nuclear magnetic resonance studies of aminoglycoside–RNA complexes" Methods Enzymol. 317:240 (2000).
17. W.D. Wilson, K. Li "Targeting RNA with small molecules" Curr. Med. Chem. 7:73 (2000).
18. D. Frishman, P. Argos "Seventy-five percent accuracy in protein secondary structure prediction" Proteins Struct. Funct. Genet. 27:329 (1997).
19. F. Österberg, M.M. Garrett, M.F. Sanner, et al "Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in autodock" Proteins Struc. Funct. Genet. 46:34 (2002).
20. M. Vorodovsky, E.V. Koonin, K.E. Rudd "New genes in old sequence: a strategy for finding genes in the bacterial genome" Trends Biochem. Sci. 19:309 (1994).
21. M. Vorodovsky, J.D.M. MacIninch, E.V. Koonin, et al "TI detection of new genes in a bacterial genome using Markov models for three gene classes" Nucleic Acids Res. 23:3554 (1995).
22. A. Krogh, M. Brown, I.S. Mian, et al "Hidden Markov models in computational biology: applications to protein modeling" J. Mol. Biol. 235:1501 (1994).
23. K.C. Chou "Prediction and classification of -turn types" Biopolymers 42:837 (1997).
24. T.J. Hubbard, J. Park "Fold recognition and ab initio structure predictions using Hidden Markov models and β-strand pair potential" Proteins Struc. Funct. Genet. 23:398–402 (1995).
25. V. Di Francesco, P.J. Munson, J. Garnier "FORESST: fold recognition from secondary structure predictions of proteins" Bioinformatics 15:131 (1999).
26. H. González-Díaz, R. Ramos de Armas, E. Uriarte "In silico Markovian bioinformatics for predicting 1Ha-NMR chemical shifts in mouse epidermis growth factor (mEGF)" Online J. Bioinf. 1:83 (2002).
27. H. González-Díaz, S.I Hernádez, E. Uriarte, et al "Symmetry considerations in Markovian chemicals 'in silico' design (MARCH-INSIDE) I: central chirality codification, classification of ACE inhibitors and prediction of s-receptor antagonist activities" Comput. Biol. Chem. 27:217-227 (2003).
28. H. González-Díaz, E. Olazábal, N. Castañedo, et al "Markovian chemicals 'in Silico' design (MARCH-INSIDE), a promising approach for computer aided molecular design II: experimental and theoretical assessment of a novel method for virtual screening of fasciolicides" J. Mol. Mod. 8:237 (2002).
29. N.G. Van Kampen "Stochastic Process in Physics and Chemistry" North-Holland, New York (1981).
30. A.T. Bharucha-Reid "Elements of theory of Markov process on the application. In McGraw-Hill Series in Probability and Statistics" McGraw-Hill Book Company, New York, pp. 167–434 (1960).
31. J.A. Freund, T. Poschel (eds) "Stochastic processes in physics, chemistry and biology" In Lecture Notes in Physics. Springer, Berlin (2000).
32. L.D. Landau, E.M. Lifshitz "Mecánica Quántica no-Relativista" In: Curso de Física Teórica. Reverté. Barcelona, v3, pp 1–49 (1963).
33. G.R. Grimmett, D.R. Stirzaker "Probability and Random Processes" Claredon Press, Oxford, pp. 194–264 (1992).
34. I .Hernández, H. González-Díaz MARCH-INSIDE version 1.0, (Markovian Chemicals 'In Silico', Design), Chemicals Bioactives Center, Central University of 'Las Villas', Cuba (2002). (This is a preliminary experimental version, a professional version will be available to the public in the near future.)

35. P.M. McPike, J. Goodisman, C.J. Dabrowiak "Footprinting and circular dichroism studies on paromomycin binding to the packaging region of the human immunodeficiency virus type-1" Bioorg. Med. Chem. 10:3663–3672 (2002).
36. R. Manhnhold, P. Krogsgaard-Larsen, H. Timmerman (eds) "Chemometric methods in molecular design" vol 2, ed. H. Van Waterbeemd, V.C.H. Weinheim (1995).
37. Statsoft Inc. STATISTICA for windows, Version 6.0 (2002).
38. S. Hua, Z. Sun "Support vector machine approach for protein subcellular localization prediction" Bioinformatics 17:721–728 (2001).
39. B.W. Matthews "Comparison of predicted and observed secondary structure of T4 phage lysozyme" Biochim. Biophys. Acta. 405:442–451 (1975).
40. R. Garcia-Domenech, J.V. de Julian-Ortiz "Antimicrobial activity in a heterogeneous group of compounds" J. Chem. Inf. Comput. Sci. 38:445 (1998).
41. M. Randic "Fitting of Nonlinear regression by orthogonalized power series" J. Comput. Chem. 14:363–370 (1993).
42. R.H. Myers "Classical and Modern Regression with Applications" PWS-KENT Publishing Company (1997).
43. P. Geladi, B.R. Kowalski "Partial least squares: a tutorial" Anal. Chem. Acta 185:1–17 (1986).
44. H. Martens T. Næs "Multivariate Calibration" Wiley, New York (1989).
45. A.J. Richard L.B. Kier "Structure activity analysis of hydrazide monoamine oxidase inhibitors using molecular connectivity" J. Pharm. Sci. 69:497–502 (1980).
46. T. Hermann, E. Westhof "Drug design and high-throughput techniques for RNA targets" Combinatorial Chem. and High Throughput Screening 3:219–234 (2000).