Available online at www.sciencedirect.com
SCIENCE DIRECT
Polymer 46 (2005) 6461–6473
polymer
www.elsevier.com/locate/polymer

Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA–QSAR for mycobacterial promoters

Humberto González-Díaz[a,b,*], Alcides Pérez-Bello[b], Eugenio Uriarte[a]

[a]Department of Organic Chemistry, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain
[b]CBQ and Department of Veterinary Medicine, Central University of Las Villas, 54830 Las Villas, Cuba

Received 19 January 2005; received in revised form 5 April 2005; accepted 12 April 2005
Available online 5 July 2005

Abstract

Stochastic molecular descriptors have been applied in QSAR studies on small molecules and polymers (including our series in Polymer) [H. González-Díaz, A.R. Ramos de, R.R. Molina, Bioinformatics 19 (2003) 2079–2087; H. González-Díaz, R.R. Molina, E. Uriarte, Bioorg Med Chem Lett 14 (2004) 4691–4695; H. González-Díaz, R.R. Molina, E. Uriarte, Polymer (I) 45 (2004) 3845–3853; H. González-Díaz, E. Olazábal, N. Castañedo, S.I. Hernádez, A. Morales, H.S. Serrano, et al., J Mol Mod 8 (2002) 237–245; H. González-Díaz, E. Uriarte, A.R. Ramos de, Bioorg Med Chem 13 (2005) 323–331; Polymer (II) (2005) accepted. [40,41,42,44,48]]. However, QSAR studies concerning multiple polymeric RNA molecules, which are among the most important biopolymers, have not been reported to date. The work described here attempts to extend this research by introducing for the first time stochastic moments for the secondary structure of polymeric RNA molecules. These moments are subsequently used to seek a QSAR model that classifies a polymeric DNA sequence as a mycobacterial promoter (mps) or not on the basis of its putative RNA secondary polymeric structure. The model correctly classified 83.7% of 132 mps and 98.89% of 274 control sequences in training. Similar results were obtained in four cross validation experiments using a re-substitution technique that showed the model to have an average 93.9% of robustness and 94.1% of predictability for the 407 sequences used. The present model (mps $= 14.2^1O_0 - 13.4^2O_2 - 1.1$), which has only two variables, compares very favorably in terms of complexity with other models previously reported by Kalate et al.—these authors used a non-linear artificial neural network and a large parameter space [R.N. Kalate, S.S. Tambe, B.D. Kulkarni, Comput Biol Chem 27 (2003) 555–564. [82]]. The model can also be back-projected to derive maps showing the influence of sub-structural RNA patterns on the biological activity of the polymer as a whole.

Keywords: QSAR; Polymer electrostatics; Polymer secondary structure

1. Introduction

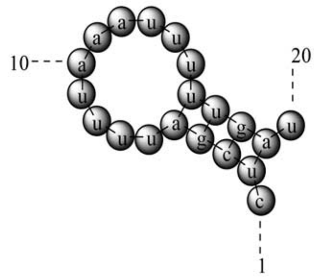
The use of molecular descriptors to derive quantitative structure–activity relationships (QSAR) is an approach of major interest. Molecular descriptors are numerical indices that codify either molecular or polymeric structures [1]. The general use of QSAR is also illustrated in the works of Roy and others [2,3]. In this sense, Gónzalez and Morales applied molecular descriptors in polymer science [4,5]. New sequences of molecular descriptors have been defined for DNA [6] and protein sequence QSAR [7–10]. However, in our opinion, classical QSARs deal with branched rather than linear polymeric molecules such as many synthetic polymers and DNA and protein sequences. For this reason, greater success can be expected for classical molecular indices when branched polymers are considered. Indeed, the branched polymer of greatest biological interest is the RNA secondary structure as described by Mathews, Turner and Zuker [11]. Nevertheless, despite the fact that more than 1600 molecular descriptors have been reported to date,

* Corresponding author. Address: Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, USC, Campus Sur. 15782 Santiago de Compostela, Spain. Tel.: +34 981 563100x14938; fax: +34 981 594 912.
E-mail address: ohumbe@usc.es (H. González-Díaz).

0032-3861/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2005.04.104

Table 1
Calculation of the stochastic molecular electrostatic moments for RNA secondary structure

Fragment of the RNA secondary structure
for the DNA promoter sequence of gene S6
from M. Smegmatis: c1ucgauuuuaaauuuuugau20.*



$$^{1}p_{c_1,u_2} = \frac{\varphi_0(u_2)}{\varphi_0(u_2)+\varphi_0(c_1)} = \frac{\varphi_0(u_2)}{\varphi_{u2}}$$

$$^{1}p_{u_2,c_1} = \frac{\varphi_0(c_1)}{\varphi_0(c_1)+\varphi_0(u_2)+\varphi_0(c_3)+\varphi_0(a_{19})} = \frac{\varphi_0(c_1)}{\varphi_{c1}}$$

$$^{1}p_{u_2,a_{19}} = \frac{\varphi_0(a_{19})}{\varphi_0(a_{19})+\varphi_0(u_2)+\varphi_0(c_1)+\varphi_0(c_3)} = \frac{\varphi_0(a_{19})}{\varphi_{a19}}$$

$$^{1}p_{a_{19},u_2} = \frac{\varphi_0(u_2)}{\varphi_0(u_2)+\varphi_0(a_{19})+\varphi_0(g_{18})+\varphi_0(u_{20})} = \frac{\varphi_0(u_2)}{\varphi_{u2}}$$

$$^{SR}\pi_0 = \mathbf{o}^T \cdot (^{1}\Pi)^0 \cdot \mathbf{o} = \begin{bmatrix} o\ (c_1) & o\ (u_2) & . & . & o\ (u_{20}) \end{bmatrix}^T \cdot \begin{bmatrix} 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & 1 \end{bmatrix} \cdot \begin{bmatrix} o\ (c_1) \\ o\ (u_2) \\ . \\ . \\ o\ (u_{20}) \end{bmatrix} = \sum_{j=1}^{n} {}^{0}p_{ii} = n$$

$$^{SR}\pi_k = \mathbf{o}^T \cdot (^{1}\Pi)^k \cdot \mathbf{o} = \begin{bmatrix} o\ (c_1) & o\ (u_2) & . & . & o\ (u_{20}) \end{bmatrix}^T \cdot \begin{bmatrix} ^{1}p_{C1,T2} & 0 & 0 & . & 0 \\ ^{1}p_{T2,C1} & ^{1}p_{T2,T2} & ^{1}p_{T2,C3} & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & ^{1}p_{T21,T21} \end{bmatrix}^k \cdot \begin{bmatrix} o\ (c_1) \\ o\ (u_2) \\ . \\ . \\ o\ (u_{20}) \end{bmatrix} = \sum_{j=1}^{n} {}^{k}p_{ii}$$

parameters that encode RNA secondary structure in terms of QSAR systems have yet to be described [12]. In particular, a very successful method for the structural characterization of both small molecules and polymers in the more diverse chemical contexts is based on the concept of moments (see, for example, Cabrera-Pérez and numerous others works [13–22]). However, the application of moments or other molecular descriptors to predict the biological activity of polymeric RNA molecules has not been reported.

Markov models are well known tools for the characterization of the structures of biomolecules. Markov models have been used to analyze biological sequence data and have also been used to find new genes from the open reading frames [23,24]. Another uses of these models are data base searching and multiple sequence alignment of protein families and protein domains. Protein turn types and subcellular locations have been successfully predicted [25–28]. Hubbard and Park [29] used amino acid sequence-based hidden Markov Models to predict secondary structures. In this sense, Krogh et al. [30] also proposed a hidden Markov Model architecture. In addition, Markov's stochastic process has been used for protein folding recognition [31]. This approach can also be used for the prediction of protein signal sequences [32,33]. Another seminal work is related to the application of Markov Chain Theory to Proteomic and Bioinformatics. Chou applied Markov Models to predict beta turns and their types, and the prediction of protein cleavage sites by HIV protease [34–37]. However, the combination of Markov models and moments theory for the generation of molecular descriptors that encode biopolymer structures has not been reported in terms of predicting the properties of viruses.

Our group has elsewhere introduced a physically meaningful Markov model that encodes molecular backbone information. This model allowed us to introduce matrix invariants such as stochastic entropies and spectral moments for the study of molecular properties. More specifically, entropy-like molecular descriptors have demonstrated flexibility in a variety of different problems including the estimation of anticoccidial activity [38] and chemically-induced agranulocytocis [39] by small-to-medium sized drug like molecules, modeling the interaction between drugs and HIV-packaging-region RNA [40], and predicting protein and virus activity [41–43]. On the other hand, the stochastic spectral moments introduced by our group have been largely used for small molecule QSAR issues including the design of fluckicidal [44], anticancer [45] and antihypertensive drugs [46]. However, the application of this approach to polymers has been restricted to simple RNA [47] or proteins [48] without consideration of multiple RNA polymeric molecules.

The work described here deals with the definition of a new Markov model, which makes use of novel stochastic moments as molecular descriptors for the RNA polymeric

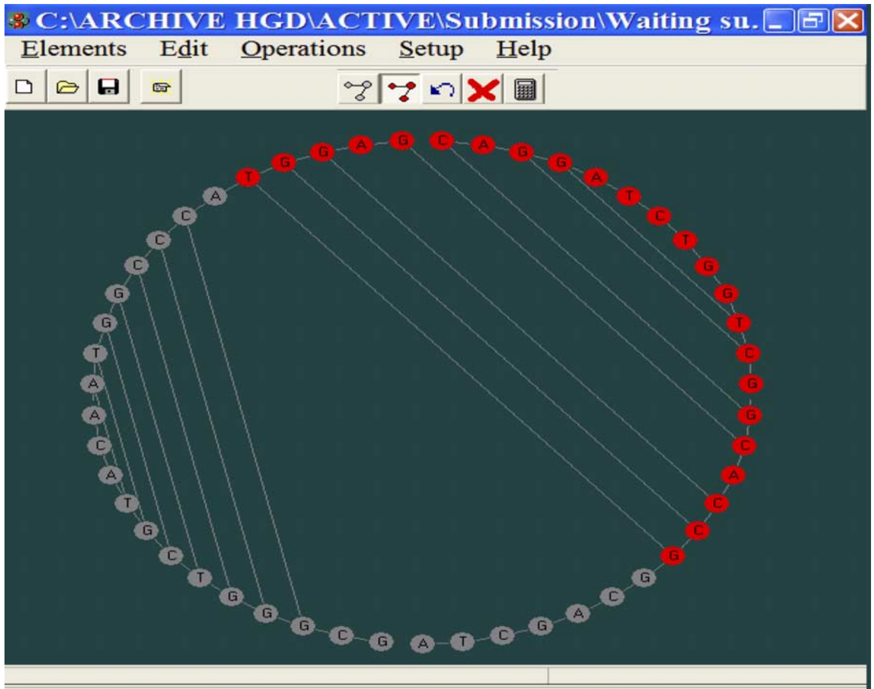


Fig. 1. BIOMARKS 1.0 interface showing the circular representation for RNA of mps T3 from *M. tuberculosis*, note main stem highlighted in red.

secondary structure. In this respect, we will consider as an illustrative example the mycobacterial promoter sequences (mps) problem, which is addressed for the first time from the point of view of RNA stochastic moments.

2. Methods

2.1. The initial probability of RNA secondary structure folding

In analogy to our previous works [38–49], the present approach employed a Markov chain (MC) model to codify information about RNA polymeric secondary structure. In this case the MC model is used to describe an electrostatic-force-driven RNA secondary structure folding process. The procedure considered as states of the MC the nucleotides (nuc) of an RNA sequence [40,47]. This MC is defined by a stochastic matrix, $^1\prod$, built as a squared table of order n, where n is the total number of nucleotides in the RNA. The elements ($^1p_{ij}$) of $^1\prod$ were the probabilities with which a truncated electrostatic interaction [50] of energy E_{ij} occurs between the ith nucleotide (nuc$_i$) and the jth (nuc$_j$) at time $t_1=1$. This time ($t_1=1$) is considered the time when the RNA 2D polymeric structure has just started to fold. In other words, $^1p_{ij}$ was conditioned to an abrupt truncation factor $\delta_{ij}=1$ if nuc$_i$ and nuc$_j$ are either covalently or hydrogen bonded or, alternatively, $\delta_{ij}=0$ if nuc$_i$ and nuc$_j$ are not adjacent within the RNA secondary structure backbone [Eq. (1)] [48,49]:

$$^1p = \frac{\delta_{ij}E_{ij}}{\sum_{k=1}^{\alpha+1}\delta_{ik}E_{ik}} = \frac{\delta_{ij}(Q_iQ_j^*/R_j)}{\sum_{k=1}^{\alpha+1}\delta_{ik}(Q_iQ_j^*/R_k)}$$

$$- = \frac{\delta_{ij}(Q_j^*/R_j)}{\sum_{k=1}^{\alpha+1}\delta_{ik}(Q_j^*/R_k)} = \frac{\delta_{ij}\varphi_j}{\sum_{k=1}^{\alpha+1}\delta_{ik}\varphi_k} \quad (1)$$

Where we summed up to all the α-neighbors of nuc$_i$ and $\varphi_i=Q_j^*/R_j$ was called the electrostatic potential at the nucleotide surface (nuc$_j$-charge-radius ratio), R_j is the radius for the nucleotide and Q_j^* is the nucleotide charge. For the sake of simplicity, the parameter R_j was considered here to be the same for all nucleotides as a rough approximation. The decision concerning which pairs of nucleotides were considered to be adjacent in the RNA secondary structure is an input of this model uploaded from the secondary structure predicted for RNA molecules using algorithms described by Mathews, Turner and Zuker [51]. The calculation of the $^1\prod$ matrix for a given mps fragment is exemplified in Table 1.

2.2. The MC model for RNA electrostatic-driven secondary structure folding

Once the initial electrostatic interactions have taken place in the process of RNA polymer chain folding ($t_1=1$), it is expected that the more favorably the RNA 'manage to' relax to a stable structure the higher will be the stability of this transcript in the cytoplasm of the cell. Thus, at this second stage the problem can deal with the calculation of the probabilities (^kp) with which the electrostatic interactions between nucleotides propagate to the other

Table 2
RNA–QSAR model robustness and predictability in cross-validation

Robustness							
Train	%	mps	Cs	Average	%	mps	Cs
mps	83.7	113	22	mps	83.2	84	17
Cs	98.9	3	268	Cs	99.3	2	202
Total	93.8	116	290	Total	93.9	86	219
cv1	%	mps	Cs	cv2	%	mps	Cs
mps	84.3	86	16	mps	83.2	84	17
Cs	99.0	2	201	Cs	100.0	0	202
Total	94.1			Total	94.4		
cv3	%	mps	Cs	cv4	%	mps	Cs
mps	82.2	83	18	mps	83.2	84	17
Cs	99.5	1	202	Cs	98.5	3	202
Total	93.8			Total	93.5		
Predictability							
Train	%	mps	Cs	Average	%	mps	Cs
mps	83.7	113	22	mps	84.4	29	5
Cs	98.9	3	268	Cs	98.9	1	67
Total	93.8			Total	94.1		
cv1	%	mps	Cs	cv2	%	mps	Cs
mps	87.9	29	4	mps	79.4	27	7
Cs	98.5	1	67	Cs	100.0	0	69
Total	95.0			Total	93.2		
cv3	%	mps	Cs	cv4	%	mps	Cs
mps	85.3	29	5	mps	85.3	29	5
Cs	97.1	2	66	Cs	100.0	0	66
Total	93.1			Total	95.0		

nucleotides nuc_j with time (t_k with $k>1$) to reach a folding equilibrium value.

The most important approximation in the present work considers that once the first electrostatic interaction takes place between two nucleotides within the polymer ribbon, the probabilities (${}^kp$) of propagation for such an interaction to other nucleotides obey Chapman-Kolmogorov equations [45]. In mathematical terms, these probabilities are the elements of the matrices ${}^k\prod$, which can be calculated as depicted in Eq. (2) [38–49]:

$$\prod = \left(\prod\right)^k \tag{2}$$

The elements of the matrices ${}^k\prod$ depend on the adjacency relationships between the nucleotides on the RNA and the charge on these nucleotides. For this reason, any molecular descriptor derived from these matrices necessarily encodes information on the secondary folding and electrostatic characteristics of the RNA polymer chain. We therefore used the moments of the ${}^k\prod$ matrices as electrostatic and secondary folding molecular descriptors for the RNA polymer [44–49]:

$${}^{SR}\pi(\varphi) = \mathrm{Tr}\left(\left(\prod\right)^k\right) = \mathbf{o}^{\mathrm{T}}\left(\prod\right)^k\mathbf{o} = \sum_{\mathrm{nuc}=j}^{n} {}^kp \tag{3}$$

where Tr is the trace operator [12–22,44–49] that indicates the operation of summing up all the probabilities (${}^kp$, self-return probabilities) within the main diagonal of these matrices. Table 1 exemplifies the calculation of ${}^{SR}\pi$. The vector $\mathbf{o}$ and its transpose $\mathbf{o}^{\mathrm{T}}$ represent a Kröcnecker notation vector, which elements $o(j)$ are equal to 1 if multiplied an element in the main diagonal of ${}^1\prod$ and 0 otherwise. This notation will be used in Section 3 for comparative purposes. All calculations of molecular descriptors were performed with our experimental software BIOMARKS 1.0 (BIOinformatics MARKovian Studio) [52]. BIOMARKS uploaded ct files generated by the software RNAstructure in order to input secondary RNA structure connectivity information necessary for the calculation of the different molecular descriptors (${}^{SR}\pi$) [53]. These ct files can also be depicted in the BIOMARKS user interface in, for example, the circular representation illustrated in Fig. 1.

2.3. QSAR and statistical analysis

The classification of polymers according to different structural properties [54] is an intriguing field of research. In this respect, there are many different techniques that are appropriate for pattern recognition and classification problems. However, linear discriminant analysis (LDA) is often preferred by researchers in QSAR, mainly on the basis of its simplicity [55,56]. In the present work we decided to use LDA in order to seek a linear discriminant function to

Table 3
Name, sequence, species, training probability, and cross-validation average probability for all the mps used

Gene name	Sequence	P	Pcv
M. tuberculosis			
T3	atcgacggccacggctggtctaggac-gaggtacccggtaacatgctgggcg	1.00	1.00
T6	ccgtccagtctggcaggccggaaa-catcggtcagcagataggctttaccag	0.98	0.99
T26	ctgcgagcatcatatgccgcgtgcgtggt-gatgcggcaggatgttggacca	1.00	1.00
T180	gatcactccgagcatgcgcccattgttgtg-catagggcaggatgccctgg	1.00	1.00
T101	agcgatcgcagccgacgtgatacct-gaccgttgttgatagtgtcggcggca	1.00	1.00
T119	ccccgtgctcgtagtaggcgtccagcc-gacccgccgctaccatgcacaagt	0.99	1.00
T125	ccgaggtaaggactgagcatgggcccga-taaagtgactattatggatttct	0.82	0.86
T129	actcgcggcagattacgcc-gacggttcctggcgtggttcaa-tattcgccga	0.99	0.95
T130	actccaacaggtcga-taacctcctgcgcctgctcgtctatgctgc-gatccg	0.95	0.96
T150	gacccccgccacgtattgacactttgcga-cacgcttttatcattttccgac	0.43	0.55
RecA	ttcggagcagccgacttgt-cagtggctgtctctagtgtcacggccaacc-gaccgat	1.00	0.86
RrnA P1	gagaacctggtgagtctcggtgccga-gatcgaacgggtatgctgttaggcgacggt-cacct	1.00	1.00
gyrA	gatgggcgaggacgtcgacgcgcggcg-cagctttatcacccgcaacgccaag-gatgttcggt	1.00	1.00
cpn60	ccccggcgatccccgtgctcaccacgggt-gatttccggggcggcatgcgttagcggac-tagc	1.00	1.00
gyrB P1	gatgtccgacgcacggcgcggtta-gatgggtaaaaacgaggccagaa-gatcggccctggcgc	1.00	1.00
gyrB P3	caaggggcctcgccatattgccgg-taggggtccgcgcgacacctacggataa-cacgtcgatc	1.00	1.00
85A	gaagttgtggttgactacacgag-cactgccgggcccagcgcctgcagtct-gacctaattcag	0.99	1.00
85A	cgcccgaagttgtggttgactacacgag-cactgccgggcccagcgcctgcagtct-gacctaattcag	1.00	1.00
gyrB P2	agcggttggcaacgatgtggtgcgatcgc-taaagatcaccgggccggcac-catcgtggcgca	1.00	1.00
rrnA PCL1	tgaccgaacctggtcttgactccattgccg-gatttgtattagactggcagggttgccc-gaaa	0.97	0.98
16S rRNA	tgaccgaacctggtcttgactccattgccg-gatttgtattagactggcagggttgcccc-gaa	0.98	0.97
glnA	tcggcatgccaccggttacgatcttgcc-gaccatggccccacaatagggccgggga-gacccggcgt	1.00	0.99

Table 3 (*continued*)

Gene name	Sequence	P	Pcv
glnA	ccaccggttacgatcttgccgac-catggccccacaatagggccgggga-gacccggcgt	1.00	1.00
KatG P_A	ggtcatctactggggtctatgtcct-gattgttcgatatccgacacttcgcgatca-catccgtgat	0.94	0.96
KatG P_A	atctactggggtctatgtcctgattgttcga-tatccgacacttcgcgatcacatccgtgat	0.82	0.85
katG P_B	gaggcggaggtcatctactggggtc-tatgtcctgattgttcgatatccga-cacttcgcgatcac	0.99	0.95
katG P_B	acgaggcggaggtcatctactggggtc-tatgtcctgattgttcgatatccga-cacttcgcgatcac	0.99	0.99
katG P_C	cctgattgttcgatatccgacacttcgcgat-cacatccgtgatcacagcccgataacac-caactcct	0.35	0.51
katG P_C	ttcgatatccgacacttcgcgatca-catccgtgatcacagcccgataacac-caactcct	0.11	0.17
purL	cggcttgtccgtttccacgcggccg-cagcgcgatggggcctagcta-gactgcctccgtgatgtctcc	1.00	0.78
purC	atctcataccagagataccagca-cagggcgccgtcgtgcggcgga-taggctggcgtgatgcgcccccgc	1.00	1.00
groE	caggaagcaagggggcgcccttgagtgc-tagcactctcatgtatagagtgctagatgg-caatcggcta	1.00	1.00
groE	caggaagcaagggggcgcccttgagtgc-tagcactctcatgtatagagtgctagatgg-caatcggcta	1.00	1.00
ahpC	tgtgatatatcacctttgcctgacagcgactt-cacggtacgatggaatgtcgtaac-caaatgc	0.52	0.64
32 kDa	acatgcatggatgcgttgagatgaggat-gagggaagcaagaatgcagcttgttga-cagggttc	0.99	0.88
10 kDa	aagcaaggggcgcccttgagtgtcag-cactctcatgtatagagtgctagatgg-caatcggctaa	0.97	0.97
10 kDa	aagcaaggggcgcccttgagtgtcag-cactctcatgtatagagtgctagatgg-caatcggct	0.99	0.98
10 kDa	aagcaaggggcgcccttgagtgtcag-cactctcatgtatagagtgctagatggca	0.94	0.95
65 kDa	gcgtaagtagcggggttgccgt-cacccggtgacccccgtttcatcccc-gatccggaggaatcac	1.00	0.98
mpt64	gagtctggtcaggcatcgtcgtcag-cagcgcgatgccctatgtttgtcgtcgact-cagatatcg	1.00	1.00
metA	tccggcccccgcgatttggc-gagcttcgtgcgtgttcggtagcctggcatt-taccgacgcggggt	1.00	1.00
rpsL	gccgcaacgcccgctttgacctgcca-gactggcggcggg-tattgtggttgctcgtgcctggcggc	1.00	1.00
38 kDa	cgtcgccggactgtcgggggacgtcaag-gacgccaagcgcggaaattgaagagca-cagaaaggtatg	1.00	1.00

Table 3 (*continued*)

Gene name	Sequence	P	Pcv
ppgk	cgggccgcagtttaaggtgagggtcatc-cacgtctcgccgaggagattcgatgac-cagcac	1.00	1.00
M. bovis BCG			
hsp60 P2	cggtgcggggcttcttgcactcggca-taggcgagtgctaagaataacgttg	1.00	1.00
rRNA	tgaccgaacctggtcttgactccattgccg-gatttgtattagactggcagggttgcccc-gaa	0.98	0.99
ahpC	tgtgatatatcacctttgcctgacagcgactt-cacggtacgatggaatgtcgcaac-caaatgc	0.56	0.66
23 K	gagtctggtcaggcatcgtcgtcag-cagcgcgatgccctatgtttgtcgtcgact-cagatatcg	1.00	0.89
mpb64	gagtctggtcaggcatcgtcgtcag-cagcgcgatgccctatgtttgtcgtcgact-cagatatcg	1.00	1.00
18K	tggcgtccgaaacacttgaggtgcggcc-cagcaaggggctacaggtttttttccttcacc-tacgga	1.00	1.00
64K	gcgtaagtagcggggttgccgt-cacccggtgacccccggtttcatcccc-gatccggaggaatcac	1.00	1.00
rpsL	gccgcaacgcccgctttgacctgcca-gactggcggcggg-tattgtggttgctcgtgcctggcggc	1.00	1.00
Mpb70	tggcgtccgaaacacttgaggtgcggcc-cagcaaggggctacaggtttttttccttcacc-tacgga	1.00	1.00
alpha	cgactttcgcccgaatcgacatttggcctc-cacacacggtatgttctggcccgagca-cacgacga	1.00	1.00
M. leprae			
16S rRNA	tagtcaacccgggacttgactcctctgctg-gatctgtattaatctggctgggttgccgaag	0.99	0.99
18Kda	cttgtctatcacaacttgcatcaatatatc-gaccagtgctatatcaaatctatgtagt-cagga	0.01	0.25
18 Kda	cttgtctatcacaacttgcatcaatatatc-gaccagtgctatatcaaatctatgtagt-cagga	0.01	0.01
28-kDa	tcaatataaccactctggtcacactaacca-tactcgtaccatcaaccgtgtggggc-taatcc	0.07	0.06
groE1	agcagcgggccggccttgagtgctag-cactcgcgtgtatagagtgctagatgg-cagtcggccag	1.00	0.77
65 kd	gaattccggaattgcactcgccttaggg-gagtgctaaaaatgatcctggcactcgc-gatc	0.92	0.94
36k	gttgggtttcctctcggagggcgcaccgc-tacgttagcgggatg	1.00	0.98
SOD	ggtgggcgcgatcatggcgcagcgttgat-tatgctagtcg	1.00	1.00
rpsL	cgccgttgggtcgctttgacctgcccgag-cagggacgggtattgtgtttctcgttcct-gacggct	1.00	1.00
M. smegmatis			
alrA	gtctgcggcctctgggacaatgggcgccg-gagattatga	1.00	1.00

Table 3 (*continued*)

Gene name	Sequence	P	Pcv
S4	aagccgaatcgagaccttttgggttcgta-cacacttgctttataagcctcg	0.42	0.56
S5	aacaagattccgttaatcgtgtctggtg-gagctggtggtaagcttgatccg	0.97	0.84
S6	catcgattttaaatttttgatagagtgcaaa-taa	0.00	0.24
S12	acctcgttatgcttctggctatttttgat-caacttttatacatgggcggtt	0.32	0.23
S14	tcaagcacccaagccaacatggttgtag-tagtcgttttaccatgtgtacct	0.19	0.22
S16	tccacgc-gaaccgcttcggcgtgccccgttttccctgt-tataatatcggcg	0.98	0.79
S18	gatcattgtcttctgttgtctttcgta-taaagttgttactg	0.10	0.32
S19	tttgatgtagccaaaggctctcaccacct-gagccatgatagtatccatccc	0.17	0.15
S21	acatggcatttttcatttaaaacaggact-caggtggtatggttgacatcga	0.99	0.79
S30	gatcagctatgttcttcagtaaaatttcggc-tatatgttggtg	0.11	0.33
S33	gatccgctcttcttatgatgccagttatgg-tatctatggttatcg	0.49	0.39
S35	aactaaagtatgtgccgtaattga-cagtgttctagattatgatgctgcatc	0.04	0.16
S65	ggcacagctcgaagttctacta-catggcttgctgaatccagtcacattact	0.22	0.17
S69	atcacgatgtcttcatgcttggctt-caatgctccggtctacaatcagttca	0.24	0.23
S119	gatcaagaagccaatgatttgttaaacg-caattaatg	0.00	0.07
gyrB	cagaatcggtgctgtcgctatctcgcggta-gactggacgacggatctcaggc	1.00	0.75
recA	agagttcgaccg-gacttgtcggtggtctgctctaacgt-cacggccaaccgatcgga	1.00	1.00
ask	gtttgcccgccgcggcgccccacgat-gaaccgcacgggctgacg	1.00	1.00
acetami-dase	ggccggcgttcacccttgacttttatttt-catctggatatatttcgggtgaatggaaagg	0.91	0.94
rrnB	ctctgacctggggatttgactcccagtttc-caaggacgtaacttattccaggtcagagc-gac	0.91	0.91
rrnA P1	gaaaacctggtcagcctcggagccga-gatcgagagagtaagctcgtaggaagcaa-gacc	0.97	0.96
rrnA P2	ctctgaccaggcgatttgcaatcgcgac-gaacctcgtattatctttat-gaagtcgccgcgga	0.93	0.94
rrnA P3	ccgggccagagcgacttgacaagc-cagccgagatcgtactaagctggc-gaggttgcctcaga	1.00	0.98
rrnA PCL1	ccggtccagagcgacttgacaagccaga-caaagcagtattaagctggcagggttgccc-caaa	0.83	0.87
rpsL	ccgccgtgcac-gagtttgtttcgtcgcggtcgcccctgg-tattgtggtggatcgtgcctggccc	1.00	0.96
rpsL	cgtgcac-gagtttgtttcgtcgcggtcgcccctgg-tattgtggtggatcgtgcctggcccgaaa	1.00	1.00

Table 3 (*continued*)

Gene name	Sequence	P	Pcv
ahpC	tgtgatatatcacctttgcctgacagcgactt-cacggcacgatggaatgtcgcaac-caaatgc	0.62	0.71
M. paratuberculosis			
pAJB303	gacgacgagggcggtggcgtcgccggtg-tagccgaacggcacgtgcgcgtaggcc-cagat	1.00	0.91
pAJB86	ccaccttactcccgatgacgttg-cacggctgggattaacggtccgcgtgctc-caggagaca	1.00	1.00
pAJB125	gcaacgagcgcatcattaaagatc-gaggcgccgggtcatgtccctt-caccccgcccagctt	0.99	0.99
pAJB300	tcgagttcaagaccctgacgctggcc-gacctcggcgcgcagccgaccgcg-cagcggtgcacg	1.00	1.00
pJB305	atccggacgggcagttgttg-gagtttctgtcggacggttggttggcgg-catttccggcgagg	1.00	1.00
pAJB304	caccaggtacacgccaagga-caacggccgtatccggtac-caacgggtgtgcgagctggacgg	1.00	1.00
P AN	ctggtgaagggtgaatcgacaggtacaca-cagccgccatacacttcgcttcatgccct-tacg	0.92	0.94
pAJB73	gatcggtgtgccgcttgaaccggcc-cagctcccgctccagggtgacgtgctc-gagctc	1.00	0.98
pAJB301	gatctggcgggcggtccagtacaccgc-gagttcgcgcacgctggccgg-cagcgtcttggacgcccg	1.00	1.00
M. fortuitum			
repA	gagctcgtgtcggaccatacaccggtgat-taatcgtggtctactaccaagc	0.84	0.88
rrnA PCL1	ccaggatgatgcaacttgacttgccggcaa-gattcgaattaagctggcggggttgccc-caaa	0.97	0.94
rrnA P1	gaaaacctgttgagcctcggagccga-gatcgaaagagtagggtcgtaaacag-cagtccgggcc	0.99	0.99
rrnA P2a	cgctgaccagccgatttgaccttgtagg-caggcccgcgctaatctttt-gaagtcgcgcggagcgg	1.00	1.00
rrnA P2b	ccgggccagagcgacttgacaagc-cagccgagatcgtactaagctggc-gaggttgcctcagaccg	1.00	1.00
rrnA P3	caggatgatgcaacttgacttgccggcaa-gattcgaattaagctggcggggttgccc-caaaacag	0.96	0.97
rrnA PCL1	actggggacgaggtcttgacgcccctgat-cagatcggtatagactggcagggttgccc-gaaa	1.00	0.99
rrnA P1	gagaacctccgcagtctcggcgccga-gatcgagagggtcgcctgaaa-catgccgtttacctgc	1.00	1.00
rrnA P2	aggggacccccctttttgactccgctca-gacgtgggctattcttctaaccacaagcc-caacgc	0.93	0.95
rrnA P3	ctggggacgaggtcttgacgcccctgat-cagatcggtatagactggcagggttgccc-gaaagcaa	1.00	0.98
M. phlei			
pKGR25	cctgtacaccctcgctgcactcgccgag-gacaagcactatcgcccc-gacgtcccggcctgg	1.00	1.00
pKGR9	accacgagcacccggtcgtcaggactgc-gacactcgatgttgtagacgcactggtg-cagcatg	1.00	1.00
pKGR38	atctggtcgacctgctcgacgaggtcgat-catcttcttcatctcgccgaacgg-gatgccctgg	1.00	1.00
ORF1	acctcatggagcacttcgaggtcactgag-cacgcccacgaactacgagaggccgtgg-gactgg	1.00	1.00
ORF2	tactttttgtaccgttcgacac-cagcggtttccgcttccttgccaatctcctg-caaacaaaccacaatg	0.34	0.51
Mycobacteriophage I3			
rrnA P4	gccaaaaccgggaatttgactcaggttcac-gaacttgatacggtttccgagcgccc-gaaag	0.88	0.90
rrnA P1	ggcgggtctagtggcggacggcgtcaca-gaggtatacgatgtgtttcatatc-gaccgcggttac	1.00	0.97
rrnA PCL1	gcccccgacccgaagttgactcaagtt-cattggacttggta-cagtggtcgggttgccctgaa	0.99	1.00
rrnA P2	gccaaaaccgggaatttgactcaagtt-caccgaacttgatacggtttc-caagtcgctcgg	0.80	0.85
rrnA P3	gccaaaaccgggaatttgactcaagtt-caccgaacttgatacggtttc-caagtcgctcggaa	0.57	0.63
pKGR1	acacagaccaggagctcgacatgaccgc-caccgccccctacagcgtcatctggttc-gaaggcaccccggat	0.99	0.83
Mycobacteriophage L5			
71 P2	tacctgtcacaaggtttgctacc-gagtggggcaggccgctacatttac-gaccgcgtaacgcca	0.99	0.99
71 P left	tttgcgattaggcttgacagccacccggc-cagtagtgcattcttgtgtcaccgcagcagc	1.00	0.99
71 P1	acaactgaatatggttccgcagacgcaac-taaattaggggtatccttgacaggcaccaa-cat	0.11	0.33
M. avium			
Avi-3	gccggcgatcgtgggctgataagtct-tatcgggcatactataagtgtagtgggaaa-tatcacct	0.96	0.75
pLR7	agccttgttggcggccaactgccggac-gatcgcggcggccatcgtcctc-gagctcggccccgtgc	1.00	0.99
M. neoaurum			
rrnA PCL1	gcgagacagagaagcttgactcgccaga-caagatagtttaagctggcagggttgcccc-gaa	0.97	0.98
rrnA P1	gaaaacctggtcagcttgggcgccgg-gatcgagcgagtacactcgtaaga-gaccggtcgagtg	1.00	0.99
rrnA P3	gcgagacagagaagcttgactcgccaga-caagatagtttaagctggcagggttgcccc-gaaacg	0.98	0.98

Table 3 (*continued*)

Gene name	Sequence	P	Pcv
rrnA P2	ctctgaccagcggatttgacttccgaagg-cacaaagttctaatctttt-gaagtcgccgcggggag	0.97	0.97
M. abscessus			
rrnA P4	ggcgggtctagtggcggacggcgtcaca-gaggtatacgatgtgtttcatatc-gaccgcggttac	0.88	0.90
rrnA P1	gcccccgacccgaagttgactcaagtt-cattggacttggta-cagtggtcgggttgccctgaa	1.00	0.97
rrnA PCL1	gccaaaaccgggaatttgactcaagtt-caccgaacttgatacggtttc-caagtcgctcgg	0.99	1.00
rrnA P2	gccaaaaccgggaatttgactcaagtt-caccgaacttgatacggtttc-caagtcgctcggaa	0.80	0.85
rrnA P3	ccaaaacccggagtttgactcaagttcacc-gaacttgatcggttcccgggccgcttacaa	0.57	0.63
M. chelonae			
rrnA P2	ccaaaacccggagtttgactcaagttcacc-gaacttgatcggttcccgggccgcttacaa	0.98	0.88
rrnA P1	ggcggggttagtggcggatggcgtcacc-gaggtatacgatgtgtttcatatc-gaccgcggtta	1.00	0.99
rrnA PCL1	ccccagaacccgaagttgactcaagtt-cattggacttggta-cagtggtcgggttgccctgaa	0.97	0.98
rrnA P3	gccaaaaccgggaatttgactcaagtt-caccgaacttgatcggtttcccagccgccc-gaaa	0.50	0.62
rrnA P4	gccaaaaccgggaatttgactcaagtt-caccgaacttgatacggtttcc-gagccgcccgaaa	0.54	0.53

decide whether a sequence is an mps or not. In this system mps (mycobacterial promoter sequence) is the output of the model and was represented by a dummy variable such that mps = 1 if the sequence is an mps and −1 if belongs to the control sequence (cs) group, which was generated at random. All of the mps cases were taken from the data set of polymer sequences collected by Kalate et al. [57]. The random generation of the control group is a widely accepted method due to the extremely low probability of creating at random a positive sequence [58]. The training quality of this model was assessed by direct inspection of different statistics such as percentages of good classification (% mps, % cs, % total), Wilks' statistics (U), Fisher ratio (F) and the probability of error [p-level (p)]. The parameters % mps and % cs were percentages of good classification of mps and cs. The total percentage of good classification is denoted % total. The quality of the model was considered acceptable if all of these percentages were >85%. Statistical signification was measured by selecting models for which the values of U and F imply that $p<0.05$ [59]. On the other hand, validation of the model was carried out by means of re-substitution cross validation and all statistical calculations were carried out with STATISTICA 6.0 [60]. The cross-validation was carried out four times with four different partitions (training and predicting sets), with 25% of all sequences being leave-out in each of these studies in such a way that each virus was out of the training set on at least one occasion [41,48].

3. Results and discussion

3.1. Mycobacterial promoter polymer sequence recognition by linear discriminant analysis

While *Mycobacteria* have a low transcription rate and a low RNA content per unit DNA, their genomes are rich in G+C monomer content [61]. Since the G+C content of a genome affects the codon usage and the promoter recognition sites in an organism, it is expected that the transcription and translation signals in *Mycobacteria* may be different from those in other bacteria such as *E. coli*. An understanding the factors responsible for the low level of transcription and the possible mechanisms of regulation of gene expression in *Mycobacteria* therefore necessitate examination of the polymer structure of mycobacterial promoters and their transcription machinery, including information about the involved RNA polymer molecules [62,63].

For the reasons outlined above, it is desirable to develop a QSAR technique to correlate RNA secondary structure information with the biological properties of sequences and the study described here aims to address this problem [64]. Fortunately, the RNA secondary structure pictures consist of two elements—letters (nucleotides) and edges (covalent and hydrogen bonds). This means that they can be split into numerous pieces (nucleotides), which are interconnected and, as indicated above, have only four possible colors: A, T, G, and C. This aspect can be automatically identified with the concept of colored graphs, which are commonly dealt with in graph theory [65,66]. It is worth referring to our previous publication in this series [H. González-Díaz, E. Uriarte, Biopolymer 2005 accepted] for an overview of these concepts [67]. From this point it is feasible to encode the information about RNA secondary structure by means of graph theoretical invariants or the same molecular descriptors.

However, this issue involves more than mathematics: The colors of this novel RNA graph have a clear physicochemical meaning. Consequently, an arbitrary graph theoretic invariant should not be selected, but one that is in agreement with the physical sense. A panoply of molecular descriptors has been used previously in QSAR studies over a long period of time. Almost all molecular descriptors are susceptible to a vector–matrixvector representation, including quadratic and linear forms. For instance, the first molecular descriptor defined in a chemical context—the Wiener index W (Eq. (4))—is a quadratic form [68]. In addition, several other classic Zagreb indices M_1 (Eq. (5)) and M_2 (Eq. (6)), Harary number H (Eq. (7)), Randic invariant χ (Eq. (8)), valence connectivity index χ^v (Eq. (9)), the Balaban index J (Eq. (10)), the MTI index (Eq. (11)), and Moreau-Boroto autocorrelation ATS_d (Eq. (12))—to mention a few examples—can all be expressed in quadratic, linear or in general vector–matrix–vector forms [12,69]. Unfortunately, many of these do not have a direct physical interpretation and have not been used for QSAR with a number of RNA molecules. This lack of physical sense can be detected in recent quadratic $q_k(X)$ (Eq. (13)), linear $f_k(X)$ (Eq. (15)), and stochastic forms $s_k(X)$ (Eq. (16)) introduced by Marrero-Ponce et al. [70–73]:

$$W = \frac{1}{2}(\mathbf{u}\cdot\mathbf{D}\cdot\mathbf{u}^T) \quad (4)$$

$$M_1 = \mathbf{v}\cdot\mathbf{A}\cdot\mathbf{u}^T \quad (5)$$

$$M_2 = \frac{1}{2}(\mathbf{v}\cdot\mathbf{A}\cdot\mathbf{v}^T) \quad (6)$$

$$H = \frac{1}{2}(\mathbf{u}\cdot\mathbf{D}^{-k}\cdot\mathbf{u}^T) \quad (7)$$

$$\chi = \mathbf{v}'\cdot\mathbf{A}\cdot\mathbf{v}'^T \quad (8)$$

$$\chi^v = \mathbf{v}''\cdot\mathbf{A}\cdot\mathbf{v}''^T \quad (9)$$

$$J = \frac{1}{2}\cdot C\cdot(\mathbf{d}'\cdot\mathbf{A}\cdot\mathbf{d}'^T) \quad (10)$$

$$\mathrm{MTI} = \mathbf{v}\cdot(\mathbf{A}+\mathbf{D})\cdot\mathbf{u}^T \quad (11)$$

$$\mathrm{ATS} = \mathbf{w}\cdot{}^{\mathbf{m}}\mathbf{B}\cdot\mathbf{w}^T \quad (12)$$

$$q_k(X) = \mathbf{w}\cdot\mathbf{M}\cdot\mathbf{w}^T \quad (13)$$

$$f_k(X) = \mathbf{w}\cdot\mathbf{M}\cdot\mathbf{u}^T \quad (14)$$

$$s_k(X) = \mathbf{w}\cdot\mathbf{S}_k\cdot\mathbf{w}^T \quad (15)$$

where, all the matrices and vectors used have been explained in detail in the literature and, for the sake of brevity, will not be explained here. In particular, the vector o (mentioned above in Section 2) may be used here to represent spectral moment descriptors as quadratic forms. These indices include the self-return walking counts srwc^k (Eq. (16)), the spectral bond moments $\mu(B)$ (Eq. (17)) and bond-weighted adjacency matrices $\mu(^dB)$ (Eq. (18)), the energy moments $\mu(H)$ (Eq. (19)), the Kirchhoff number Kf (Eq. (20)), the I_3 number (Eq. (22)), [12] and our stochastic moments $^{SR}\pi$ (Eq. (21)):

$$\mathrm{srwc}_k(\mathbf{A}) = \mathbf{o}\cdot\mathbf{A}\cdot\mathbf{o}^T \quad (16)$$

$$\mu_k(\mathbf{B}) = \mathbf{o}\cdot\mathbf{B}\cdot\mathbf{o}^T \quad (17)$$

$$\mu_k(^d\mathbf{B}) = \mathbf{o}\cdot[(\mathbf{B}+\mathbf{W})]^k\cdot\mathbf{o}^T \quad (18)$$

$$\mu_k(\mathbf{H}) = \mathbf{o}\cdot\mathbf{H}\cdot\mathbf{o}^T \quad (19)$$

$$Kf = \mathbf{o}\cdot(a\cdot\mathbf{L})\cdot\mathbf{o}^T \quad (20)$$

$$^{SR}\pi = \mathbf{o}\cdot\left[\left(\prod\right)^k\right]\cdot\mathbf{o}^T \quad (21)$$

$$I_3 = \frac{1}{k!}\sum_k^{\infty}\mu_k(\mathbf{A}(\varphi,\Psi,\omega)) = \frac{1}{k!}\sum_k^{\infty}[\mathbf{o}\cdot\mathbf{A}(\varphi,\Psi,\omega)\cdot\mathbf{o}^T] \quad (22)$$

where, all the matrices and vectors used have again been explained in detail in the literature. Unfortunately, as is the case for the other descriptors (Eqs. (4)–(15)), these spectral moments have not yet been used for QSAR with a number of RNA molecules. However, the physical sense of spectral moments is often more clearly stated [17–21] than for other indices such as the Harary, Balaban, Wiener and Marrero-Ponce indices, with some spectral moments widely used in physics. It can be seen more clearly after expansion of Eq. (22) that the stochastic spectral moments are also vector–matrix–vector forms:

$$^{SR}\pi = \mathbf{o}\cdot\left[\left(\prod\right)^k\right]\cdot\mathbf{o}^T = \left[{}^1o\,{}^2o\cdots{}^no\right] \times \begin{bmatrix} {}^1p\,{}^1p\cdots{}^1p_1p\cdots\cdots & & p\cdots{}^1p \\ & \cdots\cdots & \\ & & \cdots\cdots_1 \end{bmatrix}\cdot\begin{bmatrix} {}^1o & & o \\ & {}_2o & \\ & & \ddots_n \end{bmatrix} \quad (23)$$

However, the stochastic spectral moments introduced here (Table 1) are proportional to the probability with which the effect of the electrostatic field of one nucleotide propagates throughout the RNA backbone at distance k and returns to the initial nucleotide. Consequently, these moments encode information concerning the RNA biopolymer electrostatics and secondary structure folding (see Table 1 for details).

For this reason, we used these descriptors to find the mps QSAR. The data were processed by LDA [56,57] and the best relationship we found for predicting mps had only two variables:

$$\mathrm{mps} = 14.2O_0 - 13.4O_2 - 1.1 \quad (24)$$

$$N_{\mathrm{mps}} = 132 \quad N_{\mathrm{es}} = 274 \quad N_{\mathrm{TOTAL}} = 406$$

$$\lambda_0 = 0.93 \quad \lambda_2 = 0.41 \quad \lambda_{\mathrm{TOTAL}} = 0.44$$

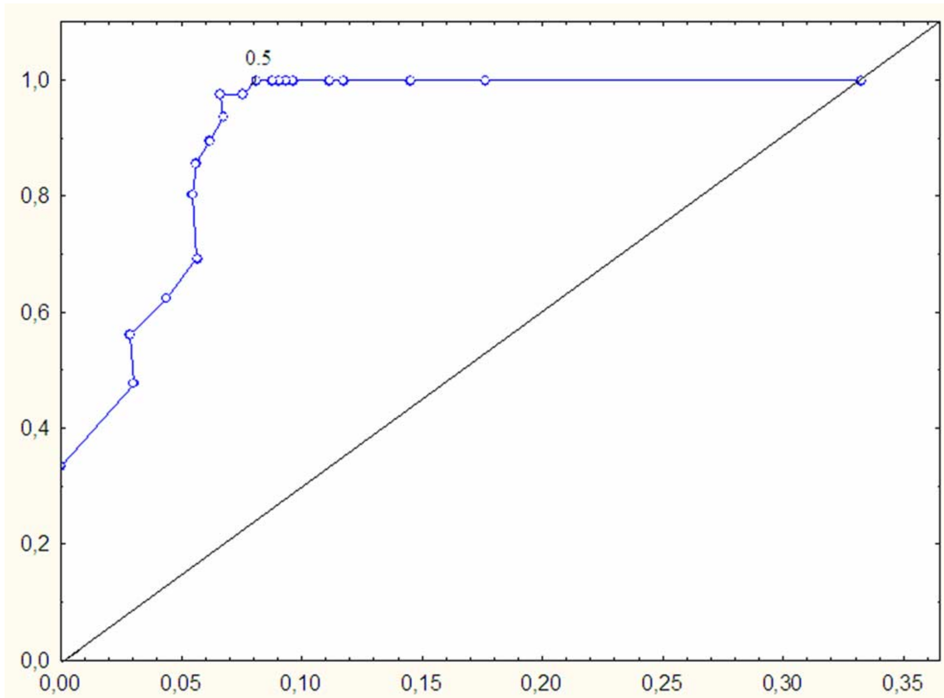


Fig. 2. ROC curve for the present RNA–QSAR model.

$F_0 = 29.0 \quad F_0 = 578.75 \quad F_{\text{TOTAL}} = 515.03$

$\%_{\text{mps}} = 83.7 \quad \%_{\text{es}} = 98.89 \quad \%_{\text{TOTAL}} = 93.83$

The *p*-level of Fisher's test for this LDA was <0.05. This means that the hypothesis of groups overlapping with a 5% error can be rejected. The equation was derived after application of Randič's orthogonalization procedure [74–77]. The symbol ^{m}O represents the orthogonal analog of $^{SR}\pi$ where m is the step at which this variable is selected in the forward stepwise analysis. Details of the overall and group-specific classifications for these series' are given in Table 2. Note that all values remain quite stable under data variation in training and predicting series [78].

This QSAR model gave an overall accuracy of 93.9% in training and 94.1% in four different and predicting series'. It is noteworthy that these values are very high for this kind of analysis [78]. The name, sequence, species, and resulting probabilities in training and cross-validation for all the studied mps are given in Table 3. Direct inspection of Table 2 shows a model accuracy by species of 95.7% for 46 mps from *M. tuberculosis*, 100% for 10 mps from *M. bovis BCG*, 66.7% for 9 mps from *M. leprae*, 57.1% for 28 mps from *M. smegmatis*, 100.0% for 9 mps from *M. paratuberculosis*, 100% for 10 mps from *M. fortuitum*, 80.0% for 5 mps from *M. phlei*, 85.7% for 7 mps from *Mycobacteriophage I3*, 66.7% for 3 mps from *Mycobacteriophage L5*, 100% for 2 mps from *M. avium*, 100% for 4 mps from *M. neoaurum*, 100% for 5 mps from *M. abscessus*, and 100% for 5 mps from *M. chelonae*.

Interestingly, *M. tuberculosis*, which is the most widely represented species (46 sequences), was predicted with a very high accuracy along with the least represented species *M. avium* (only two sequences). The worst predicted species was *M. smegmatis* (57.1%) with 28 mps. We can therefore expect that there were no outlier species with respect to the number of mps of this species used for the analysis. In addition, if the samples of the sequences were distributed in a completely random fashion between the mps and cs sets, the rate of correct identification by random assignment would generally be $1/50 \times 100 = 50\%$. Alternatively, if the distribution is weighted according to the sizes of subsets, one would expect.

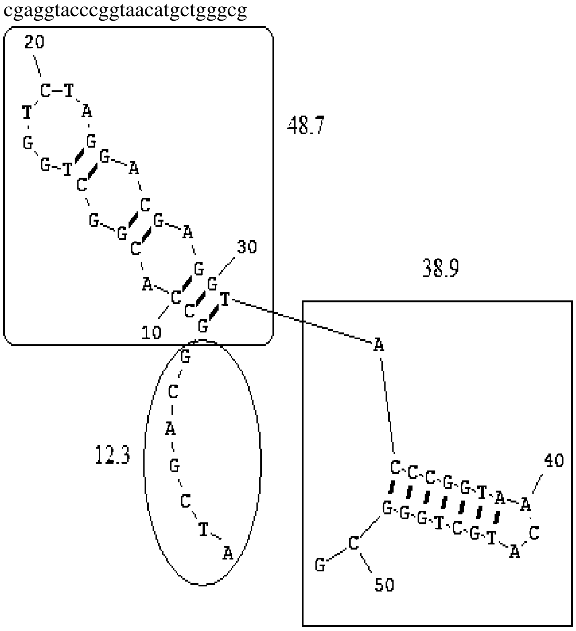
$$(N_{\text{mps}}/N_{\text{TOTAL}})^2 + (N_{\text{cs}}/N_{\text{TOTAL}})^2$$
$$= (132/406)^2 \times 100 + (274/406)^2 \times 100 = 10.6 + 45.5$$
$$= 56.1$$

Therefore, the rates of correct identification obtained in re-substitution cross validation are much higher than the corresponding completely or weighted randomized rates, which implies that the mps is well correlated with the stochastic moments used here [79].

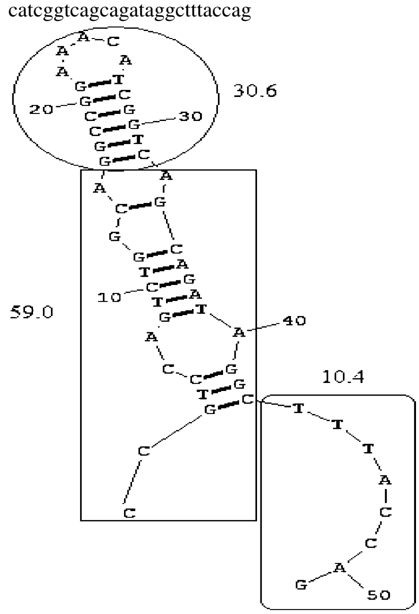
In an effort to avoid over-fitting problems we also built into the above analysis an ROC curve, see Fig. 2. It can be seen by visual inspection that the ROC curve for the present QSAR has an area under the curve that is markedly higher than the area under the random classifier ROC curve (diagonal). The results obtained in this study are therefore highly significant in statistical terms [80]. The previous analysis also takes account of the over-fitting problems in

Table 4
Different possible partitions for RNA back-projection maps of some *M. tuberculosis* mps

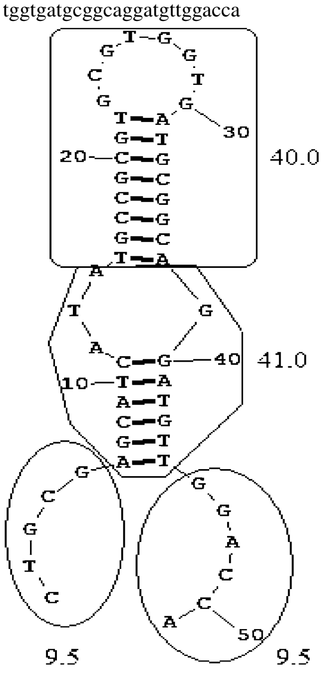
T3
atcgacggccacggctggtctagga
cgaggtacccggtaacatgctgggcg



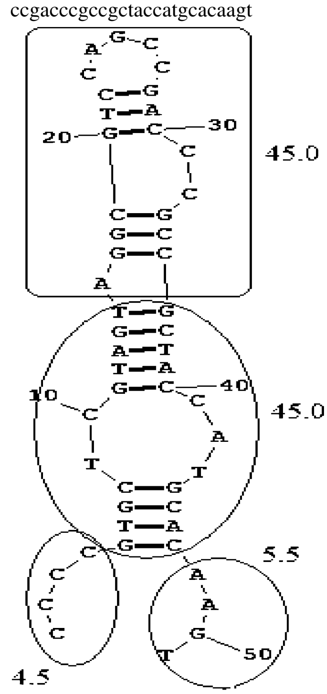
T6
ccgtccagtctggcaggccggaaa
catcggtcagcagataggctttaccag



T26
ctgcgagcatcatatgccgcgtgcg
tggtgatgcggcaggatgttggacca



T119
ccccgtgctcgtagtaggcgtccag
ccgacccgccgctaccatgcacaagt



the present QSAR [81]. Finally, as well as the accuracy of the present model, it should be noted that the model is extremely simple (mps $= 14.2^1O_0 - 13.4^2O_2 - 1.1$) and has only two variables. This compares very favorably in terms of complexity with models previously reported by Kalate et al., who used a non-linear artificial neural network and a large parameter space for the mps data collected by them [82].

3.2. Backprojection analysis for the RNA mps QSAR model

Finally, a back-projection approach was applied in an effort to gain a further insight into the role played by the different RNA motifs in mps action. The use of back-projectable approaches enables the variables in the QSAR to be projected back into molecular space and this provides biologically and chemically significant conclusions [83]. Kalate et al. studied the mps DNA sequences using a caliper randomization approach, which in our opinion can also be classified as a back-projection technique. These authors concluded that that: (i) the −35 box and its upstream region play a critical role in mycobacterial promoter function, (ii) −10 box and spacer region also contribute towards mycobacterial promoter characteristics, and (iii) for promoter recognition the −10 region is not as important as the −35 region [82].

However, results have not been reported to date concerning the secondary structure folding requirements for the putative RNA sequence of an mps in terms of the possible electrostatic interactions. In this study the back-projection is presented as a map in which the secondary RNA molecule for an mps is partitioned into different motifs. The spectral moments for these motifs are then calculated and substituted into the QSAR model to obtain the contribution of each substructure to the biological activity. The contribution values were scaled in the range 0 to 100%. The results of this analysis are represented in Table 4 for some mps from *M. tuberculosis* as an example. In these examples it can be predicted that the hairpin stems and central loops in the putative RNA structure could positively contribute to the mps activity. However, a more extensive study of all mps, which is beyond of the scope of the present paper, will be carried out.

4. Conclusions

The model described here justifies the high level of interest that researchers have in polymer QSAR studies [84]. This study also shows the importance warranted by electrostatic properties of polymers in studying the effect of polymer structure on biological activity [85]. The high level of accuracy provided by the model and the timelines of the calculations carried out further justify the use of the abrupt truncation of the electrostatic field in the QSAR with stochastic moments for RNA molecules, specifically mps activity. A similar approach has previously been used in molecular dynamic studies of other biopolymers [86–89]. Finally, this study confirms the versatility of the stochastic approach to solve problems related to polymers in biology [90].

Acknowledgements

The authors would like to express their sincere gratitude to the editorial board of Polymers for their kind attention in connection to the present series on stochastic molecular descriptors for polymers: Particular thanks go to Dr J.E. Mark.

References

[1] Kubinyi H, Taylor J, Ramdsen C. Quantitative drug design. In: Hansch C, editor. Comprehensive medicinal chemistry, vol. 4. New York: Pergamon; 1990. p. 589–643.
[2] Roy K, Ghosh G. QSAR Comb Sci 2004;23:526–35.
[3] Roy K, Leonard JT. Bioorg Med Chem 2004;12:745–54.
[4] Morales AH, González MP, Rieumont JB. Polymer 2004;45:2045–50.
[5] González MP, Dias LC, Morales AH. Polymer 2004;15:5353–9.
[6] Randić M, Vračko M, Nandy A, Basak SC. J Chem Inf Comput Sci 2000;40:1235–44.
[7] Hua S, Sun Z. Bioinformatics 2001;17:721.
[8] Khandogin J, York DM. Prot Struct Funct Bioinf 2004;56:724–37.
[9] Zhou H, Zhou Y. Prot Struct Funct Gen 2002;49:483–92.
[10] Arteca GA. J Chem Inf Comput Sci 1999;39:550–7.
[11] Mathews DH, Zuker M. Predictive methods using RNA sequences. In: Baxevanis A, Ouellette F, editors. Bioinformatics, a practical guide to the analysis of genes and proteins. New York: Wiley; 2003.
[12] Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley VCH; 2000.
[13] Cabrera-Pérez MA, Bermejo M, Gonzalez MP, Ramos R. J Pharm Sci 2004;7:1701–17.
[14] Cabrera-Pérez MA, Bermejo-Sanz M. Bioorg Med Chem 2004;12: 5833–43.
[15] Cabrera-Pérez MA, García AR, Teruel CF, Álvarez IG, Bermejo-Sanz M. Eur J Pharm Biopharm 2003;56:197.
[16] González MP, Morales AH. J Comput Aid Mol Des 2003;10:665–72.
[17] Gutman I, Rosenfield VR. Theor Chim Acta 1996;93:191–7.
[18] Jiang Y, Tang A, Hoffmann R. Theor Chim Acta 1984;66:183–92.
[19] Burdett JK, Lee S. J Am Chem Soc 1985;107:3063–82.
[20] Lee S. Acc Chem Res 1991;24:249–54.
[21] Markovic S, Gutman I. J Mol Struct (Theochem) 1991;81:81–7.
[22] Randic M. In: Schleyer PvR, editor. Encyclopedia of computational chemistry, vol. 5. New York: Wiley; 1998. p. 3018.
[23] Vorodovsky M, Koonin EV, Rudd KE. Trends Biochem Sci 1994;19: 309–13.
[24] Vorodovsky M, Macininch JD, Koonin EV, Rudd KE, Médigue C, Danchin A. Nucleic Acids Res 1995;23:3554–62.
[25] Krogh A, Brown M, Mian IS, Sjeander K, Haussler D. J Mol Biol 1994;235:1501.
[26] Chou KC. Biopolymers 1997;42:837–53.
[27] Yuan Z. FEBS Lett 1999;451:23–6.
[28] Hua S, Sun Z. Bioinformatics 2001;17:721–8.
[29] Hubbard TJ, Park J. Prot Struct Funct Gen 1995;23:398–402.
[30] Krogh A, Brown M, Mian IS, Sjeander K, Haussler D. J Mol Biol 1994;235:1501–31.
[31] Di Francesco V, Munson PJ, Garnier J. Bioinformatics 1999;15: 131–40.
[32] Chou KC. Curr Prot Pept Sci 2002;3:615–22.
[33] Chou KC. Peptides 2001;22:1973–9.
[34] Chou KC. Anal Biochem 2000;286:1–16.
[35] Chou KC. J Biol Chem 1993;268:16938–48.
[36] Chou KC. Anal Biochem 1996;233:1–14.
[37] Chou KC, Zhang CT. J Prot Chem 1993;12:709–24.
[38] González-Díaz H, Bastida I, Castañedo N, Nasco O, Olazábal E, Morales A, et al. Bull Math Biol 2004;66:1285–311.
[39] González-Díaz H, Marrero Y, Hernández I, Bastida I, Tenorio I, Nasco O, et al. Chem Res Tox 2003;16:1318–27.
[40] González-Díaz H, Ramos de AR, Molina RR. Bioinformatics 2003; 19:2079–87.
[41] González-Díaz H, Molina RR, Uriarte E. Bioorg Med Chem Lett 2004;14:4691–5.
[42] González-Díaz H, Molina RR, Uriarte E. Polymers 2004;45:3845–53.
[43] Ramos de AR, González-Díaz H, Molina RR, Uriarte E. Prot Struct Funct Bioinf 2004;56:715–23.

[44] González-Díaz H, Olazábal E, Castañedo N, Hernádez SI, Morales A, Serrano HS, et al. J Mol Mod 2002;8:237–45.
[45] González-Díaz H, Gia O, Uriarte E, Hernádez I, Ramos R, Chaviano M, et al. J Mol Mod 2003;9:395–407.
[46] González-Díaz H, Hernández SI, Uriarte E, Santana L. Comput Biol Chem 2003;27:217–27.
[47] González-Díaz H, Ramos de AR, Molina R. Bull Math Biol 2003;65: 991–1002.
[48] González-Díaz H, Uriarte E, Ramos de Armas R. Bioorg Med Chem 2005;13:323–31.
[49] Ramos de Armas R, González-Díaz H, Molina R, Uriarte E. Biopolymers 2005 [doi10.1002/bip.20202].
[50] Norberg J, Nilsson L. Biophys J 2000;79:1537–53.
[51] Mathews DH, Turner DH, Zuker M. RNA secondary structure prediction. In: Beaucage S, Bergstrom DE, Glick GD, Jones RA, editors. Current protocols in nucleic acid chemistry.
[52] González-Díaz H, Molina R, Hernández I. BIOMARKS® version 1.0, 2004. This is a preliminary experimental version. A professional version will be available to the public in the future, contact: humbertogd@vodafone.es.
[53] Mathews DH, Zuker M, Turner DH. RNAStructure version 4.0©, 2002.
[54] Zhou GP, Assa-Munt N. Prot Struct Funct Gen 2001;44:57–9.
[55] Gálvez J, García-Domenech R, De Julian-Ortiz JV, Soler R. J Chem Inf Comput Sci 1995;35:272–84.
[56] Gálvez J, García-Domenech R, De Gregorio AC, De Julian-Ortiz JV, Popa L. J Mol Graph Modell 1996;14:272–6.
[57] Kalate RN, Kulkarni BD, Nagaraja V. Biophys Chem 2002;99:77–97.
[58] Pisterer C, Mihailescu D, Smith JC, Reed J. J Med Chem 2004;47: 3723–9.
[59] Kowalski RB, Wold S. Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN, editors. Handbook of statistics. Amsterdam: North Holland Publishing Company; 1982. p. 673–97.
[60] Stat soft inc. Statistica© version 6.0; 2002.
[61] Harshey RM, Ramkrishnan T. J Bacteriol 1977;129:616–22.
[62] Nakayama M, Fujita N, Ohama T, Osawa S, Ishihama A. Mol Gen Genet 1989;218:384–9.
[63] Ohama T, Yamao F, Muto A, Osawa S. J Bacteriol 1987;169: 4770–7.
[64] Mathews DH, Zuker M. RNA secondary structure prediction. In: Clote P, editor. Encyclopedia of genetics, genomics, proteomics and bioinformatics. New York: Wiley; 2004.
[65] Balaban AT. Chemical applications of graph theory. New York: Academic Press; 1976.
[66] Trinajstič N. Chemical graph theory. Boca Raton: CRC Press; 1992.
[67] González-Díaz H, Uriarte E, Biopolymers; 2005 in press. doi: 10.1002/bip.20234.
[68] Wiener H. J Am Chem Soc 1947;69:17.
[69] Estrada E. Chem Phys Lett 2001;336:248.
[70] Marrero-Ponce Y, González-Díaz H, Romero-Zaldivar V, Torrens F, Castro EA. Bioorg Med Chem 2004;12:5331.
[71] Marrero-Ponce Y. J Chem Inf Comput Sci 2004;44:2010–26.
[72] Marrero-Ponce Y. Bioorg Med Chem 2004;12:6351–69.
[73] Marrero-Ponce Y, Montero-Torres A, Romero-Zaldivar C, Iyarreta-Veitía M, Mayón-Peréz M, García-Sánchez R. Bioorg Med Chem 2004 [doi10.1016/j.bmc.2004.11.008].
[74] Randič M. J Chem Inf Comput Sci 1991;31:311–20.
[75] Randič M. New J Chem 1991;15:517–25.
[76] Randič M. J Mol Struct (Theochem) 1991;233:45–59.
[77] Randič M. J Comput Chem 1993;4:363–70.
[78] Kowalski RB, Wold S. Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN, editors. Handbook of statistics. Amsterdam: North Holland Publishing Company; 1982. p. 673–97.
[79] Zhou GP, Kutbuddin D. Prot Struct Funct Bioinf 2003;50:44–8.
[80] Swets JA. Science 1988;240:1285–93.
[81] Hawkins DM. J Chem Inf Comput Sci 2004;44:1–12.
[82] Kalate RN, Tambe SS, Kulkarni BD. Comput Biol Chem 2003;27: 555–64.
[83] Stilf N, Baumann K. J Med Chem 2003;46:1390.
[84] Zhou GP. J Prot Chem 1998;17:729–38. Kundu S, Gupta-Bhaya P. J Mol Struct (Theochem) 2004;668:65–73.
[85] Norberg J, Nilsson L. Quart Rev Biophys 2003;36:257–306.
[86] Esteve V, Blondelle S, Celda B, Perez-Paya E. Biopolymers 2001;59: 467.
[87] Navarro E, Fenude E, Celda B. Biopolymers 2004;73:229.
[88] Navarro E, Fenude E, Celda B. Biopolymers 2002;64:198.
[89] Monleon D, Celda B. Biopolymers 2003;70:212.
[90] Freund JA, Poschel T. Stochastic processes in physics, chemistry, and biology. In: Lecture notes in physics. Berlin: Springer-Verlag; 2000.