Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 16 (2006) 547–553
Bioorganic & Medicinal Chemistry Letters

QSAR study for mycobacterial promoters with low sequence homology

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

[a]*Department of Organic Chemistry, University of Santiago de Compostela, 15782, Spain*
[b]*Chemical Bioactives Center and Department of Veterinary Medicine, Central University of 'Las Villas', 54830, Cuba*
[c]*Department of Ultrasound Medicine, Calixto, Las Tunas, 77400, Cuba*

Received 28 June 2005; revised 13 October 2005; accepted 18 October 2005
Available online 4 November 2005

Abstract—The general belief is that quantitative structure–activity relationship (QSAR) techniques work only for small molecules and, protein sequences or, more recently, DNA sequences. However, with non-branched graph for proteins and DNA sequences the QSAR often have to be based on powerful non-linear techniques such as support vector machines. In our opinion, linear QSAR models based on RNA could be useful to assign biological activity when alignment techniques fail due to low sequence homology. The idea bases the high level of branching for the RNA graph. This work introduces the so-called Markov electrostatic potentials ${}^{k}\xi_{M}$ as a new class of RNA 2D-structure descriptors. Subsequently, we validate these molecular descriptors solving a QSAR classification problem for mycobacterial promoter sequences (mps), which constitute a very low sequence homology problem. The model developed ($mps = -4.664 \cdot {}^{0}\xi_{M} + 0.991 \cdot {}^{1}\xi_{M} - 2.432$) was intended to predict whether a naturally occurring sequence is an mps or not on the basis of the calculated ${}^{k}\xi_{M}$ value for the corresponding RNA secondary structure. The RNA-QSAR approach recognises 115/135 mps (85.2%) and 100% of control sequences. Average predictability and robustness were greater than 95%. A previous non-linear model predicts mps with a slightly higher accuracy (97%) but uses a very large parameter space for DNA sequences. Conversely, the ${}^{k}\xi_{M}$-based RNA-QSAR encodes more structural information and needs only two variables.

There are different reasons to believe that the transcription and translation signals in *Mycobacteria* may be different from those in other bacteria such as *Escherichia coli*. Therefore, understanding the factors responsible for the low level of transcription and the possible mechanisms of regulation of gene expression in *Mycobacteria* requires examination of the structure of mycobacterial promoter sequences (mps) and their transcription machinery, including information concerning the RNA macromolecules involved. Unfortunately, mps present a very low sequence homology and mathematical methods to assign biological activity based on sequence alignment are not of practical use in this case.[1–3]

Different mathematical methods have been used for the analysis of genome information. The group of Professor Grau[4,5] has reported results on genome algebras. Markov models[6–9] are also well-known tools for analyzing biological sequence data. However, advances have not been reported concerning the treatment of this macromolecular structure–activity problem based on RNA secondary structure.

A real possibility to address this problem involves structure–activity relationships for naturally occurring RNA macromolecules with Markov molecular descriptors.[10] The use of molecular descriptors to derive quantitative structure–activity relationships (QSAR)[11] is an approach of major interest. Molecular descriptors[12] are numeric indices that codify either molecular or macromolecular structure. In this sense, González[13,14] and Morales[15] have applied molecular descriptors in macromolecular science. Furthermore, studies published by Roy, Toropov and co-workers[16–18] and others illustrate the use of the QSAR

Keywords: QSAR; RNA secondary structure; Sequence homology; Markov model; Mycobacterial promoters; Electrostatic potential.
* Corresponding author. Tel.: +34981563100x14938; fax: +34981594912; e-mail: gonzalezdiazh@yahoo.es

0960-894X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.10.057

approach. New sequences of molecular descriptors that can be extended to other biomacromolecules have been defined for DNA by Randič et al.[19] and by Hua and Sun[20] for protein sequence QSAR. Nevertheless, in spite of the potential use of classical molecular descriptors, they have not been applied in this area of science. In our opinion, the problem is that classical QSARs, such as those reported by Cabrera-Pérez et al.,[21–23] deal with branched rather than linear molecules such as DNA and protein sequences. For this reason, one may expect a higher success for classical molecular indices in branched biomacromolecules. The study of branched biomacromolecules with QSAR techniques could be also very useful for sequences with low homology. However, it must be remembered that the more commonly known branched biomacromolecule is the RNA secondary structure.[24]

Researchers worldwide have reported increasing interest in the characterization of the RNA macromolecular structure by computational techniques.[25,26] In this context, we propose here that 2D-RNA-QSAR is a promising field within biomacromolecule research. New analogues of our stochastic molecular descriptors will be introduced for the RNA secondary structure.[27–36] Two preliminary studies into secondary QSAR of RNA macromolecules have also been published.[37,38] These studies focus only on local properties of a single RNA molecule. As a consequence, the main aim of the present paper was to introduce in RNA-QSAR studies the Markov electrostatic potentials ($^{k}\xi_{M}$) previously used for protein QSAR.[39] In this sense, we intend to predict whether a naturally occurring DNA sequence is an mps or not on the basis of the $^{k}\xi_{M}$ calculated for its putative RNA secondary structure. Consequently, a more specific, but still important, aim of this work is to introduce a novel approach to predict mps. This work has led to a 2D-RNA-QSAR to discriminate between two groups comprising several RNA macromolecules, including 135 mps and 450 control sequences (cs).

In the work described here, we used the MC model theory to encode the 2D-RNA structure. We take into consideration long-range electrostatic interactions and secondary folding of the macromolecule. Norberg and Nilsson[40] have remarked on the importance of truncating long-range interactions throughout space to study folding and biological activity of nucleic acids. In this study, long-range interactions are allowed here to propagate stepwise throughout the 2D-RNA ribbon. This approach uses a MC for the propagation of long-range electrostatic interactions in a step-by-step manner during folding. Accordingly, this model uses the nucleotide-nucleotide short-range electrostatic interaction $^{1}\mathbf{\Pi}$ matrix (with elements p_{ij}). $^{1}\mathbf{\Pi}$ was built up as a squared table of order n, where n represents the number of nucleotides in the RNA molecule.

The elements of $^{1}\mathbf{\Pi}$ ($^{1}p_{ij}$), defined to codify information about the electrostatic interaction between nucleotides, were defined as[27–37]

$$ {}^{1}p_{ij} = \frac{\delta_{ij}\cdot\varphi_0(j)}{\varphi_j} = \frac{\delta_{ij}\cdot\varphi_0(j)}{\sum_{m=1}^{\alpha+1}\delta_{ij}\cdot\varphi_0(m)} = \frac{\delta_{ij}\cdot\frac{q_j}{d_{ij}}}{\sum_{m=1}^{\alpha+1}\delta_{im}\cdot\frac{q_m}{d_{im}}} = \frac{\delta_{ij}\cdot q_j}{\sum_{m=1}^{\alpha+1}\delta_{im}\cdot q_m}, \quad (1) $$

where δ_{ij} is the Kronecker symbol, which equals 1 for covalently or hydrogen bonded nucleotides and 0 otherwise. The value q_j is the electrostatic charge for the jth nucleotide; and d_{ij} is the topologic distance. The topologic distance is always equal to 1 due to δ_{ij} cutting off for long-range interactions. The sum carries over all directly interacting α nucleotides placed in the same row of $\mathbf{\Pi}$. This sum enables us to calculate the partial potential φ_j. This last value is the denominator in the probability expression (see the previous publication in this series for details). Chapman–Kolgomorov equations were used to calculate the vector $^{A}\boldsymbol{\pi}_k$ of absolute probabilities $^{A}p_k$(j). The $^{A}p_k$(j) are the probabilities with which short-range electrostatic interactions reach every jth nucleotide at distance k within the 2D-RNA framework. The sum of numerous successive short-range interactions thus results in long-range indirect interactions between nucleotides (see references for similar models)

$$ {}^{A}\boldsymbol{\pi}_k = {}^{A}\boldsymbol{\pi}_0\cdot(^{1}\mathbf{\Pi})^{k}, \quad (2) $$

where $^{A}\boldsymbol{\pi}_0$ is the vector of the initial probabilities $^{A}p_0(j)$ with which the jth nucleotide begins an interaction. This vector can be calculated using Eq. 1, but summing up to the n nucleotides rather than α. The result of the sum is called the total initial potential φ. The value of $^{A}p_k(j)$ depends on the specific nucleotides (identified by q_j) and on the connectivity between the nucleotides in the RNA molecule. Therefore, we can assert that any function having $^{A}p_k(j)$ values as arguments may encode information on the 2D-RNA structure. In this sense, we introduce here for the first time the molecular Markov electrostatic potentials ($^{k}\xi_{M}$). These new molecular descriptors may be considered as 2D-RNA backbone molecular descriptors[39]

$$ {}^{k}\xi_{M} = \sum_{j=1}^{n} {}^{A}p_k(j)\cdot\varphi_0(j) = {}^{A}\boldsymbol{\pi}_0\cdot{}^{k}\mathbf{\Pi}\cdot{}^{0}\boldsymbol{\varphi} = {}^{A}\boldsymbol{\pi}_0\cdot(^{1}\mathbf{\Pi})^{k}\cdot{}^{0}\boldsymbol{\varphi}. \quad (3) $$

The calculation of $^{k}\xi_{M}$ was carried out using our in-house software BIOMARKS 1.0©[41] (Bioinformatics Markovian studies). This software inputs the ct files generated by the software RNAStructure 4.0.[42] These files contain information concerning the secondary structure of the RNA macromolecules. These parameters represent the electrostatic interaction potential for nucleotides at a topologic distance equal to k or less. An RNA macromolecule depicted at the BIOMARKS 1.0 © interface is represented in Figure 1.

An example of the calculation of $^{k}\xi_{M}$ is shown in more detail in Table 1. Note that for the calculation of $^{0}\xi_{M}$, $(^{1}\mathbf{\Pi})^{0}$ becomes an identity matrix with 1 in the main

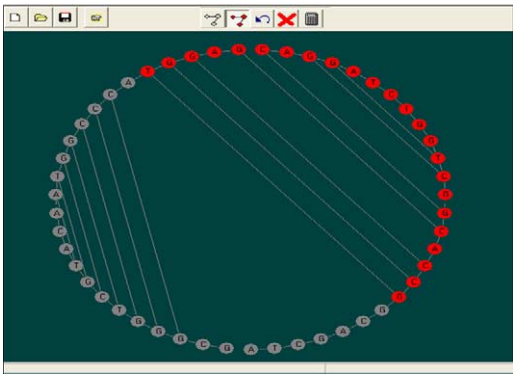


Figure 1. BIOMARKS 1.0 interface showing the circular representation for a folded RNA macromolecule of mps T3 from *Mycobacterium tuberculosis*, note the main stem highlighted in red.

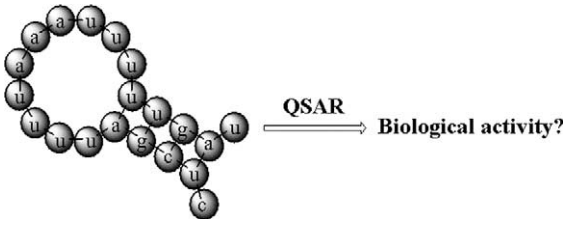
diagonal. In addition, the matrix presents 0 off-diagonal elements—indicating that we do not consider nucleotide-nucleotide interactions (isolated nucleotides). Conversely, $^{1}\mathbf{\Pi}$ is used for the calculation of $^{1}\xi_{M}$ taking into consideration direct interactions between covalently or hydrogen bonded nucleotides such as U_2-C_1 and U_2-C_3.

However, an example of a real numeric calculation is probably the best way to understand these indices. A detailed step-by-step calculation of $^{A}\boldsymbol{\pi}_0$, $^{1}\mathbf{\Pi}$, $^{0}\xi_{M}$ and $^{1}\xi_{M}$ for a fragment of a DNA sequence is depicted as concisely as possible in Table 1SM of the supplementary material. These steps are as follows:[39]

1. Obtain the DNA sequence and transform it into the corresponding RNA sequence.
2. Upload the RNA sequence into the software RNA-Structure 4.0[42] and build the RNA secondary structure.
3. Set the initial potentials $\varphi_0(j)$ that will be used for each kind of base j; in this step, one may consider either electrostatic potentials (as in this work) or other weights.
4. Build the vector $^{0}\boldsymbol{\varphi}$, whose elements are the initial potentials $\boldsymbol{\varphi}_0(j)$ of all the nucleotides in the RNA molecule.
5. Calculate the total initial potential φ as the sum of the initial potentials for all nucleotides in the RNA molecule.
6. Build the vector $^{A}\boldsymbol{\pi}_0$.
7. Calculate the partial potentials φ_j as sums of the initial potentials φ_0(j) considering only interacting nucleotides. These potentials are used as denominators in $^{1}\mathbf{\Pi}$ elements.
8. Build the $^{1}\mathbf{\Pi}$ matrix.
9. Calculate from $^{0}\xi_{M}$ to $^{5}\xi_{M}$ (six molecular descriptors in total).

Table 1. Definition of the $^{k}\xi_{M}$ molecular descriptors

Fragment of the RNA secondary structure for the DNA promoter sequence of gene S6 from *M. smegmatis*: $c_1$ucgauuuuaaauuuuugau$_{20}$.[a]



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

$$^{1}p_{u_2,c_1} = \frac{\varphi_0(c_1)}{\varphi_0(c_1)+\varphi_0(u_2)+\varphi_0(c_3)+\varphi_0(a_{19})} = \frac{\varphi_0(c_1)}{\varphi_{C1}}$$

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

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

$$^{0}\xi_{M} = {}^{0}\boldsymbol{\pi}^{T}\cdot(^{1}\mathbf{\Pi})^{0}\cdot{}^{0}\boldsymbol{\varphi} = \begin{bmatrix} ^{A}p_0(c_1) & ^{A}p_0(u_2) & . & . & ^{A}p_0(u_{20}) \end{bmatrix}^{T}\cdot\begin{bmatrix} 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & 1 \end{bmatrix}\cdot\begin{bmatrix} \varphi_0(c_1) \\ \varphi_0(u_2) \\ . \\ . \\ \varphi_0(u_{20}) \end{bmatrix} = \sum_{j=1}^{n} {}^{A}p_0(j)\cdot\varphi_0(j)$$

$$^{k}\xi_{M} = {}^{0}\boldsymbol{\pi}^{T}\cdot(^{1}\mathbf{\Pi})^{k}\cdot{}^{0}\boldsymbol{\varphi} = \begin{bmatrix} ^{A}p_0(c_1) & ^{A}p_0(u_2) & . & . & ^{A}p_0(u_{21}) \end{bmatrix}^{T}\cdot\begin{bmatrix} ^{1}p_{c1,u2} & 0 & 0 & . & 0 \\ ^{1}p_{u2,c1} & ^{1}p_{u2,u2} & ^{1}p_{u2,c3} & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & . & . & . & ^{1}p_{a19,u20} \end{bmatrix}^{k}\cdot\begin{bmatrix} \varphi_0(c_1) \\ \varphi_0(u_2) \\ . \\ . \\ \varphi_0(u_{20}) \end{bmatrix} = \sum_{j=1}^{n} {}^{A}p_k(j)\cdot\varphi_0(j)$$

[a] These codes are the used for a classical representation of a nucleic acid sequence. Please note that there are only four letters 'a, t, g and c' for a DNA sequence, using 'u' instead of 't' in the case of RNA. The letters represent different classes of nucleic acid bases and the number used immediately after a base indicates, when used, the position of the base in the sequence.

As can be noted the new indices are topologic in nature. Several of these indices have previously been used in QSAR studies over a long period of time. Topologic molecular descriptors have potential for RNA-QSAR problems but have not been applied yet to it. Other very interesting molecular descriptors are the quadratic, linear, and stochastic indexes recently introduce by Marrero–Ponce.[43–45] Our molecular descriptors are defined similar to other in the sense of the use of graph-theoretic concepts but use thermodynamic or electrostatic potentials analogies.[46,47]

Several authors have studied the mps problem from the point of view of DNA. For instance, Mulder et al.[48] listed −35 and −10 DNA regions of a few mycobacterial promoters. *Mycobacteriophage I3* and *M. paratuberculosis* mps have been studied by Ramesh and Gopinathan[49] and Bannantine et al.,[50] respectively. Kremer et al.[51] studied the DNA sequences essential for transcription in promoters like *M. tuberculosis 85A*. It is possible that DNA promoters with a high GC content in the −10 region are the true representatives of the mycobacterial type.[52] An analysis of *M. smegmatis* and *M. tuberculosis* promoters by Bashyam et al.[53] showed that there are similarities to *E. coli* 70 promoters. Strohl[54] studied DNA mps for *Streptomyces* promoters.

O'Neill and Chiafari[55] have also made efforts to develop statistical algorithms for sequence analysis and motif prediction. Two studies by Mulligan and McClure[56] and Mulligan et al.[57] pointed out that the variations within individual promoter sequences are responsible for the unsatisfactory results yielded by the promoter-site-searching algorithms. It can therefore be inferred that recognition of mps requires a powerful technique that is capable of unravelling those hidden pattern(s) in the structure difficult to identify visually.

Almost all the previous models have focused on DNA analysis and sequence alignment rather than RNA macromolecular secondary structure.[58] Our QSAR model attacks the pms problem from the 2D-RNA-QSAR point of view. QSAR techniques[59] have been classically used to seek models that encode structural patterns often hidden to a 'first eagle eye inspection'. Among these molecular indices, the stochastic molecular indices derived using Markov models stand out due to their potential for work with biomacromolecules.[60] Based on this reason we selected Markov models for the present RNA-QSAR study. The RNA secondary structure used was the lower energy structure predicted by the Mathews and Zuker model.[61] In the first instance, the statistical significance of the results must be discussed before arriving at any conclusions concerning the biology involved.

Linear discriminant analysis (LDA)[62,63] was used to classify RNA macromolecules as mps or cs. In the LDA, the output was a dummy variable mps = 1 when a sequence lies within the mps class or mps = 0 otherwise. In this problem, the inputs were the Markov electrostatic potentials (${}^{k}\xi_{M}$) with k in the range [0, 5]. The best discriminant equation found was:

$$\text{mps} = -4.664 \cdot {}^{0}\xi_{M} + 0.991 \cdot {}^{1}\xi_{M} - 2.432$$
$$N = 585 \quad \lambda = 0.41 \quad F = 38.8 \quad p < 0.001 \qquad (4)$$

where λ is Wilk's statistic, N is the number of RNA sequences studied, F is Fisher's statistics and p is the p-level (probability of error). This latter factor means that the hypothesis of groups overlapping with a 5% error can be rejected. A high Matthews' regression coefficient ($C = 0.903$)[20] was observed. This high C value[20] indicates a strong linear relationship between the structural descriptors and the classification of the RNA sequences. The significance of the two variables (${}^{0}\xi_{M}$ and ${}^{1}\xi_{M}$) in the model was demonstrated with the stepwise analysis (see Table 2). Conversely, the remnant four descriptors ${}^{2}\xi_{M}$, ${}^{3}\xi_{M}$, ${}^{4}\xi_{M}$, and ${}^{5}\xi_{M}$ do not have a significant relationship with the mps characteristic. The use of only six molecular descriptors to model a data set of 585 sequences prevents us by large from chance correlation. In physical terms, the above results show that, as in other studies,[64] there is a relationship between the electrostatic potential of the RNA molecule and its biological activity. However, in this case not all the electrostatic interactions affect the activity in the same way. The RNA-QSAR predicts that the possibility of a sequence acting as an mps decreases by a factor of 4.664 per unit of electrostatic potential considering isolated nucleotides (${}^{0}\xi_{M}$). Conversely, variations of electrostatic potential (${}^{1}\xi_{M}$) due to secondary structure folding[65] (direct covalent and/or hydrogen bonds) increase by a factor of only 0.991 the possibility of sequence to act as mps. Finally, long-term electrostatic interaction potentials (${}^{2}\xi_{M}$, ${}^{3}\xi_{M}$, and ${}^{4}\xi_{M}$) do not correlate with the mps activity. The detailed results of the forward stepwise analysis are given in Table 2.

The re-substitution technique was used to validate the model. Four different training and external predicting series groups were created interchanging at random 25% of the RNA molecules among them. The re-substitution accuracies and the average re-substitution accuracy (rs-average) were rs1 = 95.9%, rs2 = 96.6%, rs3 = 96.6% and rs4 = 96.5%, respectively, with the average rs-average = 85.7. Details of the classification matrices and other parameters for training and re-substitution experiments are given in Table 3.

Overfitting is a very interesting aspect that is sometimes ignored in QSAR.[66] This phenomenon can be detected by inspecting model robustness after removing the data in cross-validation. The robustness of the model was determined by carrying out the same four cross-validation experiments as mentioned above. These involved

Table 2. Forward stepwise LDA summary results

Variables	% total	% mps	% control	λ	F	P
${}^{0}\xi_{M}$	93.16	82.22	96.44	0.41	842.97	0.00
${}^{0}\xi_{M}$				0.40	859.49	0.00
${}^{1}\xi_{M}$	96.58	85.19	100.00	0.41	38.80	0.00
${}^{0}\xi_{M}$				0.48	626.18	0.00
${}^{1}\xi_{M}$				0.38	3.04	0.08
${}^{2}\xi_{M}$	96.58	85.19	100.00	0.38	0.92	0.34

Table 3. Robustness summary results

Train (96.6%)	%	mps	cs	Average (96.5%)	%	mps	cs
mps	85.2	**115**	20	mps	84.9	**86**	15
cs	100.0	0	**450**	cs	99.9	1	**337**
rs1 (96.6%)	%	mps	cs	rs2 (96.6%)	%	mps	cs
mps	86.3	**88**	14	mps	85.1	**86**	15
cs	99.7	1	**338**	cs	100.0	0	**342**
rs3 (96.3%)	%	mps	cs	rs4 (96.3%)	%	mps	cs
mps	84.2	**85**	16	mps	84.2	**85**	16
cs	100.0	0	**334**	cs	100.0	0	**335**

Table 4. Predictability summary results

Train (96.6%)	%	mps	Control	Average (96.4%)	%	mps	Control
mps	85.2	**115**	20	mps	85.3	**29**	5
Control	100.0	0	**450**	control	99.8	0	**111**
rs1 (95.9%)	%	mps	Control	rs2 (96.6%)	%	mps	Control
mps	84.8	**29**	5	mps	82.4	**29**	5
Control	97.3	1	**110**	Control	99.3	0	**111**
rs3 (96.6%)	%	mps	Control	rs4 (96.5)	%	mps	Control
mps	85.3	**29**	5	mps	85.3	**29**	5
Control	100.0	0	**111**	Control	100.0	0	**111**

re-substitution (interchange) of the training and predicting series and the results are given in Table 4. In this approach, it is important to avoid over-fitting phenomenon gaining control over other parameters such as, for example the variable-to-cases ratio ρ have to be >4. The coefficient ρ[67] for the present LDA model was $\rho = N/(N_v + 1) \times N_g = 585/(2 + 1) \times 2 = 97.5$. Where the number of variables is $N_v = 2$ and the number of groups is $N_g = 2$. Finally but yet importantly, we check out chance correlation.[68] Herein, only six variables were explored and two of them entered in the model based on a large database of 585 cases. Consequently, one can expect no chance correlation for the present study.[69]

Testing of the model fit to data and its robustness—although very important—is not the only characteristic of an acceptable QSAR. Details of the overall accuracy of the model are shown in Table 4—not for the data retained to perform the robustness study (Table 3) but for the RNA molecules leave-out from the data as external predicting series. It can be seen that the model maintains a similar average performance of 84.9% accuracy for mps and 99.9% for cs.

The data for mps name, sequences, training and cross-validation probabilities for all the RNAs used in this work are given in Table 2SM and Table 3SM of the supplementary material. Finally, as far as the simplicity of the model is concerned, the present linear QSAR model (two variables: ${}^0\xi_M$ and ${}^1\xi_M$) compares very favourably to a previous non-linear model.[70] This non-linear model presented only a slightly higher accuracy (97%) but makes use of very large space. The success of our RNA-QSAR model can be explained considering that 2D-RNA structure molecular descriptors encode not only sequences but molecular branching.

In accordance with the aims of the work presented here, two main conclusions can be drawn from the results and discussion. First, the 2D structure of RNA can be encoded with ${}^k\xi_M$ to develop QSAR studies in the presence of low sequence homology, as in the mps problem. Second, there is a very simple linear QSAR model for mps prediction that involves the first two members of the ${}^k\xi_M$ series (${}^0\xi_M$, ${}^1\xi_M$). Also the method proved to be a success for RNA-QSAR as other previous and new molecular descriptors for other QSAR problems.[71–73]

Acknowledgments

González-Díaz, H. thanks the Xunta de Galicia (BTF20301PR) for partial financial support. Thanks are given to Professor Léon Ghosez for his kind attention and an unknown referee for his useful comments.

Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.bmcl.2005.10.057.

References and notes

1. Harshey, R. M.; Ramkrishnan, T. *J. Bacteriol.* **1977**, *129*, 616.
2. Nakayama, M.; Fujita, N.; Ohama, T.; Osawa, S.; Ishihama, A. *Mol. Gen. Genet.* **1989**, *218*, 384.
3. Ohama, T.; Yamao, F.; Muto, A.; Osawa, S. *J. Bacteriol.* **1987**, *169*, 4770.

4. Sanchez, R.; Morgado, E.; Grau, R. *WSEAS Trans. Biol. Biomed.* **2004**, *1*, 190.
5. Sanchez, R.; Morgado, E.; Grau, R. *MATCH* **2004**, *52*, 29.
6. Chou, K. C. *Biopolymers* **1997**, *42*, 837.
7. Di Francesco, V.; Munson, P. J.; Garnier, J. *Bioinformatics* **1999**, *15*, 131.
8. Vorodovsky, M.; Macininch, J. D.; Koonin, E. V.; Rudd, K. E.; Médigue, C.; Danchin, A. *Nucleic Acids Res.* **1995**, *23*, 3554.
9. Hughey, R.; Krogh, A. *CABIOS* **1996**, *12*, 95.
10. Yuan, Z. *FEBS Lett.* **1999**, *451*, 23.
11. Kubinyi, H.; Taylor, J.; Ramdsen, C. Quantitative drug design. In Hansch, C., Ed.; Comprehensive Medicinal Chemistry; Pergamon, 1990; vol. 4, pp 589–643.
12. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley VCH: Weinheim, Germany, 2000.
13. González, M. P.; Dias, L. C.; Morales, A. H. *Polymer* **2004**, *45*, 5353.
14. González, M. P.; Morales, A. H.; Molina, R.; García, J. F. *Polymer* **2004**, *45*, 2773.
15. Morales, A. H.; González, M. P.; Rieumont, J. B. *Polymer* **2004**, *45*, 2045.
16. Roy, K.; Ghosh, G. *QSAR Comb. Sci.* **2004**, *23*, 526.
17. Roy, K.; Leonard, J. T. *Bioorg. Med. Chem.* **2004**, *12*, 745.
18. Toropov, A. A.; Roy, K. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 179.
19. Randič, M.; Vračko, M.; Nandy, A.; Basak, S. C. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1235.
20. Hua, S.; Sun, Z. *Bioinformatics* **2001**, *17*, 721.
21. Cabrera-Pérez, M. A.; Bermejo-Sanz, M. *Bioorg. Med. Chem.* **2004**, *12*, 5833.
22. Cabrera-Pérez, M. A.; Bermejo, M.; Gonzalez, M. P.; Ramos, R. *J. Pharm. Sci.* **2004**, *7*, 1701.
23. Cabrera-Pérez, M. A.; García, A. R.; Teruel, C. F.; Álvarez, I. G.; Bermejo-Sanz, M. *Eur. J. Pharm. Biopharm.* **2003**, *56*, 197.
24. Mathews, D. H.; Zuker, M. RNA secondary structure prediction. In *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*; Clote, P., Ed.; Wiley: New York, 2004.
25. Ruan, J.; Stormo, G. D.; Zhang, W. *Bioinformatics* **2004**, *20*, 58.
26. Ieong, S.; Kao, M.-Y.; Lam, T.-W.; Sung, W.-K.; Yiu, S.-M. *J. Comp. Biol.* **2003**, *10*, 981.
27. González-Díaz, H.; Molina, R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
28. González-Díaz, H.; Molina, R.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691.
29. González-Díaz, H.; Bastida, I.; Castañedo, N.; Nasco, O.; Olazabal, E.; Morales, A.; Serrano, H. S.; Ramos de, A. R. *Bull. Math. Biol.* **2004**, *66*, 1285.
30. González-Díaz, H.; Gia, O.; Uriarte, E.; Hernádez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Mod.* **2003**, *9*, 395.
31. González-Díaz, H.; Hernández, S. I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
32. González-Díaz, H.; Olazábal, E.; Castañedo, N.; Hernádez, S. I.; Morales, A.; Serrano, H. S.; González, J.; Ramos de, A. R. *J. Mol. Mod.* **2002**, *8*, 237.
33. González-Díaz, H.; Uriarte, E.; Ramos de, A. R. *Bioorg. Med. Chem.* **2005**, *13*, 323.
34. Gia, O.; Marciani-Magno, S.; González-Díaz, H.; Quezada, E.; Santana, L.; Uriarte, E.; DallaVia, L. *Bioorg. Med. Chem.* **2005**, *13*, 809.
35. Ramos de, A. R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Proteins* **2004**, *56*, 715.
36. González-Díaz, H.; Marrero, Y.; Hernández, I.; Bastida, I.; Tenorio, I.; Nasco, O.; Uriarte, E.; Castañedo, N.; Cabrera-Pérez, M. A.; Aguila, E.; Marrero, O.; Morales, A.; González, M. P. *Chem. Res. Tox.* **2003**, *16*, 1318.
37. González-Díaz, H.; Ramos de, A. R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
38. González-Díaz, H.; Ramos de, A. R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
39. Saíz-Urra, L.; González-Díaz, H.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 3641.
40. Norberg, J.; Nilsson, L. *Acc. Chem. Res.* **2002**, *35*, 465.
41. González-Díaz, H.; Molina, R.; Sanchez, I. BIOMARKS ©, 2004, version 1.0.
42. Mathews, D. H.; Zuker, M.; Turner, D. H. RNAStructure ©, 2002, version 4.0.
43. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
44. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
45. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, M.; Mayón-Peréz, M.; García-Sánchez, R. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
46. González-Díaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 1119.
47. González-Díaz, H.; Agüero, G.; Cabrera, M. A.; Molina, R.; Santana, L.; Uriarte, E.; Delogu, G.; Castañedo, N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 551.
48. Mulder, M. A.; Zappe, H.; Steyn, L. M. *Tuber. Lung Dis.* **1997**, *78*, 211.
49. Ramesh, G.; Gopinathan, K. P. *Indian J. Biochem. Biophys.* **1995**, *32*, 361.
50. Bannantine, J. P.; Barletta, R. G.; Thoen, C. O.; Andrews, R. E., Jr. *Microbiology* **1997**, *143*, 921.
51. Kremer, L.; Baulard, A.; Estaquier, J.; Content, J.; Capron, A.; Locht, C. *J. Bacteriol.* **1995**, *177*, 642.
52. Parbhane, R. V. *Analysis of DNA Sequences, Modeling Sequence Dependent Features and their Biological Roles Dissertation*; University of Pune: Pune, India, 2000.
53. Bashyam, M. D.; Kaushal, D.; Das Gupta, S. K.; Tyagi, A. K. *J. Bacteriol.* **1996**, *178*, 4847.
54. Strohl, W. R. *Nucleic Acids Res.* **1992**, *20*, 961.
55. O'Neill, M. C.; Chiafari, F. *J. Biol. Chem.* **1989**, *264*, 5531.
56. Mulligan, M.; McClure, W. R. *Nucleic Acids Res.* **1986**, *14*, 109.
57. Mulligan, M. E.; Hawley, D. K.; Entriken, R.; McClure, W. R. *Nucleic Acids Res.* **1984**, *12*, 789.
58. Ewens, J. W.; Grant, R. G. *Statistical methods in Bioinformatics, an introduction*; Springer-Verlag: New York, 2003.
59. Randič, M.. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Ed.; Wiley: New York, 1998; Vol. 5, p 3018.
60. Ramos de, A. R.; González-Díaz, H.; Molina, R.; González, M. P.; Uriarte, E. *Bioorg. Med. Chem.* **2004**, *12*, 4815.
61. Mathews, D. H.; Zuker, M. Predictive methods using RNA sequences. In *Bioinformatics, A Practical Guide to the Analysis of Genes and Proteins*; Baxevanis, A., Ouellette, F., Eds.; Wiley: New York, 2003.
62. Manhnhold, R.; Krogsgaard, L.; Timmerman, H. (Eds.), *Chemometric methods in molecular design*. Van Waterbeemd, H., ed. Vol. 2, VCH, Weinheim, 1995.
63. Statsoft Inc. STATISTICA, **2002**, *version 6.0*.
64. Zuberek, J.; Wyslouch-Cieszynska, A.; Niedzwiecka, A.; Dadlez, M.; Stepinski, J.; Augustyniak, W.; Gingras, A.-C.; Zhang, Z.; Burley, S. K.; Sonenberg, N.; Stolarski, R.; Darzynkiewicz, E. *RNA* **2003**, *9*, 52.

65. Mathews, D. H.; Turner, D. H.; Zuker, M. RNA secondary structure prediction. In *Current Protocols in Nucleic Acid Chemistry*; Beaucage, S., Bergstrom, D. E., Glick, G. D., Jones, R. A., Eds.; Wiley: New York, 2000.
66. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
67. García-Domenech, R.; de Julian-Ortiz, J. V. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 445.
68. Livingstone, D. J.; Salt, D. W. *J. Med. Chem.* **2005**, *48*, 661.
69. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 663.
70. Kalate, R. N.; Tambe, S. S.; Kulkarni, B. D. *Comput. Biol. Chem.* **2003**, *27*, 555.
71. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de, A. R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Int. J. Mol. Sci.* **2004**, *5*, 276.
72. Marrero-Ponce, Y. *Bioorg. Med. Chem.* **2004**, *12*, 6351.
73. González, M. P.; Terán, M. C. *Bull. Math. Biol.* **2004**, *66*, 907.