

**QSAR study for macromolecular RNA folded secondary structures of mycobacterial promoters with low sequence homology.**

[G003]

Humberto González-Díaz,<sup>a,b,\*</sup> Alcides Pérez-Bello,<sup>b</sup>  
Eugenio Uriarte,<sup>a</sup> and Yenny González-Díaz.<sup>c</sup>

<sup>a</sup> *Department of Organic Chemistry, University of Santiago de Compostela, 15782, Spain.*

<sup>b</sup> *Chemical Bioactives Center and Department of Veterinary Medicine, Central University of 'Las Villas', 54830, Cuba.*

<sup>c</sup> *Department of Ultrasound Medicine, Calixto, Las Tunas, 77400, Cuba.*

---

**Abstract.**

The general belief is that quantitative structure-activity relationships (QSAR) techniques work only for small molecules and, proteins 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 in RNA could be useful to assign biological activity when alignment techniques fail due to low sequence homology. The idea bases in 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 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.

---

**Keywords:** QSAR; RNA secondary structure; Sequence homology; Markov model; Mycobacterial promoters; Electrostatic potential.

**\*Corresponding author:** Humberto González-Díaz **e-mail:** [gonzalezdiazh@yahoo.es](mailto:gonzalezdiazh@yahoo.es), **telephone:** +34 981 563100 ext. 14938, **fax:** +34 981 594 912.

There are different reasons to believe that the transcription and translation signals in *Mycobacteria* may be different from those in other bacteria such as *E. 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.<sup>1-3</sup>

Different mathematical methods have been used for the analysis of genome information. The group of Professor Grau<sup>4,5</sup> has reported results on genome algebras. Markov models<sup>6-9</sup> 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.<sup>10</sup> The use of molecular descriptors to derive quantitative structure-activity relationships (QSAR)<sup>11</sup> is an approach of major interest. Molecular descriptors<sup>12</sup> are numeric indices that codify either molecular or macromolecular structure. In this sense, González<sup>13,14</sup> and Morales<sup>15</sup> have applied molecular descriptors in macromolecular science. Furthermore, studies published by Roy, Toropov<sup>16-18</sup> and others illustrate the use of the QSAR approach. New sequences of molecular descriptors that can be extended to other biomacromolecules have been defined for DNA by Randić *et al.*<sup>19</sup> and by Hua and Sun<sup>20</sup> 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.*,<sup>21-23</sup> deal with branched rather than linear molecules such as DNA and protein sequences. For this reason, one may expect 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.<sup>24</sup>

Researchers worldwide have reported increasing interest in the characterization of the RNA macromolecular structure by computational techniques.<sup>25,26</sup> In this context, we propose here that 2D-RNA-QSAR is a promising field within biomacromolecules research. New analogues of our stochastic molecular descriptors will be introduced for the RNA secondary structure.<sup>27-36</sup> Two preliminary studies into secondary QSAR of RNA macromolecules have also been published.<sup>37,38</sup> These studies focus only on local properties of a single RNA molecule. As a consequence, the main aim of the present paper is to introduce in RNA-QSAR studies the Markov electrostatic potentials ( $^k\xi_M$ ) previously used for proteins QSAR.<sup>39</sup> 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<sup>40</sup> 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  $^1\Pi$  matrix (with elements  $p_{ij}$ ).  $^1\Pi$  was built up as a squared table of order  $n$ , where  $n$  represents the number of nucleotides in the RNA molecule.

The elements of  $^1\Pi$  ( $^1p_{ij}$ ), defined to codify information about the electrostatic interaction between nucleotides, were defined as:<sup>27-37</sup>

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

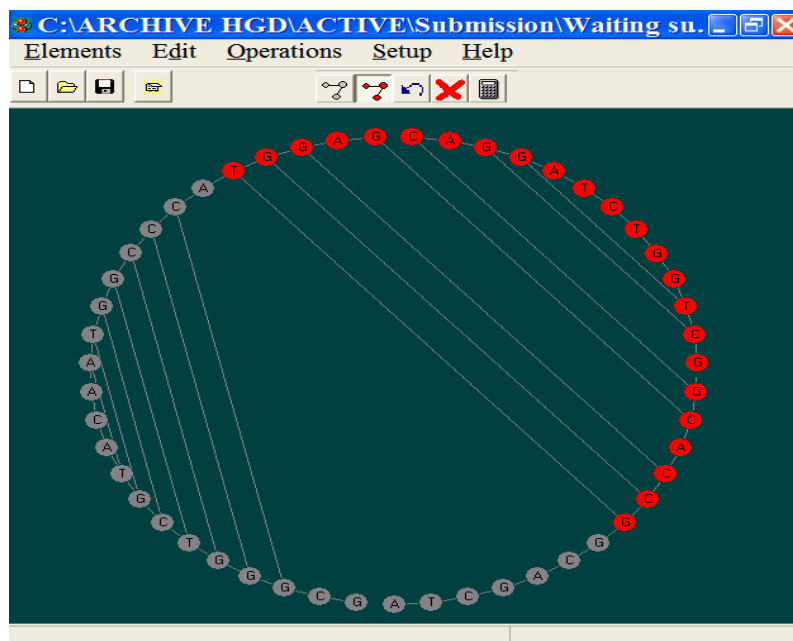
Where,  $\delta_{ij}$  is the Kronecker symbol, which equals 1 for covalently or hydrogen bonded nucleotides and 0 otherwise. The value  $q_j$  is the electrostatic charge for the  $j^{\text{th}}$  nucleotide; and  $d_{ij}$  is the topologic distance. The topologic distance is always equal to 1 due to  $\delta_{ij}$  cutting-off for long-range interactions. The sum carries over all directly interacting  $\alpha$  nucleotides placed in the same row of  $\mathbf{\Pi}$ . This sum enables us to calculate the partial potential  $\varphi_j$ . This last value is the denominator in the probability expression (see the previous publication in this series for details). Chapman–Kolmogorov equations were used to calculate the vector  ${}^A \pi_k$  of absolute probabilities  ${}^A p_k(j)$ . The  ${}^A p_k(j)$  are the probabilities with which short range electrostatic interactions reach every  $j^{\text{th}}$  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):

$${}^A \pi_k = {}^A \pi_0 \cdot ({}^1 \Pi)^k \quad (2)$$

Where,  ${}^A \pi_0$  is the vector of the initial probabilities  ${}^A p_0(j)$  with which the  $j^{\text{th}}$  nucleotide begins an interaction. This vector can be calculated using equation (1), but summing up to the  $n$  nucleotides rather than  $\alpha$ . The result of the sum is called the total initial potential  $\varphi$ . 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 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:<sup>39</sup>

$${}^k \xi_M = \sum_{j=1}^n {}^A p_k(j) \cdot \varphi_0(j) = {}^A \pi_0 \cdot {}^k \Pi \cdot {}^0 \varphi = {}^A \pi_0 \cdot ({}^1 \Pi)^k \cdot {}^0 \varphi \quad (3)$$

The calculation of  ${}^k \xi_M$  was carried out using our in-house software **BIOMARKS 1.0** ©<sup>41</sup> (**BIO**informatics **MARK**ovian **S**tudies). This software inputs the ct files generated by the software RNAstructure 4.0.<sup>42</sup> 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.



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

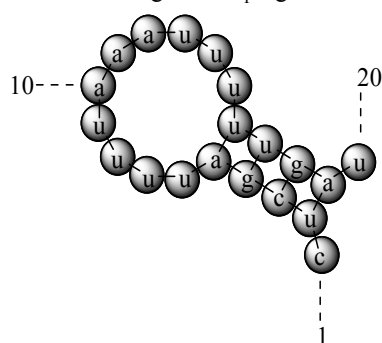
An example of the calculation of  ${}^k\xi_M$  is shown in more detail in **Table 1**. Note that for the calculation of  ${}^0\xi_M$ ,  $({}^1\Pi)^0$  becomes an identity matrix with 1 in the main diagonal. In addition the matrix presents 0 off-diagonal elements – indicating that we do not consider nucleotide-nucleotide interactions (isolated nucleotides). Conversely,  ${}^1\Pi$  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  ${}^A\pi_0$ ,  ${}^1\Pi$ ,  ${}^0\xi_M$  and  ${}^1\xi_M$  for a fragment of a DNA sequence is depicted as concisely as possible in **Table 1SM** of the supplementary material at the end of the manuscript. These steps are as follows:<sup>39</sup>

1. Obtain the DNA sequence and transform it into the corresponding RNA sequence.
2. Upload the RNA sequence into the software RNAStructure 4.0<sup>42</sup> 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  ${}^0\varphi$ , whose elements are the initial potentials  $\varphi_0(j)$  of all the nucleotides in the RNA molecule.
5. Calculate the total initial potential  $\varphi$  as the sum of the initial potentials for all nucleotides in the RNA molecule.
6. Build the vector  ${}^A\pi_0$ .
7. Calculate the partial potentials  $\varphi_j$  as sums of the initial potentials  $\varphi_0(j)$  considering only interacting nucleotides. These potentials are used as denominators in  ${}^1\Pi$  elements.
8. Build the  ${}^1\Pi$  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*:  $c_1ucgauuuuuuuuuugau_{20}$ .\*



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

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

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

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

$${}^0\xi_M = {}^0\pi^T \cdot ({}^1\Pi)^0 \cdot {}^0\varphi = [{}^A P_0(c_1) \quad {}^A P_0(u_2) \quad \dots \quad {}^A P_0(u_{20})]^T \cdot \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 1 \end{bmatrix} \cdot \begin{bmatrix} \varphi_0(c_1) \\ \varphi_0(u_2) \\ \dots \\ \varphi_0(u_{20}) \end{bmatrix} = \sum_{j=1}^n {}^A P_0(j) \cdot \varphi_0(j)$$

$${}^k\xi_M = {}^0\pi^T \cdot ({}^1\Pi)^k \cdot {}^0\varphi = [{}^A P_0(c_1) \quad {}^A P_0(u_2) \quad \dots \quad {}^A P_0(u_{21})]^T \cdot \begin{bmatrix} {}^1P_{c_1,u_2} & 0 & 0 & \dots & 0 \\ {}^1P_{u_2,c_1} & {}^1P_{u_2,u_2} & {}^1P_{u_2,c_3} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & {}^1P_{a_{19},u_{20}} \end{bmatrix}^k \cdot \begin{bmatrix} \varphi_0(c_1) \\ \varphi_0(u_2) \\ \dots \\ \varphi_0(u_{20}) \end{bmatrix} = \sum_{j=1}^n {}^A P_k(j) \cdot \varphi_0(j)$$

\* These codes are the used for a classical representation of a nucleic acid sequences. Please note that there are only four letters “a, t, g, 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.<sup>43-45</sup> 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.<sup>46,47</sup>

Several authors have studied the mps problem from the point of view of DNA. For instance, Mulder *et al.*<sup>48</sup> listed –35 and –10 DNA regions of a few mycobacterial promoters. *Mycobacteriophage I3* and *M. paratuberculosis* mps have been studied by Ramesh and Gopinathan<sup>49</sup> and Bannantine *et al.*,<sup>50</sup> respectively. Kremer *et al.*<sup>51</sup> 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.<sup>52</sup> An analysis of *M. smegmatis* and *M. tuberculosis* promoters by Bashyam

*et al.*<sup>53</sup> showed that there are similarities to *E. coli* 70 promoters. Strohl<sup>54</sup> studied DNA mps for *Streptomyces* promoters.

O'Neill and Chiafari<sup>55</sup> have also made efforts to develop statistical algorithms for sequence analysis and motif prediction. Two studies by Mulligan and McClure<sup>56</sup> and Mulligan *et al.*<sup>57</sup> 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.<sup>58</sup> Our QSAR model attacks the pms problem from the 2D-RNA-QSAR point of view. QSAR techniques<sup>59</sup> 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.<sup>60</sup> Based in 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.<sup>61</sup> 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)<sup>62,63</sup> 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:

$$mps = -4.664 \cdot {}^0\xi_M + 0.991 \cdot {}^1\xi_M - 2.432 \quad (4)$$

$$N = 585 \quad \lambda = 0.41 \quad F = 38.8 \quad p < 0.001$$

Where,  $\lambda$  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 ( $C = 0.903$ )<sup>20</sup> was observed. This high  $C$  value<sup>20</sup> 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 prevent us by large from chance correlation. In physical terms the above results show that, as in other studies,<sup>64</sup> 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<sup>65</sup> (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 ( ${}^2\xi_M$ ,  ${}^3\xi_M$ ,  ${}^4\xi_M$ ) do not correlate with the mps activity. The detailed results of the forward stepwise analysis are given in **Table 2**.

**Table 2.** Forward stepwise LDA summary results

Variables	% Total	%mps	%control	$\lambda$	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

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**.

**Table 3.** Robustness summary results

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

Over-fitting is a very interesting aspect that is sometimes ignored in QSAR.<sup>66</sup> 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 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  $\rho$  have to be  $> 4$ . The coefficient  $\rho$ <sup>67</sup> for the present LDA model was  $\rho = N/(Nv + 1) \times Ng = 585 / (2 + 1) \times 2 = 97.5$ . Where, the number of variables was  $Nv = 2$  and the number of groups was  $Ng = 2$ . Finally but yet importantly, we check out chance correlation.<sup>68</sup> 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 not chance correlation for the present study.<sup>69</sup>

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.

**Table 4.** Predictability summary results

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

The data for mps name, sequences, training and cross-validation probabilities for all the RNAs used in this work are given in as supplementary materials as **Table 2SM** and **Table 3SM** at the end of the manuscript. 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.<sup>70</sup> This non-linear model presented only 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. Firstly, 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. Secondly, 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 as success for RNA-QSAR as other previous and new molecular descriptors for other QSAR problems.<sup>71-73</sup>

## Acknowledgements

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

## 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 *Comprehensive Medicinal Chemistry*, Ed. C. Hansch. Pergamon. 1990, vol. 4, p. 589-643.
12. Todeschini, R.; Consonni, V. 2000. *Handbook of molecular descriptors*, Weinheim, Germany, Wiley VCH.
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. Randić, 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* P. Clote ed., John Wiley & Sons, NY. 2004.
25. Ruan, J.; Stormo, G.D.; Zhang, W. *Bioinformatics* **2004**, *20*, 58.
26. Jeong, S.; Kao, M.-Y.; Lam, T.-W.; Sung, W.-K. and 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.; Hernández, 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.; Hernández, 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.; Enriken, 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, NY, 2003.
59. Randić, M. Encyclopedia of Computational Chemistry. Schleyer, P. v. R., Ed. John Wiley & Sons, 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 A. Baxevanis and F. Ouellette eds., John Wiley & Sons, NY, 2003.
62. Manhnhold, R.; Krogsgaard, L.; Timmerman, H. (eds) *Chemometric methods in molecular design*. Vol 2, ed. Van Waterbeemd, H., 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* S. Beaucage, D.E. Bergstrom, G.D. Glick and R.A. Jones eds., John Wiley & Sons, NY, 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. Artificial neural networks for prediction of mycobacterial promoter sequences. *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.

**SUPPLEMENTARY MATERIAL**

**Table 1SM.** Steps for the calculation of  ${}^0\xi_M$  and  ${}^1\xi_M$  to a RNA molecule.

Step	Result
1 – DNA to RNA	atagcgtat → auagcguau
2- Derive 2D-RNA structure using the software: RNAStructure.	
3- Set potentials	$\varphi_a = 4.7, \quad \varphi_u = 4.6, \quad \varphi_g = 4.8, \quad \varphi_c = 4.5$
4- Build vector ${}^0\varphi$	$[\varphi_{a1}, \varphi_{u2}, \varphi_{a3}, \varphi_{g4}, \varphi_{c5}, \varphi_{g6}, \varphi_{u7}, \varphi_{a8}, \varphi_{u9}] = [4.7, 4.6, 4.7, 4.8, 4.5, 4.8, 4.6, 4.7, 4.6]$
5- Calculate total potential ( $\varphi$ )	$\varphi = 3(\varphi_u + \varphi_a) + 2\varphi_g + \varphi_c = 3(4.6 + 4.7) + 2 \cdot 4.8 + 4.5 = 42$
6-Build ${}^A\pi_0$ vector	$\left[ \frac{\varphi_{a1}}{\varphi}, \frac{\varphi_{u2}}{\varphi}, \frac{\varphi_{a3}}{\varphi}, \frac{\varphi_{g4}}{\varphi}, \frac{\varphi_{c5}}{\varphi}, \frac{\varphi_{g6}}{\varphi}, \frac{\varphi_{u7}}{\varphi}, \frac{\varphi_{a8}}{\varphi}, \frac{\varphi_{u9}}{\varphi} \right] = \left[ \frac{4.7}{42}, \frac{4.6}{42}, \frac{4.7}{42}, \frac{4.8}{42}, \frac{4.5}{42}, \frac{4.8}{42}, \frac{4.6}{42}, \frac{4.7}{42}, \frac{4.6}{42} \right]$ $= [0.112, 0.109, 0.112, 0.114, 0.11, 0.114, 0.109, 0.112, 0.109]$
7- Calculate partial potentials $\varphi_j$	$\varphi_1 = \varphi_a + 2\varphi_u = 2 \times 4.7 + 4.6 = 140 \quad \varphi_2 = 3\varphi_a + \varphi_u = 3 \times 4.7 + 4.6 = 187 \quad \varphi_3 = 2\varphi_u + \varphi_a + \varphi_g = 2 \times 4.6 + 4.7 + 4.8 = 187$ $\varphi_4 = \varphi_a + \varphi_g + \varphi_c = 4.7 + 4.8 + 4.5 = 140 \quad \varphi_5 = 2\varphi_g + \varphi_c = 2 \times 4.8 + 4.5 = 141 \quad \varphi_6 = \varphi_u + \varphi_c + \varphi_g = 4.6 + 4.5 + 4.8 = 139$ $\varphi_7 = 2\varphi_a + \varphi_g + \varphi_u = 2 \times 4.7 + 4.8 + 4.6 = 188 \quad \varphi_8 = 3\varphi_u + \varphi_a = 3 \times 4.6 + 4.7 = 185 \quad \varphi_9 = 2\varphi_a + \varphi_u = 2 \times 4.7 + 4.6 = 140$
8-Build the ${}^1\Pi$ matrix	${}^1\Pi = c5 \begin{bmatrix} \frac{\varphi_A}{\varphi_1} & \frac{\varphi_U}{\varphi_1} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\varphi_U}{\varphi_1} \\ \frac{\varphi_A}{\varphi_2} & \frac{\varphi_U}{\varphi_2} & \frac{\varphi_A}{\varphi_2} & 0 & 0 & 0 & \frac{\varphi_U}{\varphi_2} & 0 & 0 \\ 0 & \frac{\varphi_U}{\varphi_3} & \frac{\varphi_A}{\varphi_3} & \frac{\varphi_G}{\varphi_3} & 0 & 0 & \frac{\varphi_U}{\varphi_3} & 0 & 0 \\ 0 & 0 & \frac{\varphi_A}{\varphi_4} & \frac{\varphi_G}{\varphi_4} & \frac{\varphi_C}{\varphi_4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\varphi_G}{\varphi_5} & \frac{\varphi_C}{\varphi_5} & \frac{\varphi_G}{\varphi_5} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\varphi_C}{\varphi_6} & \frac{\varphi_G}{\varphi_6} & \frac{\varphi_U}{\varphi_6} & 0 & 0 \\ 0 & 0 & \frac{\varphi_A}{\varphi_7} & 0 & 0 & \frac{\varphi_G}{\varphi_7} & \frac{\varphi_U}{\varphi_7} & \frac{\varphi_A}{\varphi_7} & 0 \\ 0 & \frac{\varphi_U}{\varphi_8} & 0 & 0 & 0 & 0 & \frac{\varphi_U}{\varphi_8} & \frac{\varphi_A}{\varphi_8} & \frac{\varphi_U}{\varphi_8} \\ \frac{\varphi_A}{\varphi_9} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\varphi_A}{\varphi_9} & \frac{\varphi_U}{\varphi_9} \end{bmatrix} = \begin{bmatrix} 0.34 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 \\ 0.25 & 0.25 & 0.25 & 0 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0.25 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0.34 & 0.33 & 0.33 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.34 & 0.32 & 0.34 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.32 & 0.35 & 0.33 & 0 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0.26 & 0.24 & 0.25 & 0 \\ 0 & 0.25 & 0 & 0 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0.34 & 0.32 \end{bmatrix}$
9- Calculate ${}^0\xi_M$	${}^0\xi_M = {}^0\pi^T \cdot ({}^1\Pi)^0 \cdot {}^0\varphi = \sum_{j=1}^n {}^A p_0(j) \cdot \varphi_0(j) = 3x({}^A p_0(a)xq_A + {}^A p_0(u)xq_U) + 2x{}^A p_0(g)xq_G + {}^A p_0(c)xq_C$ $= 3x(0.112 \times 4.7 + 0.109 \times 4.6) + 2x0.114 \times 4.8 + 0.11 \times 4.5 = 4.67$
10- Calculate ${}^1\xi_M$	${}^k\xi_M = {}^0\pi^T \cdot ({}^1\Pi)^k \cdot {}^0\varphi = \sum_{j=1}^n {}^A p_k(j) \cdot \varphi_0(j) = ({}^A p_1(a1) + {}^A p_1(a3) + {}^A p_1(a8))x\varphi_a + ({}^A p_1(u2) + {}^A p_1(u7) + {}^A p_1(u9))x\varphi_u +$ $+ ({}^A p_1(g4) + {}^A p_1(g6))x\varphi_g + {}^A p_1(c5)x\varphi_c = 1235$

**Table 2SM.** Mycobacterial promoters name, sequences, training and cross-validation probabilities

Gene name	Sequence*	P	Pcv
<i>M. tuberculosis</i>			
<i>T3</i>	atcgacggccacggctggtctaggacgaggtaccggtaacatgctgggcg	1.00	1.00
<i>T6</i>	ccgtccagtctggcaggccggaacatcggcagcagataggctttaccag	0.98	0.99
<i>T26</i>	ctgcgagcatcatatgccgctgctggtgatcgggcaggatggtggacca	1.00	1.00
<i>T180</i>	gatcactccgagcatgcgccattgttgatagggcaggatgccctgg	1.00	1.00
<i>T101</i>	agcgatcgagccgacgtgatacctgaccgttgtgatagtgctggcggca	1.00	1.00
<i>T119</i>	ccccgtgctgtagtaggctccagccgacccgccgtaccatgcacaagt	0.99	1.00
<i>T125</i>	ccgaggttaaggactgagcatgggcccgataaagtgactattatggattct	0.82	0.86
<i>T129</i>	actcggcgagattaccgccgacggctcctggcgtggtcaatattccgca	0.99	0.95
<i>T130</i>	actccaacaggctgataacctcctgcgcctgctcgtctatgctgcgatccg	0.95	0.96
<i>T150</i>	gacccccgccacgtattgacactttgcgacacgctttatcattttccgac	0.43	0.55
<i>RecA</i>	ttcggagcagccgactgtcagtggtctctagtgtcacggccaaccgaccgat	1.00	0.86
<i>RrnA P1</i>	gagaacctggtgagctcggcgccgagatcgaacgggtatgctgtaggcgacggcacct	1.00	1.00
<i>gyrA</i>	gatgggcgaggacgtcagcgcggcgagctttatcaccgcaacccaaggatgttcggt	1.00	1.00
<i>cpn60</i>	ccccggcgatccccgtgctcaccacgggtgattccggggcgggatgcgtagcggactagc	1.00	1.00
<i>gyrB P1</i>	gatgtccgacgcacggcgcggttagatgggtaaaaacgagccagaagatcggccctggcgc	1.00	1.00
<i>gyrB P3</i>	caaggggcctcgccatattgccggtagggtccgcgcgacacctacggataacacgtcgtac	1.00	1.00
<i>85A</i>	gaagtgtggtgactacacgagcactccgggcccagcgcctgcagtctgacctaattcag	0.99	1.00
<i>85A</i>	cgccccgaagtgtggtgactacacgagcactccgggcccagcgcctgcagtctgacctaattcag	1.00	1.00
<i>gyrB P2</i>	agcggttggcaacgatgtggtgcatcgtaaagatcaccgggcccggcaccatcgtggcgca	1.00	1.00
<i>rrnA PCL1</i>	tgaccgaacctggtcttgactccattgccggatttatttagactggcagggttggcccga	0.97	0.98
<i>16S rRNA</i>	tgaccgaacctggtcttgactccattgccggatttatttagactggcagggttggcccga	0.98	0.97
<i>glnA</i>	tcggcatgccaccggttacgatcttgccgacctggccccacaatagggccggggagaccggcgt	1.00	0.99
<i>glnA</i>	ccaccggttacgatcttgccgacctggccccacaatagggccggggagaccggcgt	1.00	1.00
<i>KatG P<sub>A</sub></i>	ggatcactactgggtctatgtcctgattgttcgatatccgacacttcgcgatcacatccgtgat	0.94	0.96
<i>KatG P<sub>A</sub></i>	atcactggggtctatgtcctgattgttcgatatccgacacttcgcgatcacatccgtgat	0.82	0.85
<i>katG P<sub>B</sub></i>	gaggcggaggatcactactgggtctatgtcctgattgttcgatatccgacacttcgcgatcac	0.99	0.95
<i>katG P<sub>B</sub></i>	acgaggcggaggatcactactgggtctatgtcctgattgttcgatatccgacacttcgcgatcac	0.99	0.99
<i>katG P<sub>C</sub></i>	cctgattgttcgatatccgacacttcgcgatcacatccgtgatcacagcccgataacaccaactcct	0.35	0.51
<i>katG P<sub>C</sub></i>	ttcgatatccgacacttcgcgatcacatccgtgatcacagcccgataacaccaactcct	0.11	0.17
<i>purL</i>	cggcttgctcgtttccacgcccgcagcgcgatggggcctagctagactgcctccgtgatgtctcc	1.00	0.78
<i>purC</i>	atctcataccagagataaccagcacaggcgccgctcgtcggcggataggctggcgtgatgcccccg	1.00	1.00
<i>groE</i>	caggaagcaagggggcgcccttgagtctagcactctcatgtatagagtgtgatggcaatcggcta	1.00	1.00
<i>groE</i>	caggaagcaagggggcgcccttgagtctagcactctcatgtatagagtgtgatggcaatcggcta	1.00	1.00
<i>ahpC</i>	tgtgatatatcactttgctgacagcacttcacgggtacgatggaatgtcgtaaccaaatgc	0.52	0.64
<i>32 kDa</i>	acatgcatggatgcgttgagatgaggatgagggaaagcaagaatgcagcttgttgacagggttc	0.99	0.88
<i>10kDa</i>	aagcaagggggcgcccttgagtgtcagcactctcatgtatagagtgtgatggcaatcggctaa	0.97	0.97

10kDa	aagcaaggggccccttgagtgtcagcactctcatgtatagagtgctagatggcaatcggct	0.99	0.98
10kDa	aagcaaggggccccttgagtgtcagcactctcatgtatagagtgctagatggca	0.94	0.95
65 kDa	gcgtaagtagcgggggtgccgtcaccgggtgacccccgtttcatccccgatccggaggaatcac	1.00	0.98
mpt64	gagtcctggtcaggcatcgtcgtcagcagcgcgatgccctatgtttgctcgtcactcagatatacg	1.00	1.00
metA	tccggccccgcgatttggcgagcttcgtcgtgttcggtagcctggcattfaccgacgcgggggt	1.00	1.00
rpsL	gccgcaacccccgctttgacctgccagactggcggcggggtattgtggttgctcgtgcttgccggc	1.00	1.00
38 kDa	cgtcgccggactgtcgggggacgtcaaggacgccaagcgcggaaattgaagagcacagaaaggtatg	1.00	1.00
ppgk	cgggccgcagtttaaggtgagggtcatccacgtctcggcaggagattcgatgaccagcac	1.00	1.00

*M. bovis BCG*

<i>hsp60 P2</i>	cggtgcggggcttcttgactcggcatagggcagtgctaagaataacgttg	1.00	1.00
<i>rRNA</i>	tgaccgaacctggcttgactccattgccggatttatttagactggcagggtgccccgaa	0.98	0.99
<i>ahpC</i>	tgtgatatacactttgctgacagcagcttacggtagatggaatgtcgcaaccaaatacg	0.56	0.66
23K	gagtcctggtcaggcatcgtcgtcagcagcgcgatgccctatgtttgctcgtcactcagatatacg	1.00	0.89
mpb64	gagtcctggtcaggcatcgtcgtcagcagcgcgatgccctatgtttgctcgtcactcagatatacg	1.00	1.00
18K	tggcgtccgaaacacttgaggtgcggcccagcaaggggctacaggtttttccttcacctacgga	1.00	1.00
64K	gcgtaagtagcgggggtgccgtcaccgggtgacccccgtttcatccccgatccggaggaatcac	1.00	1.00
rpsL	gccgcaacccccgctttgacctgccagactggcggcggggtattgtggttgctcgtgcttgccggc	1.00	1.00
Mpb70	tggcgtccgaaacacttgaggtgcggcccagcaaggggctacaggtttttccttcacctacgga	1.00	1.00
alpha	cgactttcgcccgaatcgacatttggcctccacacacgggtatgttctggcccagcacacgacga	1.00	1.00

*M. leprae*

<i>16S rRNA</i>	tagtcaaccgggacttgactcctctgctggatctgtattaatctggctgggtgccgaag	0.99	0.99
<i>18Kda</i>	cttgtctatcacaacttgcatcaatataatcgaccagtctatatcaaatctatgtagtcagga	0.01	0.25
<i>18 Kda</i>	cttgtctatcacaacttgcatcaatataatcgaccagtctatatcaaatctatgtagtcagga	0.01	0.01
28-kDa	tcaatataaccactctggtcacactaaccatactgaccatcaaccgtggtgggtaatacc	0.07	0.06
groE1	agcagcgggcccggccttgagtgctagcactcgcgtgtatagagtgctagatggcagtcggccag	1.00	0.77
65 kd	gaattccggaattgcactcgccttaggggagtgctaaaaatgatcctggcactcgcgatc	0.92	0.94
36k	gttgggttctctcggagggcgcaccgctacgttagcgggatg	1.00	0.98
<i>SOD</i>	ggtgggcgcgatcatggcgcagcgttgattatgctagtcg	1.00	1.00
<i>rpsL</i>	cgccgttgggtcgtttgacctgcccagcaggacgggtattgtttctcgttctgacggct	1.00	1.00

*M. smegmatis*

<i>alrA</i>	gtctcggcctctgggacaatgggcgcccggagattatga	1.00	1.00
<i>S4</i>	aagccgaatcgagaccttttgggttcgtacacacttgctttataagcctcg	0.42	0.56
<i>S5</i>	aacaagattccgtaatcgtgtctggtggagctggtggaagcttgatccg	0.97	0.84
<i>S6</i>	catcgattttaaattttgatagagtgcaataa	0.00	0.24
<i>S12</i>	acctcgttatgcttctggctattttgatcaactttatacatgggcgggt	0.32	0.23
<i>S14</i>	tcaagcaccgaagccaacatggtttagtagtcgtttaccatgtgtacct	0.19	0.22
<i>S16</i>	tccacgcgaaccgcttcggcgtgccccgtttccctgtataatatacggcg	0.98	0.79
<i>S18</i>	gatcattgtcttctgttcttctgtataaagttgttactg	0.10	0.32
<i>S19</i>	ttttagtagccaaaggctctcaccactgagccatgatagtatccatccc	0.17	0.15
<i>S21</i>	acatggcattttcattttaaacaggactcaggtggtatggttgacatcga	0.99	0.79
<i>S30</i>	gatcagctatgttctcagtaaaattcggctatatgttgggtg	0.11	0.33

<i>S33</i>	gatccgctcttctatgatgccagttatggatctatggttatcg	0.49	0.39
<i>S35</i>	aactaaagtatgtgccgaattgacagtggtctagattatgatgctgcatc	0.04	0.16
<i>S65</i>	ggcacagctcgaagttctactacatggcttgctgaatccagtcacattact	0.22	0.17
<i>S69</i>	atcacgatgtctcatgcttggcttcaatgctccggctctacaatcagttca	0.24	0.23
<i>S119</i>	gatcaagaagccaatgatttgtaaacgcaattaatg	0.00	0.07
<i>gyrB</i>	cagaatcgggtgctgctctatctcgcggtagactggacgacggatctcaggc	1.00	0.75
<i>recA</i>	agagttcgaccggacttgctgggtgctgcttaacgtcacggccaaccgatcggg	1.00	1.00
<i>ask</i>	gtttgcccgcgcggcgccccacgatgaaccgcacggggtgacg	1.00	1.00
<i>acetamidase</i>	ggccggcgttcaccctgacttttatttcatctggatataattcggggaatggaaagg	0.91	0.94
<i>rrnB</i>	ctctgacctggggatttgactcccagttccaaggacgtaacttattccaggtcagagcgac	0.91	0.91
<i>rrnA P1</i>	gaaaacctgggtcagcctcggagccgagatcgagagagtaagctcgtaggaagcaagacc	0.97	0.96
<i>rrnA P2</i>	ctctgaccagcggattgcaatcgcgacgaacctcgtattatctttatgaagtcgccgcgga	0.93	0.94
<i>rrnA P3</i>	ccgggccagagcgcacttgacaagccagccgagatcgactaagctggcgaggtgcctcaga	1.00	0.98
<i>rrnA PCL1</i>	ccggtccagagcgcacttgacaagccagacaaagcagtattaagctggcaggggtgccccaaa	0.83	0.87
<i>rpsL</i>	ccgccgtgcacgagttgtttcgtcgcggcgcgccctggattgtggggatcgtgcctggccc	1.00	0.96
<i>rpsL</i>	cgtgcacgagttgtttcgtcgcggcgcgccctggattgtggggatcgtgcctggccccgaaa	1.00	1.00
<i>ahpC</i>	tgtgatatacaccttggctgacagcgcacttcacggcagatggaatgtcgcaaccaaatgc	0.62	0.71
<i>M. paratuberculosis</i>			
<i>pAJB303</i>	gacgacgagggcgggtggcgtcgcgggtgtagccgaacggcacgtgcgcgtaggccagat	1.00	0.91
<i>pAJB86</i>	ccacctactcccgatgacgttgacggctgggattaacggctccgcgtgctccaggagaca	1.00	1.00
<i>pAJB125</i>	gcaacgagcgcacattaagatcgagggcgggtatgtccctcaccgcccagctt	0.99	0.99
<i>pAJB300</i>	tcgagttcaagacctgacgctggccgacctcggcgcgcagccgaccgcgcagcgggtgcacg	1.00	1.00
<i>pJB305</i>	atccggacgggcagttgttgagttctgtcggacgggtggtggcgacatttccggcgagg	1.00	1.00
<i>pAJB304</i>	caccaggtacacgcaaggacaacggccgtatccggtaccaacgggtgtgcgagctggacgg	1.00	1.00
<i>P AN</i>	ctggggaagggtgaatcgacaggtacacagccgccatacacttcgcttcatgcccttacg	0.92	0.94
<i>pAJB73</i>	gatcgggtgcccgttgaaccggcccagctcccgtccagggtgacgtgctcgagctc	1.00	0.98
<i>pAJB301</i>	gatctggcggggcgggtccagttacaccgcgagttcgcgcacgctggccggcagcgttggacgcccg	1.00	1.00
<i>M. fortuitum</i>			
<i>repA</i>	gagctcgtgctggaccatacaccgggtgattaatcgtgggtactactaccaagc	0.84	0.88
<i>rrnA PCL1</i>	ccaggatgatgcaactgacttgccggcaagattcgaattaagctggcgggggtgccccaaa	0.97	0.94
<i>rrnA P1</i>	gaaaacctgttgaccctcggagccgagatcgaaagagtagggctgtaaacagcagtcggggcc	0.99	0.99
<i>rrnA P2a</i>	cgctgaccagccgatttgacctgtagggcagcccgctaatctttgaagtcgcgcggagcgg	1.00	1.00
<i>rrnA P2b</i>	ccgggccagagcgcacttgacaagccagccgagatcgactaagctggcgaggtgcctcagaccg	1.00	1.00
<i>rrnA P3</i>	caggatgatgcaactgacttgccggcaagattcgaattaagctggcgggggtgccccaaaacag	0.96	0.97
<i>rrnA PCL1</i>	actggggacgaggtcttgacgcccctgatcagatcggtatagactggcaggggtgccccgaaa	1.00	0.99
<i>rrnA P1</i>	gagaacctccgcagctcgcggcggcagatcgagagggctgcctgaaacatgccgtttacctgc	1.00	1.00
<i>rrnA P2</i>	aggggaccccccttttgactccgctcagacgtgggctattcttcaaccacaagccccaacgc	0.93	0.95
<i>rrnA P3</i>	ctggggacgaggtcttgacgcccctgatcagatcggtatagactggcaggggtgccccgaaagcaa	1.00	0.98
<i>M. phlei</i>			
<i>pKGR25</i>	cctgtacacctcgtgcaactcgcggaggaagcactatgccccgacgtcccggcctgg	1.00	1.00
<i>pKGR9</i>	accacgagcaccggctcgtcaggactcgcgacactcgtgttagacgactgggtgcagcatg	1.00	1.00

<i>pKGR38</i>	atctggtcgacctgctcgacgaggtcgatcatcttcttcatctcgccgaacgggatgccctgg	1.00	1.00
<i>ORF1</i>	acctcatggagcacttcgaggtcactgagcacgccacgaactacgagaggccgtgggactgg	1.00	1.00
<i>ORF2</i>	tacttttgtaccgttcgacaccagcgggttccgcttccctgccaatctcctgcaaaacaaccacaatg	0.34	0.51
<i>Mycobacteriophage I3</i>			
<i>rrnA P4</i>	gccaaaaccgggaatttgactcaggttcacgaactgatacggttccgagcggccgaaag	0.88	0.90
<i>rrnA P1</i>	ggcgggtctagtggcggacggcgtcacagaggtatac gatgtgttcatatcgaccgcggttac	1.00	0.97
<i>rrnA PCL1</i>	gccccgaccggaagttgactcaagttcattggacttggtagctggcgggttgcctgaa	0.99	1.00
<i>rrnA P2</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatacggttccaagtcgctcgg	0.80	0.85
<i>rrnA P3</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatacggttccaagtcgctcggaa	0.57	0.63
<i>pKGR1</i>	acacagaccaggagctcgacatgaccgccaccgcccctacagcgtcatctggttcgaaggcac cccggat	0.99	0.83
<i>Mycobacteriophage L5</i>			
<i>71 P2</i>	tacctgtcacaaggtttgctaccgagtggggcaggccgctacatttacgaccgcgtaacgcca	0.99	0.99
<i>71 P left</i>	tttgcgattagggcttgacagccaccggccagtagtgacattcttgtgacaccgcagcagc	1.00	0.99
<i>71 P1</i>	acaactgaatatggttccgacagcgaactaaattaggggtatccttgacaggcaccaacat	0.11	0.33
<i>M. avium</i>			
<i>Avi-3</i>	gcccggcagcgtgggctgataagtcttatcgggcactataaagtgtagtgggaaatcacct	0.96	0.75
<i>pLR7</i>	agccttgttggcggccaactgccggacgatcggggcgccatcgtcctcgagctcggccccgtgc	1.00	0.99
<i>M. neoaurum</i>			
<i>rrnA PCL1</i>	gcgagacagagaagcttgactcggcagacaagatagtttaagctggcaggggtgccccgaa	0.97	0.98
<i>rrnA P1</i>	gaaaacctggtcagcttggcgccgggatcagcaggtacactcgtaaagagaccggtcagatg	1.00	0.99
<i>rrnA P3</i>	gcgagacagagaagcttgactcggcagacaagatagtttaagctggcaggggtgccccgaaacg	0.98	0.98
<i>rrnA P2</i>	ctctgaccagcggatttgacttccgaaggcacaagaagttaacttttgaagtcgccgsgggag	0.97	0.97
<i>M. abscessus</i>			
<i>rrnA P4</i>	ggcgggtctagtggcggacggcgtcacagaggtatac gatgtgttcatatcgaccgcggttac	0.88	0.90
<i>rrnA P1</i>	gccccgaccggaagttgactcaagttcattggacttggtagctggcgggttgcctgaa	1.00	0.97
<i>rrnA PCL1</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatacggttccaagtcgctcgg	0.99	1.00
<i>rrnA P2</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatacggttccaagtcgctcggaa	0.80	0.85
<i>rrnA P3</i>	ccaaaaccgggatttgactcaagttcaccgaactgatcggttcccgggcccgttaciaa	0.57	0.63
<i>M. chelonae</i>			
<i>rrnA P2</i>	ccaaaaccgggatttgactcaagttcaccgaactgatcggttcccgggcccgttaciaa	0.98	0.88
<i>rrnA P1</i>	ggcgggttagtggcggatggcgtcaccgaggtatac gatgtgttcatatcgaccgcggtta	1.00	0.99
<i>rrnA PCL1</i>	ccccagaaccggaagttgactcaagttcattggacttggtagctggcgggttgcctgaa	0.97	0.98
<i>rrnA P3</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatcggttcccagcccgggaaa	0.50	0.62
<i>rrnA P4</i>	gccaaaaccgggaatttgactcaagttcaccgaactgatacggttccgagcccgggaaa	0.54	0.53

\* In this work we used a classical representation of a nucleic acid sequences. Please note that there are only four letters “a, t, g, c” for a DNA sequence. The letters represent different classes of nucleic acid bases.



**Table 3SM.** Mycobacterial promoters and control group name, sequences, training and cross-validation probabilities

number	Sequence*	P	Pcv
1	tacaataattggttaaattatataaccttctgtacacagtgc	1.00	0.89
2	tacatagccttggtgaccgaaaattccctagctag	0.99	0.99
3	agttctccgcattcacgggattataacgcatgtaagaaaaattaaatagcagctgtt	0.99	0.99
4	tccgcgtgggtattagcttatatgaccaacaattaaatgag	0.98	0.99
5	tttgtgtgagccataatttcgttgatccgagtcac	0.78	0.84
6	attcccaatttagcaagagctccatttacaagtcaaca	1.00	0.93
7	cgagttgatacttgacatggattataataaaaagatctatactgctatgagaggagtag	0.98	0.99
8	agaaggctttatatgaagactttgagatctatgacaataacagatgtagatatac	1.00	0.99
9	tgaatttcaatcagtatctatggagccatgtatagatggctctcgtacttattatcattggttaagt	0.90	0.92
10	gcaactccgggataacaaactcgagaggtagtacaacaacacctcaaacat	1.00	0.97
11	tccaagccgtccaaaatgagctattaaaggctttataaaagtaactatcccataag	1.00	1.00
12	ctcggtttaataaattaactgtggcctagagatcgttg	0.81	0.87
13	ctgagtgaaataatttacttacatgttctcctgattata	1.00	0.94
14	caagaaccgggagtatcaccggtatataaagactagtcgtgat	0.95	0.97
15	tatgggcaaagtatatatagcgagatgtaacatagactagatat	1.00	0.98
16	tcagatcccctgcgggaagatcttagtatacacctttgctcag	0.79	0.85
17	tcaatactacagttgagattctataaaaaatcctgcttaaagaactcaaaaa	1.00	0.94
18	tccactagcaaaatcagtggttgaaagagctaaatgaaacggagctactctatatattgtaccagct	0.97	0.98
19	acgctaataaataggctcgcgcgataaacgagctaaagaaaacgattaac	1.00	0.99
20	tatgattccccaaaaccaatacgggttagtatacagattgcacgtc	1.00	1.00
21	tggcataaatccgattctgaaagttagaattcaatcatcgagagcgaatcttcgcatt	0.98	0.98
22	gaactctcgtagatcctcttttcagggtgtgtacgaattatagaatc	0.96	0.96
23	attcacatcagcttcattccctctcgtctctataccctt	1.00	0.98
24	taacccttactaacaggttgacttaaatgcttaactgatacagagacgctgcactagcgtgccta	0.88	0.91
25	tcgccacagcttctaataagaaagtctgctactttccatccacaatattatggcttctgta	0.95	0.93
26	gtctacatttagtccactcaaaaataggttttccaataagacactactgcacgtcgga	0.99	0.98
27	ccctcaagatattaacagacaactaccatatttcaaaftaacctgtacagttttgtc	1.00	1.00
28	gtgttactttcacctatagaccgtctgagggcattcatatcaa	0.95	0.97
29	ctcctactgcaggatataagtggatctggaatatgtgacgtataaacacggtgt	0.67	0.75
30	cacgcgaatacggatgttttataacagataagattaaccaaagcagcggcttcgtatttggectcatt	0.95	0.88
31	gttcaaagcactcctctaatatcgttttatcgacgtctgctatacatctatagtgcccaacattcgacaca	0.96	0.96
32	ttgaaatacatgattagagaaacgattctttctagcgacagaaaaataactgacaataagcgtgttaaa	1.00	0.99
33	agaaccaatcggaaccatttattacgagtgactcaaaaaggaaatttagtttgaccattttattccc	1.00	1.00
34	ggtattccggacagctggaatataactgattttttgatcataacatagttttccgataagctat	0.98	0.98
35	taaaagtgttttcggcaaaagtgaatagtatatgtcgtagaactttaggcttgcagtaagattgg	0.71	0.77
36	attttaaagcttgccatataactacacctaggcaaacgttgtcgaatacttttaccgtgcccgat	0.97	0.91

37	atcagaaggacgtcttacgttcagctcaccaattcttggtaatacaaggcgataaatctcta	0.98	0.98
38	aacgatgggctcaatacgtactagtgtgtttggttaatcattcgcattaatttacagatac	0.95	0.96
39	cggcgtagtgcaaatgactcattacattcccacagccgttatcgactatcttatcttgcaatg	0.75	0.80
40	ccatcggagtacacgactaaaacatgaatttgcgaactcgtttgagcaatggccaatcatt	0.94	0.89
41	agaaatgcaccaatgtttgaatcttagggaccgctttatgggttagcgtctatctaa	0.86	0.88
42	tctaaatgttagtaatcacccctacacagagagaattaaacattgatctagctagatcg	1.00	0.96
43	caacagtcgaaattgccaaattgaatagcgtccctaattgaaatgtgtatagggattg	0.96	0.97
44	ctaccatatgtactttattagaaggcacacgagtaacttccgcagatttcccaa	0.98	0.98
45	aatacacaaaatataatctattttatcagcacttaggctgaccggccggga	1.00	0.99
46	ctaagtccgactatatctatataccgtctccgcaccagaattgtcatgatcgac	0.96	0.97
47	aagaacgtagttgctactggatcgtcgcaggctcccagaagaaccataaactg	0.67	0.74
48	ctgttgcaacgattaataatgacgactgggctttgtaaccctggttcaactctt	0.79	0.76
49	tcagccaggaagtctatgcagtaggatttctctcttatggtttatcaga	0.80	0.80
50	aaacatcgtaatatgttagatatgtgtgtttcaaaccttattgcacga	1.00	0.94
51	agatctacaaattaggttgaccttaagtgttatcacaatacgattggcgg	0.97	0.98
52	caattcccctgtgcaactcatttaggacatcgtagttgatataagtattt	0.99	0.98
53	gcgtcaatgatttcttacttttctacaactcagtgacttctctaga	0.99	0.99
54	cagcgtccttgttagactctatatgatcctattatgggccattatc	0.96	0.97
55	agaaagaaatagaggcattcgaagatattaacgggtacaaccattcac	1.00	0.99
56	aacaggggttctgttgactattttaattctccatttctctgaag	0.99	0.99
57	tccagattcgaacttattacatgatgaatcgtctcaaagtgagacct	1.00	0.99
58	tgcgaaaagtactaccaatcgaaaaacaagagcagtaagtggcagtt	0.99	0.99
59	atatggaacaaatttaacatttgagatttgaataaataac	1.00	1.00
60	gaagtgtgcttcgtgggaagattttagcaattgtcaacgtgc	1.00	1.00
61	gtgaactaactcaaccacgcgtctagaatcttaaagta	1.00	1.00
62	aacaacctcgcagttaccgcttcttaactcttagacta	1.00	1.00
63	gtaacattggacggcaacgattaagtaaggttagacatc	0.96	0.97
64	ctttgtgtagagagtggttttcttggatacaaatattt	0.95	0.96
65	taatccttaaagcctcaggtcgtctttacagcagt	0.98	0.97
66	tctacttctggatagatagtatctttctccaagtt	1.00	0.99
67	tttgtgctctttctacaaaaatgaaccctccga	1.00	1.00
68	gaaattccgcctgtgagaaatccctgcacatgatg	0.88	0.92
69	aataattgcgacgtcatctactgaactatttgattctactgtcatgcatcatgacttgataaagaaaaga	0.99	0.96
70	agttacatcctacgatcaaacagccgctgagagttttatataaacctcgtatctatttcattatagaaa	0.99	0.99
71	tagggatgatatagttttgaatataccaaaataacttctcgtctcgcgaaaaccatcacagaagatatga	0.98	0.98
72	tttgttcccgcctacaaaagccacctccacagtaaaagttgctttcgtacaatcgggaattacaaatgg	0.87	0.90
73	gtagttagttaatcttactcgcccgaattgagctattttagatacaggaagacagcaatagtagccgg	0.60	0.66
74	tcatttcaaacttttctagaagatagtgcataggttcataataaagcaaaaaataagagaatcttact	1.00	0.90
75	gtcgttatgatcagctcctgttgcgatcaaccgcctcaatctgtaaacccatcagctaaagaaaaga	0.82	0.87
76	attatatacttctatatacgtacatatgtgtattatgttaatgctgatccctacgctgctaacggc	0.98	0.94
77	ttaaaaagcatcgagggaaatggaacatgttattgcacgaaaccttctggagattagttaattgac	0.92	0.93

78	gaatgtacttattcctatftctttacaatacttcttattcgcataatgtcagtcgaatgtactcatat	1.00	0.98
79	tatctcgcaggggaacaatggaatattcctccacgaaagataaatcaaggtttctctatacaat	0.99	0.99
80	cttatgacttaagaaaggacataagcggggtatcgaggcttggacatgtcaataacttaa	0.97	0.98
81	gatgaactgacacgttcatcaatgtctatgtccaaacgatccttagaagaagataagcaac	0.99	0.99
82	catcttgatagcattgacttaatacatgaggctgaaagatcaatacgtcgcgatgacgccacc	0.91	0.93
83	tgccatgctcagttatccattccttaatggaaacctcagccagtgtaagattctatftaaatt	0.99	0.97
84	atactcattgggcgcaactaccgatccacgtcttattcttaagagaataaagtaaaaca	1.00	1.00
85	gagtccaacctaggacaacgatccaggagatattccatataatgaggatagacgac	0.94	0.96
86	ctaatacgaataatgagatctaaagtgactcactacattgcctaataatataatccat	1.00	0.98
87	tgtataaaagtggactattacagggagttcgttctccggcatttaccgaagcagctact	0.78	0.84
88	tgatgtttatctcgcaatgtcaatgatcgtacgtgagaaacaaatatcctgagta	0.98	0.93
89	tctttgtccaagcaagtttgctttgagtcagttatgacattggaatactacatg	0.82	0.86
90	agttattcttcccacttcagtcactgattagatatggcctagggttaaaaattgg	0.97	0.93
91	ttataccttgtaaacgagcctacgcgtgtacaacactgagtcataaagaagag	0.98	0.98
92	ccgaacttaatcgtacatgagtcctcaatgtgacaataaaaaatctcgcacct	1.00	0.99
93	attccaatactttatgtctatcatgaaccagtatactggtagcaaatggc	1.00	1.00
94	agtaaagctacctatataattgctgcttaagtagaagttaaaagaatagg	1.00	1.00
95	ttaacatactataggggtcatagccctggtaataaactcttctactcaat	1.00	1.00
96	tgtatcaagttattacacaaatftcataccatttccgccatgtttat	1.00	1.00
97	aaagataatgtcattcttcttgggctcgtaacagttcaatgacaa	1.00	1.00
98	gtttgtatggactaggcactagatgaaattatcactgatcctca	0.99	0.99
99	agttcgccatgagtattccgaatcattgactaactagaacaaacaac	1.00	1.00
100	cgctcagttaatttagctctcgaaggaagttcttacgcatgtgtc	0.86	0.90
101	gcttaaaagagcacacactatctgggcctgattgatataaaggaagg	0.89	0.88
102	tgaattaggcacagaacgtgctaggagatatgatgtcataaataaag	0.99	0.96
103	acttagagggtgcggaatatggccatatccgacaatttaggtagtagt	0.89	0.92
104	taagcataltttactattcgtctacgtcggagacctccacagatat	0.99	0.96
105	tttagcacttgcttatatgcagcacttttctatcggatccgggtt	0.79	0.85
106	tataatgtttacceattattaacctactttttgacgttaagc	1.00	0.94
107	gaaatcaataaccagacttttagtactagtggtagtctctcggcgc	0.93	0.95
108	acaatgattagcctgagcaaaccttgagtttagcagtagtccac	0.98	0.97
109	catcgttggtagagcttccatcagttgtattctcacttatg	0.72	0.80
110	ggactatgcttetaagataatctcggcatacacttacggt	0.97	0.90
111	ccgggtaacctagtgtcaagcaaacgaatcaactagtaaatt	1.00	0.99
112	tgaattacaaaacggtatagatagcaacctttggctcattgat	1.00	1.00
113	caatgtgacatactttatccccaattgggtgacacttaa	1.00	1.00
114	aaacgaatcttataaccgagagaatctgacaccagtgtaa	1.00	1.00
115	caaatcgatattccaacgataaacctgcttctctctg	1.00	1.00
116	atcagaagacatctactcgagctactttttacaaaatat	1.00	1.00
117	ccctgaaatactagctgattcatgacagaggaccatact	0.98	0.99

118	ccagctggactgattttgtaatttaagtgcactc	0.98	0.98
119	taatcaaacctacagtagaaccaaggaccagtccg	1.00	0.99
120	agaatctcattaaaaaggtaggtgattatTTTTG	1.00	1.00
121	ctTTTTccaccatgcatgtaccctcctaggtat	0.98	0.99
122	gagaaggataacaaataattctcctgTTtagaacatctaagactctaataagaggctatttgcgaagag	0.98	0.98
123	tacttagccgattcagcactaccttcaaagggtgcgaataaccacgttacaactgactagcgttacaatac	0.94	0.95
124	accgtacagaaaaagatatcaagcgttgatcaatatagtctacttaataaggagcttgataccagcatcatc	0.98	0.97
125	gttcagaaatccagttatgcttattgaaactataggtgcatataacagtttaattaaatctccaatcggt	1.00	0.99
126	acggcccttgctgttaaatatactgggaacactgaaaactagatattattgtttcaatacaaat	0.99	0.99
127	tacaaacgaaaaaataagagttactattacctatccccagctaagttctggtagagcgttttctaa	0.99	0.99
128	gactctttctctcgcaataaatggctgtataaaatattcactttttatctgcatacgattaacatg	1.00	0.99
129	tactgataagttatagtcacgtgtattccgacgtaattgacaattccgagaagaaccagttacacg	0.93	0.95
130	ctcgccgacccaaaaatatcaatttgcagatgttaacaatgaattccgtgaggttaagataatccg	0.98	0.97
131	tgctagttgccaacattttcgtgaccattgacatctggtaaagtgtacatgatcaattgatttgaat	0.92	0.94
132	gaactccctagtcgaacatgagactactagaattagacgttctgaaacgatctacaaagtttatctt	0.99	0.97
133	aacgaacagttcagtggttatagcaagctaaagctaaacccttcgagatctgaataaatTTA	1.00	0.99
134	gttcttaaagatccctgattttgacgttataactcggcatgctggtagatagaagatagagtag	0.72	0.79
135	agaccacattagttggattcaagttatcgcgatcgaaccggacttactggttataga	0.79	0.78
136	agaaggactatcatctttatcgattattggcaggtcttttaggtacgcatccattgaa	0.90	0.87
137	tagcctattaaatacaggaaataactataacatctaaaaaaggaaattgtcgggtgcgta	1.00	0.97
138	ctagtaacgtatgcgattgaacggtagtactgatttaagcgatgtgcacaatcagtagt	0.68	0.76
139	gttttagagtgccgaacgtttatatactacgaataattcgggtaccttctgatacgcact	0.85	0.81
140	ataggatagtcattattataatcttctggtatgcatatggtacagcaatatacgtgat	0.99	0.95
141	gcagtttaacttcatagtttacatgcggtaaaaagaacatcaagtatttaaca	1.00	1.00
142	ctttctcaatcatgctatactaggtcgeatccacgatacgtcagagcataagaaa	0.98	0.99
143	aaatcactggcgtggaacaactgcaaattgatactggtgtcatcttcccctca	0.89	0.92
144	cattgtgcgactatgcttacaaccgaatacataaggacattaaaggagaaaaaca	1.00	0.97
145	tatagaagaaaccataaacacgaactccatattagtagcacgcctagaccgtgata	1.00	1.00
146	caaatctacactatagaaccgggccgaatatctgtgattgatttctgaagtcc	0.95	0.97
147	tacaatagctcaacaaatggtcagtaaaaaatttgcgacggaagtcacacatcct	1.00	0.99
148	tcgaccaagatccacattttctgcaatgacttttagctaaaactcgattca	1.00	1.00
149	gtatatattgtgaagcataatgcacagtacaccattaacgctaatctaata	1.00	1.00
150	gtctctaggcctcatgtgcctacaattttctagctaagactccgtgactaa	0.86	0.90
151	atagttacagatgctcacactgtgaatattacatcgaataacgcccggtgcc	0.98	0.94
152	gaccgacatctaaatattgaaactaatggaacggctctcgcgtgcattt	0.86	0.89
153	actgtctacctcgtgcttttgaaccgaaataactggctatgatc	0.91	0.89
154	tcctacgaagtacaaagaattcgtaatagcgttttcgatcatgagc	0.99	0.96
155	gccaggatctgagaatatctatgattgaaactaaattgcgtttatct	0.99	0.99
156	gaaatagcactactccgctcagaagtggttagtctcaattcaccag	0.85	0.89
157	cattctcaaaattaccaagacatacaaaagtcagaattaaacaatatg	1.00	0.96

158	tctacatctacacaaaacgtattggcgtacggtagttcaatcacg	0.98	0.99
159	acaatctggtattcacgtttgataagtcatatacgtcatgaca	1.00	0.99
160	agcataaatctctcactgccatagaagttgttcgcagtctgc	0.94	0.95
161	cactagtaaattatgaatctctactgggcgatatcgaagatt	1.00	0.98
162	taggttgatttacagtaattcggcttgcaggctccaca	0.80	0.86
163	cgcttgattatcgtctccacccaagaatgatatgattta	0.99	0.93
164	actgaatcttagtgactttcactccgccgccttaa	0.91	0.93
165	aagcaatggttgattgatgaaatattgagtcttgt	0.97	0.95
166	cctcatccagcgtatacatccctggcactacgcgtact	0.82	0.87
167	gtcaagcctatctcttagtgcactctgatgtacaaagaat	1.00	0.94
168	gaaccaattgctccctcgcacgcgttatggctttaat	0.86	0.90
169	tcttaatgcagctcttttattgtattcattatctc	1.00	0.96
170	catagagtctacagtaacaagcagttgtaccgggc	0.87	0.91
171	aggcctgaaggattttattgtatgagagagatatt	0.96	0.93
172	tattattctaattcatgaattgatgacatgtggagcacttccttgtaatagataagggtagatctgggt	0.83	0.86
173	aaaacatttaagaataaataggttaaggcaggctcgcctgtctcaatagagcaagtaatagtactgaa	0.92	0.90
174	atthataaaggcactgcacttaaacgagataaagtccaatgactcgtaaaagttttgcagtttagact	0.99	0.98
175	cgattcatcgattgtcaacactaccaacattactcgtgcatacattacaatcctcagattgaag	0.99	0.99
176	cggtatgaaactctatttacaaccatccttaacgcttataaaggcgcacttcaagtgtac	0.98	0.99
177	cgaccacttaagactggcctgtatcgtttcgtttaagtgtacaatcagccgattaacc	0.80	0.85
178	atagatcataggttataatcactcctggagtgaacagaaacagtgaggacacgctctatc	0.91	0.88
179	tgtttatcggagaaacatatagcaactttcaatgagtaccacctgaacgtaggacatcg	0.96	0.94
180	aaccaagaacttgactagcaaatgctacgctaactgacccttattcaaccttca	1.00	0.99
181	tagtcgaatcacttgaagatagtttagtgaatactaattgtccccttgcctcata	0.99	0.99
182	gagcattcatatatctaataatgttcattcaagctattaaggatctagaatcgaac	1.00	1.00
183	taactttgtttatthaagcgtgtttatatagaataaagttagaaaatctatgat	1.00	1.00
184	ccaatgcgcgtttttatcaatgctttcaggagtcatcgaatattaacc	0.98	0.98
185	acccttgaccatgactgaaatgctcagaaaagattgaagcgggtgaaagc	0.86	0.89
186	ctgacgctagaccaattgaagagcctataacacgctgtatcagaaaaag	0.99	0.95
187	agaatcgtggaatacattagagacatcacgttctttaaactgactgaggg	0.96	0.97
188	agtacactcagttccccgaacgacttcaacattttggggagggtaaa	0.81	0.85
189	gattgccaacattacaggagcttatggaaagcagcctagcgccttta	0.78	0.78
190	ttccctggctctatatctacaacataagatactgctgacagtctatg	0.99	0.93
191	ttaatcatagtcaagattcttatgtagtatacaggggaaagca	1.00	1.00
192	agatgatgaaagacctaaccgactaacagcgggaatcagctata	1.00	1.00
193	gtgaatcgttaaggagcttattcacttacgggtctctgacaaa	0.88	0.91
194	aaactgggatgagacaatctacaagatctgtaacgttttag	0.99	0.96
195	gaatglatatttgggtggcattacgttcgctatattaatctcag	0.97	0.97
196	tcccgcgcaagctattataaattatgttataatgctgattagt	1.00	0.99
197	accagcattgacgggaaagtgcttagtatgttcaaacg	0.92	0.95
198	aactctggtaagctgatgaaattatggfactgtaccctg	0.93	0.92

199	gcgatacctactcaccaggcaagagtcaccagt	0.77	0.82
200	gtgaatacgaaccattcaggagaacttataatcgt	1.00	0.93
201	gtaaagcattatgctgtaagtagtcccgataagat	0.99	0.99
202	ttcccacgaccctaacgcaattgtacctgtgagac	0.93	0.95
203	gttctagattcgggtacaaggttaacatagcctg	0.97	0.96
204	caaccgctatcagggtatagaaagaaatatacgt	1.00	0.99
205	ttcacgtctcatggccaattggatttatggca	0.92	0.95
206	gacttcgaagaatccataaccgtaagaaaaactttt	1.00	0.97
207	gcaaacgatgatagtttgaagataagttacgtaa	1.00	1.00
208	tccttgcgatgtaaagtagcgtgttacgataac	0.95	0.96
209	acagatgattgaccacgggttacctatcttttactttt	0.81	0.84
210	actcgaataagctgaatgaattactctatagtagt	0.99	0.95
211	tcatataaataaacgtgtgcatacgtcaaggctgtct	0.96	0.97
212	gaataaatgtttttaaactgaccaatgtatctgtac	0.97	0.96
213	acggccctaatcccgtcaccaaactactttttgaa	0.67	0.74
214	tgtgaattgaaaatgaggcttcacaaaaatggatt	0.77	0.74
215	catttaaaatttagctgttagctccagaatgcatat	0.99	0.94
216	gcttaagtatacccgatgtaaatgctttacagccag	0.91	0.93
217	aaatccacacggtatcacaacagtttacgggctt	0.99	0.97
218	tgccttacttttattgggtacctgtttctacaaaac	0.91	0.93
219	gtctagtagagccatattgaaaggaattcaaaaag	0.96	0.94
220	agcccagataacgctttgggttcataaaactgccat	0.95	0.95
221	atcagagaatactgaagctctatgttgacttaacac	1.00	0.99
222	gatgttgataagttagtaagaatcagaaatgagtc	0.98	0.98
223	agtctactcgttaattgtgaacgaaaatcgtctga	0.93	0.94
224	cacattgttgatcagcattcggctatagacgccaa	0.95	0.94
225	cctctcttaactacaatagtagtactttacaagttc	1.00	0.98
226	aattcaactaacacaactttagcaaatctacagtac	1.00	1.00
227	acttattataatgctgcagtctatacgaacagatgt	0.98	0.98
228	acctggcattccttaataaataacgaaatctacctc	0.94	0.95
229	tgacggagtttattcagtgaccatatccggttagtt	0.77	0.82
230	ttaggaactttagattgtgtttgctagataaattag	0.99	0.93
231	gttcgttcgataaagctcttagcactattcgttatt	1.00	0.99
232	tctcagcggacaagtaccgtctacatttaagctagc	0.57	0.68
233	atataacatattagtcgaacaggaaggtaagtaca	0.99	0.88
234	aggttcgaagaactcagaataagacggtaaagagg	0.99	0.99
235	cataactgaatcgatcattctataacctgcgttga	1.00	1.00
236	cattttaaagggtggaccatgcgacacacaccatac	0.97	0.98
237	aaatatcgcgtcgaagttgcaaagtagaccactaga	1.00	0.99
238	aatgcttgataacgctgggtgtagatataaaatga	0.99	0.99
239	taaaatgactatgcatcacagacagaaccttggcat	1.00	1.00

240	atataattctacttcgctggaaaaataagagtactaatcttac	1.00	1.00
241	agaagaggaatgatttgctgctgccattatctggtaacaacg	0.87	0.91
242	tactgccgaatgatagcacgtccccaattattatgaacatggt	0.99	0.95
243	tcaccagtaaaaatgagtagccgaaaagagcgtttccta	1.00	0.99
244	ccattaatattatattcttgagaaaaactcggcacg	1.00	1.00
245	tctgtgccaagcataaacgattggatcatgctaata	0.99	0.99
246	gtacacatcgtaacgtagatagcgtatacaaaa	1.00	1.00
247	cataaagaaattgtacatgtaattttccatagca	1.00	1.00
248	tacatccatgattggctggtgtgtgtttattagc	1.00	1.00
249	tgtcaatataagttccagtagtttcagactatcg	1.00	1.00
250	cttttgctgacaatctacttatagatataacttcaataacttctcttaataaatatcagcgcgctttt	1.00	1.00
251	gtcggaaaatctatatctggagagaccatttataatcggaaataatgcacatagaacaatctatcacatag	1.00	1.00
252	cttatcgtatcctaattaatagcaaatcctggaatcacgatatagtcgtttaaaggccttatctccat	0.99	0.99
253	atatctaaatgctaaggtaccaacagactggttactgtagatacatcttacgatgtgcgtattacc	0.95	0.96
254	gatacttagagggtttaaagtagtactacagacggaggatagatataatatacttagacgtatcaagc	0.97	0.96
255	aataagtaggcgtccttataagcatccatctccagatcgtcaggtctctgcactatgaacacgtcac	0.77	0.82
256	gaaagaaccatataaatcaacaacagtcacaattttcgaatttatgatgatccagctaaatccttt	1.00	0.94
257	catgaattaacttcgaaatgtatccaacaagattatgactggatagctctatcatgcgcaa	0.99	0.99
258	tatgaaaattgttgagcagcagtcagagatacatcgggtgtaatagaatcatctattgatac	0.94	0.96
259	ggccacatggtaaaactaatcagtaatttcaccccgttggtgacctaccaataggtc	0.85	0.87
260	ccccactaccctatatatctgtacaatataatftaattaagagtaggacggagtgtc	0.99	0.95
261	gctgctccatgcctgtttaaactgtacaaggctattcagattagctacttgata	0.95	0.96
262	atctacctgttacccttacatcgatttgctacagttaagagctagagttatcgca	0.97	0.96
263	gcgatcagatagactcatcgatccgtacagtggtcttaaccacacaaaaata	0.99	0.99
264	gccatggttataactaacgtcctcaggttgcaactgaagaatagaaattcagagc	0.92	0.94
265	ctaacaggtgatcaggatgtcgtaaaccgtttctattccagagtagtcaactc	0.83	0.86
266	aaatgaaggacgatatactctattccagggttctagcgggtctatacacattgcc	0.81	0.81
267	aaggattaggcgactacaaatattgagatggccatagtttacgacgcaattgat	0.96	0.92
268	catcgggagtattaactaatccggacagctaaatattgtattcgtgcaaatcgagt	0.95	0.95
269	ttaaatcaaaaccctttaccggtcaaatccacacatggaaaactaaatcgcgat	1.00	0.99
270	gcgtattagtttactccgaaaaacgtcctaccgaaaaaacctattttaatcgt	0.99	0.99
271	agctttgcccactaacgtcgcacgggataatccaattctctatgacaattgcgt	0.80	0.85
272	taggtcgcgatggttagtgatcgaaaaagcgtataccaataaacattg	1.00	0.94
273	ctaatgattagtctatttttttaattattttgtttaagtaacgaa	1.00	1.00
274	agctccttgccgtaccgcagcaaaatagtactaatgacggtaaccga	0.90	0.93
275	aacaaatataagacttagtcaactcagcagagacgaatgatta	1.00	0.97
276	tttgatcagaacctacggcttatgcttaataattctggtgatgatc	0.91	0.94
277	ttcgaggacactatcgagaactcatggcacgagcgtataaatcaac	0.90	0.91
278	aggcaaaagatgacactaagggtggcatgagttattatgtctgagatc	0.85	0.86
279	tcgtgacttacaacatcaatcgaaaaatacagctactattaacat	1.00	0.96
280	tgaaccgttgagctcaagattatgagagcaataatgacacgcct	0.99	0.99

281	ataggtctgtgcatagacaagaacgtctattcagtaaaccagttct	0.99	0.99
282	aacacctactccggtactctcatttctcgttcttagtcaatcg	0.97	0.98
283	ttgaattgatcgcctcctattctacatactcataaacgtgtcg	0.99	0.99
284	aatttagatttacgatatgagacaaaaaacatcgtatgtcggg	1.00	1.00
285	ttaggtcagcccagagaaatattgagtacccgttcgaa	0.95	0.96
286	agcatcaaatgtttgtgaaaaaactacagtataatgga	1.00	0.98
287	aagctagtttgtgacaccaaacggtaacatcagaatca	1.00	1.00
288	cctatgataatacagagcattcttaatctgacagag	1.00	1.00
289	gagaccccgctactcctgaagtaattaaaggcacaatattcagataattaagacttttctgttctat	0.98	0.98
290	taatctgcctaatacaatagaacatagatttatgtggcggatgtccgatatttctattccatagcggag	0.81	0.85
291	gttattctcatcatatcgacaatgcaataggttcataagagaagatcaatgatgtggcttaggtgtaacg	0.86	0.84
292	acttcagcataagatcaagtggatacaaaagttaaagactgtgacactgcagtaattgtaacgaggagag	0.95	0.93
293	taagagagctttataataaaaatttgcggatgataatctcaccgcgactaatcgattgttctgatagaat	0.98	0.98
294	gcaaatcaagtatatgattccctagtaagtattgttatcaaaacgcactaaggaatgcagcgtact	0.99	0.99
295	ccatataaaacataatattacatcaccgaaaactcctgctcaaatfgtgtggtggaagctgtatgt	0.99	0.99
296	aagtctgtgctatgaacccatctcgggtctcgttctaatcattcatgtaaatgcaatggtaatagat	0.85	0.89
297	tacaaagtcttatccgtttgtgtacaacaaaagggttctgagataggatgatgccctaagactat	0.96	0.93
298	tgttacttaagattataaccagcgagacttagttgttctcattccgacctatgaatcatt	0.97	0.97
299	ccgaatgtttttttcttatagagaatcttattgcgtgttaaattctaaccattaagaact	1.00	0.99
300	ttgtccaggtcagagtgtagcatttgacacatactccaacgttttagactatca	0.86	0.89
301	gacgacagatggagaattctgctttaatatcagtaatgaaacgatttgcagaga	0.98	0.94
302	tacagtattgaatctaaggtaataatcctataagttcgaaactcggcttttaggac	0.99	0.99
303	attataaatgctaacaacgcacttgaggtaagttcaataaacataaaagtgagaat	1.00	1.00
304	tcacctcgtcaatfaatccttttacaagttaaagggacgtaatatctatcgggt	0.99	0.99
305	acgcgcatttaattgttatcgatgatttgggtggactagagatatgatcgaactat	0.89	0.91
306	tcgtattatatggtgataaataaacaccactccggagtgtagcagcaagcttt	0.98	0.95
307	tgtgtgattcataaacatgccttccctcctgaacctaagctacaagagtaaattt	0.99	0.98
308	atcccagaaattttacgttttaggtcactatagattcacatgtgaaatgata	1.00	1.00
309	aaaggacaactaataaatacatcaaccctcgttcataatagtacttaa	1.00	1.00
310	aactctaactttaagatataggaggattttgataagctacatataaggcgt	0.99	0.99
311	aaaaagttgctaattcgtgtgtagggtagcaagactaacattaaacgttt	0.99	0.99
312	agagaattccaagcacatgtctatgtacattacttatgaatttaaaa	1.00	1.00
313	attggaacggacgtgcattgtcaagtataatacttttggcttta	0.95	0.96
314	gtactgcttgcgcacgtagaaaaaaagctgctgaacgaactccctag	0.78	0.83
315	atcatcgtgcataatacaggtaaagggcataaaagtagcagccatg	0.90	0.87
316	gaatgattagatgatcgtatgtaaatgtattcgttgatcgggtg	0.78	0.82
317	gttcagcctctaagataagtacgcagcgaagatcagtggtgaaact	0.89	0.86
318	aaccatcattttgcagtaaatcatagattacaacaaggcctacact	1.00	0.97
319	agaagagcaggggtgcaaaaagaatcataatgataatttaactgtct	1.00	1.00
320	gtaaacaatctctttatcttctaggtataacctattacactcaatt	1.00	1.00
321	tacctatcggttccgaatacgtgcttaactataatgtttac	0.99	1.00



322	cgataaccgcacatttgcgtaaataaataatftagagaggtag	1.00	1.00
323	tctcagtttatgccagcggaaatccccagatccagtataaag	0.97	0.98
324	tgcataaactctattatacacaattaaagcgggaccaacgat	1.00	0.99
325	cactgcaacccttaaacacacgtgagaaaatatgaaata	1.00	1.00
326	gaggcatagtatcgcactgaatatgtggaatgtctacg	0.72	0.80
327	tcgttgttctcaatgtgagtgtagtcttattgtaagat	0.79	0.77
328	ctgattgtagagctccgcaggcaaatccaaatccgaat	0.96	0.91
329	tacaatatcgatagatggcaccgcagagccccttacctgtcattcacggatcaattgctttggagaact	0.92	0.93
330	acctgtcattcacggatcaattgctttggagaact	0.97	0.95
331	agataagttggattatatgcgctcgattcctatct	0.97	0.97
332	ttatatcatatgtaaagaatctattcacttgctct	1.00	0.99
333	ttagacttgagaaggtagcgaatataagctacctcggtattaagttgctgaaaataataactcctactaa	0.99	0.99
334	gcaactcattccgcctaaacttaataactagctttcgtggacctgtttggctcggtatcaatcaa	0.74	0.80
335	ataaacctagaatcatgttcgtgcgcaagatacttgacacgtggatctataaactattaatata	0.99	0.93
336	atcatgtattattgttgaacctgtacgaggatcacctggaccgctcaagagttccataaatgtc	0.80	0.85
337	tctcttaacgccattaacttacataccttgataaattgcaagaccttctagtataactaacttg	1.00	0.95
338	tgtatagattgacaactcttgtgtctgaattatgtcagctagaacattatcaaaggaacaaact	0.99	0.99
339	tatatgtcccattcactactacccttaaatataattacacaacagcatcatcctactagatagaat	1.00	1.00
340	gtctgctgattcatgatacaataagaactagattacagcattcgcctaaggttaccttatga	0.99	0.99
341	atlttggtgtctgactattttactggagtagtgttcaaccgaataactcttagatggtaac	0.89	0.92
342	tattatagacagatacacaactctactcgttcaaatcaccacgaaccgattactatgtcccag	1.00	0.97
343	ctgattatthaagtaaatcagggattcacatagcagatctattatagccaaggacgggg	0.85	0.89
344	acgtacctagtaaccgacgtgtagtcttccatattgtctgggattttcatgttca	0.68	0.72
345	agacagacatatgtgataacattttacaagtctagctgtctgcaatgactagta	0.99	0.91
346	acattatgtgataacgatctcatagataaaatgggtgtgatgggtgttcaactg	0.98	0.98
347	ggtaaacactcgttgatcataaaaaattgtaagcgaatgatatgcgtagatt	1.00	0.99
348	atlttcctactagcacagtttagcatgtctttagtaaaaatttagttattct	1.00	1.00
349	ctgtacagaataacttggtacaatcattatgctacacttttagttgactaa	1.00	1.00
350	aaagaatgattttggtagagttaattattgctaccttcagtgaacaatc	1.00	1.00
351	ggtacgtgtttatcaataatagtaactcactattatcctatcggtgc	0.99	0.99
352	tgtacataaccagtttgcatcgcgttggtggcaactatccaca	0.80	0.86
353	cttagttctagcccttggttaatcatataaatgtatcttaatttt	1.00	0.94
354	tatacettatcgagactagatattaatagcacaagatataagcat	1.00	1.00
355	agtattctgccagtagttaaataatgaattcaatctgttacct	1.00	1.00
356	tattaatattgaaaacgatcttattcatgagagaccagcagctagct	1.00	1.00
357	gaaataaataaattattttattaaattgccgggtgattcatc	1.00	1.00
358	attcacgctgaagacataagtaataggattttccctgggttac	0.99	0.99
359	tttatagccgtacgtacggcgtcatagcaagcaagaacag	0.90	0.92
360	tggtgagactaaattttgatttaaaccgtgtgacagttatg	0.53	0.63
361	cacctggaattcgttactgcctagcatgctcagagatatgt	0.88	0.78
362	gcctggatgatataattgtatcaggcctaattgtagtgaat	0.96	0.93

363	atgaaagactcaaagggcctcgaatctcgcgattaattcct	1.00	0.99
364	tactaggggttaaatatttctacagattggtattggtc	0.97	0.98
365	aagggggaccgctcgggaattacctaattcagattacc	0.68	0.77
366	agtctttccatgcagtttaattatggacgattgtacgg	0.90	0.84
367	ctatacactttacatacacaagtctcgttttgagtcg	1.00	0.97
368	tacgcaattatactaccaggcactttgtaacttttt	1.00	1.00
369	agggattttgtgttaaaagtctgatgaaatcacccct	1.00	1.00
370	tgatatctactcttgtagtaacacaatggcagtcgt	0.99	0.99
371	ttaagtattcaagtatagattagttattctcaca	1.00	1.00
372	ttttattcattaaagagacgtagcagttgatgat	1.00	1.00
373	ttacattcatccacaagaggttaagcttagatTTTT	1.00	1.00
374	ttatctatgatcaaacagcataaagtgcctctgtt	1.00	1.00
375	tacttagtacgtaacgggccgacatgatcagtgat	0.81	0.87
376	tcacctggtagctgctaaatctgatagttgaaatt	0.99	0.94
377	cgtccgctaaccaaatcattattttagcgttgcgatatgtggagagtagataatcaaccctattaatcac	0.97	0.97
378	aaatatagcctttatctactcttatggtgcggcaggtccgagaattattgacaagatcactgaagattca	0.88	0.90
379	actgttacatactcaacggaacacgtaaacacactgctctacgataagtccttgcggtatccgggtat	0.90	0.89
380	ttgtaatacccttaataagagatctacaatgccactactaaatcccgaatcatcttgcgctataca	1.00	0.97
381	aacttgaaagccggaagcggaaataaactttgtattaactgtccttaaagttctattacgatggtag	0.96	0.97
382	tccatgcctttatcgcattttaatgggtattcttagatattcatcatgtgcctactaagtcttaagcc	0.97	0.96
383	aaaaggaccatccagattaacaaggaccacacttttgacagtctagttgtccctcccagtagctgt	0.88	0.90
384	gcagcaagcttcatacaaacacccaattcctagttaaagttgataattcaactgccaaacgcg	0.98	0.95
385	atgataagttaaatcgcatatgtcttaaatgggtatgaataaaaacgtcctggaagattaaat	1.00	0.99
386	gaaagtcggcctctctagataaacactataccatagaagactcctgcataagtgattcct	0.94	0.96
387	cagtctgagtagacatagaattagcatgtatatcccgcgtagtaattctattcggtaga	0.95	0.95
388	cttcttaaatagctatgggccgagtacacacagaaccgggtcagctagcatatatcctc	0.70	0.77
389	cacacagttttatcgaacgctgctatctatcatcatgaaccctgtctttgaagcttg	0.93	0.87
390	cgccaattcgaattaaatcgcataggactgtctacgtgtctcgaacgacctagc	0.82	0.85
391	attttagcgaagggcggaccatcgaagataaccgattatacccactcacatcttg	0.88	0.86
392	actagtagccgattgaaaaggactgaattcctagatgcatttcaactactcct	0.98	0.95
393	gtcagaagatacttaataatattgatctatggaccctgaaacctgctga	1.00	0.99
394	gtgtttacattgtcgattatatcatcacaaggcccagcatatatttagtaa	1.00	1.00
395	tgacagtgaagttaatcatatgcttattatatttatactggttctcacia	1.00	1.00
396	actgtaattcttatattagcttttctgttttacattataagtaccccca	1.00	1.00
397	aagaaatgtcgcagcaaaaaaggcaacaatttagataaccagcgtttggac	0.99	0.99
398	atatgaagaccacacaagtctgaattgtgtttgaaaccggtgctaaaac	0.97	0.98
399	agaaattaagagccaacctaaagaaatgcgatgggcgttgctatcatac	0.96	0.96
400	ggtgttacatgaaaacctagataaccgtctaatagattcataaaata	1.00	0.99
401	tgtaattacagaaacgatctgtggacgcatatgccaatatgcgattc	0.98	0.98
402	ggaaataaccaggattactttgatatgactttagctgtaccataac	1.00	0.99
403	aaagggctaaaatgagcaacaatagagacccccatttgcttcgtatt	0.99	0.99

404	atcacttccgggtgctgttatcagaatatggttaaattgtta	0.99	0.99
405	tggagatttaatagatatctccttgctattgcacatcgcaaa	0.99	0.99
406	agtatgcgaatactccaccctcacaaaaactaatcaatgcc	1.00	1.00
407	atacttaaatgttgatatttgcggcgggttaagtgcagac	0.88	0.92
408	catagtacattcttatgacaaaagctatgattgtaatttctc	1.00	0.96
409	agcgaagcagcactaagaagcttgaatgcgtggatgcataacc	0.76	0.83
410	ttgagaaaagtcgttatgcaaacggagtcccattatctattg	0.98	0.92
411	cactattcggccacaacttaacgtacacattacgcagaaggggt	0.98	0.98
412	ccggttcataacatatacagatacgaataggtactctca	1.00	0.99
413	tcacgatttccgaaaaaacgtatcccaaaagagaggg	0.98	0.99
414	tttggaagtaacagatctgaatcccgaatccactatg	1.00	0.99
415	cccataaatatatactgcgaagttaaacttaaaagg	1.00	1.00
416	ctaacaatgattgacgaggtcgcaatgagtaaate	0.99	0.99
417	cgtaactagacatatttgggttaattcgttacta	1.00	1.00
418	tgatgcttcatagcagtgtttaggttgaccgtatgactcaatgtaatttagccaatcttcaattca	0.92	0.94
419	acgcaaaaacgttctagtagtatagctaaaccaataatataattaaataatgttcgacaacatcaacaaaa	1.00	0.98
420	taagacctactaggaaaacgacaaaatcttacaacggttccaccatccaacacatgactctcttagag	0.99	1.00
421	ggfccgagggataaaagactgatagatcctcaataagataatggtaaaagactgagggtaatcct	0.83	0.87
422	attgaatttcacacactttacagagtagataataatctccgcacagctattcgtgtgaatggctctat	0.96	0.93
423	caggagtcgccctaattgagctaacgaaccaggacattcgcgcaacgcacatcattattcatgaat	0.71	0.77
424	atgattgggttttatgacgggtatactatgaccacaactaaatgtcttctgtaggtgttcccca	0.87	0.83
425	atcaaatagagagagcggaaaaacttgatctaatcagggcatgaaagggtttttatcgtcta	0.97	0.94
426	ttacattaaatgaaaacctctagtaataagtatctgtcgatcgaaaccggtaagtaaaatg	1.00	0.99
427	aaacgggtctactggatgtagtactaattgtatatgaagaaaaccactaaatttatgccag	0.99	0.99
428	gaaggttaaagtcgaattgagcattcctgatttttagagcatttagcctcaaggtaataat	0.99	0.99
429	cgcagtgaagagtaggtattcagtacgcgacaccggaccgtaagcatacaattatagatt	0.61	0.70
430	gctcgcttacatactgcagcctaataacaacacgtatctagtccatagttattgtggta	0.93	0.85
431	cttgaatattgacttaacagttacaatataagaattaactgacgaagagtgccgcccgtg	0.96	0.95
432	cccaaaagagatcgccaatgccaacataccatgataactttatacaacctgatcgg	0.99	0.98
433	tgtcctgatctctaccggattggactagactagtaaatattgaaatacattgggtgaat	0.92	0.94
434	gtaacgtgcttaagcaattaccacaactctcatggatagtagatagattgtaagt	0.99	0.97
435	acattagctgcaaacctctagcgtacaagccttctgtagcagtagcagcataaa	0.96	0.97
436	ttagcgtgggaaacagttgatataagccgcactaactaagcgcaaatgtgctgaa	0.91	0.93
437	tatcctgcataattactttctagattatacatattgtatcatataggtgcagct	1.00	0.97
438	acctcgtgttattcggagtatacaagttcccggtttctgtgaagtagatttatt	0.67	0.75
439	agtataaatctccatatacacgcattcgtcatgatttatagccccaca	1.00	0.91
440	cttttttaactagctgaaaatgtattcacttaacatagacgtaagta	1.00	1.00
441	tatcttctccctagctcgatcaciaaactgactcgtgggtgttcatccagcg	0.58	0.69
442	agtcagcggacattgggccttatgaatctgttaattctagttttata	0.94	0.84
443	tcgaactgcaacaaaattattcacacaatattacctgttccagct	1.00	0.98
444	tttgctaatcatccagcaacaagtttaccataatgtaacgcg	1.00	1.00

445	tacaccataactaatatcaccecgacagactcgacctcaactcgg	1.00	1.00
446	aaatactagcattatctcactccacatttcctgtgtaactat	1.00	1.00
447	tacatggtaaagtatcgcggctcgtagtgtacacggtaa	0.55	0.68
448	tttatagtcagcagattgtttgcttatatatgtatgac	0.99	0.87
449	acatacgaataatgtcatcttgtctgtaaggacagatt	1.00	0.99
450	ttaatcagcggcggacgtaattgagtagccttca	0.75	0.82

---

\* In this work we used a classical representation of a nucleic acid sequences. Please note that there are only four letters “a, t, g, c” for a DNA sequence. The letters represent different classes of nucleic acid bases.