SciForum
Mol2Net

Computational Study of Mycobacterial Promoters with Low Sequence Homology

Alcides Pérez-Bello
Veterinary Medicine Department, Central University of ‘Las Villas’, 54830, Cuba. E-Mail: alcidopb@yahoo.com.

Received: 29 September 2015 / Accepted: 29 September 2015 / Published: 2 December 2015

Abstract: This communication shows a classification model for prediction of 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 model predicted 115/135 mps (85.2%) and 100% of control sequences (cs). The detailed results have been published in detail in: Bioorg Med Chem Lett. 2006 Feb;16(3):547-53, the present is a short communications.

Keywords: Mycobacteria; Protein synthesis promoters; RNA secondary structure

1. Introduction

Harshey and Ramkrishnan stated that *Mycobacteria* have a low transcription rate and a low RNA content per unit DNA and that their genomes are rich in Guanine and Cytosine (g + c) content. Given that the g + c content of a genome affects the codon usage and the promoter recognition sites in an organism, Nakayama *et al.*, and Ohama *et al.* predicted 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. Different mathematical methods have been used for the analysis of genome information. The group of Professor Grau has reported results on genome algebras. Markov models 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 from the point of view of the corresponding RNA structure.

A real possibility to address this problem involves the analysis of structure-activity relationships for naturally occurring RNA macromolecules, synthetic polymers and small molecules in general with Markov molecular descriptors. For this reason, one may expect higher success for classical molecular indices in branched biomacromolecules. However, it must be remembered that the more commonly known branched biomacromolecule is the RNA secondary structure as described by Mathews and Zukker.

Researchers worldwide have reported increasing interest in the characterization of biomacromolecules, particularly the RNA macromolecular structure, by computational techniques. 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 and these descriptors have been largely applied to small molecules and biomacromolecules. Two preliminary studies into secondary QSAR of RNA macromolecules have also been published, but these 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. 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 the macromolecular secondary structure of its putative RNA. 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 the first 2D-RNA-QSAR to discriminate between two groups comprising several RNA macromolecules, including 135 mycobacterial promoters and 450 control sequences.

2. Results and Discussion

Several authors have studied the mycobacterial promoter sequence problem from the point of view of DNA. Mulder *et al.* listed −35 and −10 DNA regions of a few mycobacterial promoters. *Mycobacteriophage I3* and *M. paratuberculosis* promoter sequences and their similarity with the *E. coli* promoters have been studied by Ramesh and Gopinathan and Bannantine *et al.*, respectively. Kremer *et al.* 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[52] are the true representatives of the mycobacterial type. An analysis of *M. smegmatis* and *M. tuberculosis* promoters by Bashyam *et al.* showed that there are similarities to *E. coli* 70 promoters; however, in this case the −35 regions showed greater sequence variability. Strohl studied DNA promoter sequences for *Streptomyces* promoters.

O'Neill and Chiafari have also made efforts to develop statistical algorithms for sequence analysis and motif prediction by searching for homologous regions or by comparing the sequence information with a consensus sequence. Two studies by Mulligan and McClure and Mulligan *et al.* pointed out that the variations that exist within individual promoter sequences are primarily responsible for the unsatisfactory results yielded by the promoter-site-searching algorithms, which in essence perform statistical analysis. It can therefore be inferred that recognition of mycobacterial promoter sequences requires a powerful technique that is capable of unravelling those hidden pattern(s) in the biomacromolecule structure – patterns that are difficult to identify visually.

Linear Discriminant Analysis was used to classify RNA macromolecules as mycobacterial promoter sequence (mps) or control group sequence (cs). In the development of the LDA the output was a dummy variable, mps, which codifies whether a sequence lies within the mps class (mps = 1) or belongs to the cs group (mps = 0). In this problem the inputs were the Markov electrostatic potentials (${}^{k}\xi_{M}$) of interaction between nucleotides located with respect to each other at a topologic distance k within the 2D-RNA backbone, with *k* it is in the range [0, 5]. The ${}^{k}\xi_{M}$ are parameters derived by means of a Markov chain model and are used here as molecular descriptors to encode RNA secondary structure (see methods section for details). The best discriminat equation found to discriminate between mps and the control group was:

$$mps = -4.664 \cdot {}^{0}\xi_{M} + 0.991 \cdot {}^{1}\xi_{M} - 2.432 \qquad (1)$$

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

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) <0.001. 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) was observed and this high C value indicates a strong linear relationship between the structural descriptors of the biomacromolecules 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 original work). Conversely, the second order potential ${}^{2}\xi_{M}$ does not have a significant relationship with the mps characteristic or RNA sequences. In physical terms the above results show that, as in other studies, 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 on considering isolated nucleotides (${}^{0}\xi_{M}$). Conversely, the variations of global electrostatic potential (${}^{1}\xi_{M}$) due to secondary structure folding[65] as a result of direct covalent and/or hydrogen bonds between nucleotides increase by a factor of only 0.991 with respect to the possibility of RNA being encoded as an mps. Finally, long-term electrostatic interaction potentials between nucleotides at distances longer than 1 (${}^{2}\xi_{M}$, ${}^{3}\xi_{M}$, ${}^{4}\xi_{M}$) do not correlate with the mps activity. The detailed results of the forward stepwise analysis are given in the original work.

Jack-knife cross validation (cv) experiments were performed by the re-substitution technique, leaving out four different groups selected at random and containing 25% of the RNA molecules. The cross validation accuracies and the average cross validation accuracy (cv-average) were cv1 = 95.9%, cv2 = 96.6%, cv3 = 96.6% and cv4 = 96.5%, respectively, with the average Cv-average = 85.7.

The testing of the model fit to data and its robustness – although very important – is not the only characteristic of an acceptable QSAR.

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 of the original work. Finally, as far as the quality of the model is concerned, we would like to point out that the present linear QSAR model compares very favourably to a previous non-linear model reported by Kalate *et al.* in terms of simplicity (two variables: ${}^{0}\xi_{M}$ and ${}^{1}\xi_{M}$). This non-linear model presented only slightly higher accuracy (97%) but makes use of very large space parameters to describe DNA sequences rather than RNA structure. The success of our RNA-QSAR model, which uses only two variables, can be explained by considering that RNA structure molecular descriptors encode not only sequences (as is the case for DNA linear sequence descriptors) but also molecular branching.

The present paper introduces the simplest up-to-date reported method to predict mycobacterial promoters. With this ultimate aim in mind, we changed the classical point of view and used RNA 2D-macromolecular descriptors instead of DNA sequence analysis. In this sense, this work opens a new way for the application of classical QSAR approaches to biomacromolecules.

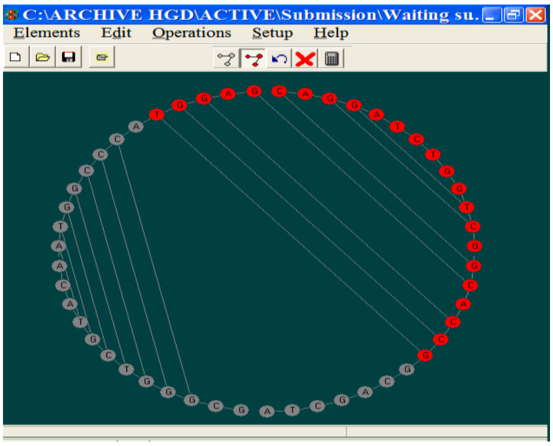

Figure 1. Circular representation for a folded RNA macromolecule of mps T3 from *M. tuberculosis*, note main stem highlighted in red.

4. Conclusions

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 (${}^{0}\xi_{M}$, ${}^{1}\xi_{M}$).

Conflicts of Interest

State any potential conflicts of interest here or “The authors declare no conflict of interest”.

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 Acid 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. 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.
14. Ruan, J.; Stormo, G.D.; Zhang, W. *Bioinformatics* **2004**, *20*, 58.
15. Ieong, S., Kao, M.-Y, Lam, T.-W., Sung, W.-K. and Yiu, S.-M. *J. Comp. Biol.* **2003**, *10*, 981–995.
16. González-Díaz, H.; Molina, R.; Uriarte, E. *Polymer* **2004**, *45*, 3845.
17. González-Díaz, H.; Molina, R.; Uriarte. E. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4691.
18. 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.
19. 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.
20. González-Díaz, H.; Hernández, S.I.; Uriarte, E.; Santana, L. *Comput. Biol. Chem.* **2003**, *27*, 217.
21. 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.
22. González-Díaz, H.; Uriarte, E.; Ramos de A, R. *Bioorg. Med. Chem*. **2005**, *13*, 323.
23. Gia, O.; Marciani-Magno, S.; González-Díaz, H.; Quezada, E.; Santana, L.; Uriarte, E.; DallaVia, L. *Bioorg. Med. Chem*. **2005**,*13*, 809.
24. Ramos de A, R.; González-Díaz, H.; Molina, R.; Uriarte, E. *Prot. Struct. Func. Bioinf.* **2004**, *56*, 715.
25. 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.
26. González-Díaz, H.; Ramos de A, R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
27. González-Díaz, H.; Ramos de A, R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
28. Saíz-Urra, L.; González-Díaz, H.; Uriarte, E. *Bioorg. Med. Chem.* **2005**, *13*, 3641.
29. Norberg, J.; Nilsson, L. *Acc. Chem. Res.* **2002**, *35*, 465.
30. González-Díaz, H.; Molina, R.; Sanchez, I. BIOMARKS ©, **2004**, *version 1.0*.
31. Mathews, D.H.; Zuker, M.; Turner, D.H. RNAStructure ©, **2002**, *version 4.0*.
32. Marrero-Ponce, Y.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F.; Castro, E.A. *Bioorg. Med. Chem.* **2004**, *12*, 5331.
33. Marrero-Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010.
34. Marrero-Ponce, Y.; Montero-Torres, A.; Romero-Zaldivar, C.; Iyarreta-Veitía, M.; Mayón-Peréz, M.; García-Sánchez, R. *Bioorg. Med. Chem.* **2005**, *13*, 1293.
35. González-Díaz, H.; Cruz-Monteagudo, M.; Molina, R.; Tenorio, E.; Uriarte, E. *Bioorg. Med. Chem*. **2005**, *13*, 1119.
36. 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.
37. Mulder, M.A.; Zappe, H.; Steyn, L.M., *Tuber. Lung Dis.* **1997**, *78*, 211.
38. Ramesh, G.; Gopinathan, K.P., *Indian J. Biochem. Biophys.* **1995**, *32*, 361.
39. Bannantine, J.P.; Barletta, R.G.; Thoen, C.O.; Andrews, R.E., Jr., *Microbiol.* **1997**, *143*, 921.

40. Kremer, L.; Baulard, A.; Estaquier, J.; Content, J., Capron, A.; Locht, C. *J. Bacteriol.* **1995**, *177*, 642.