Available online at www.sciencedirect.com

ScienceDirect

Bioorganic & Medicinal Chemistry 15 (2007) 2544–2550

Bioorganic & Medicinal Chemistry

On the applicability of QSAR for recognition of miRNA bioorganic structures at early stages of organism and cell development: Embryo and stem cells

Humberto González-Díaz,[a,b,*] Santiago Vilar,[a,c] Lourdes Santana,[a]
Gianni Podda[b] and Eugenio Uriarte[a]

[a]*Faculty of Pharmacy, University of Santiago de Compostela, Santiago de Compostela 15782, Spain*
[b]*Dipartimento Farmaco Chimico Tecnologico, Universita' Degli Studi di Cagliari, Cagliari 09124, Italy*
[c]*Department of Pharmaceutical Sciences, University of Padova, I-35131 Padova, Italy*

Received 10 October 2006; revised 24 January 2007; accepted 31 January 2007
Available online 2 February 2007

Abstract—Quantitative structure–activity-relationship (QSAR) models have application in bioorganic chemistry mainly to the study of small sized molecules while applications to biopolymers remain not very developed. MicroRNAs (miRNAs), which are non-coding small RNAs, regulate a variety of biological processes and constitute good candidates to scale up the application of QSAR to biopolymers. The propensity of a small RNA sequence to act as miRNA depends on its secondary structure, which one can explain in terms of folding thermodynamic parameters. Then, thermodynamic QSAR can be used, for instance, for fast identification of miRNAs at early stages of development such as embryos and stem cells (called here esmiRNAs), and gain clarity inside cellular differentiation processes and diseases such as cancer. First, we calculated folding free energies (ΔG), enthalpies (ΔH), and entropies (ΔS) as well as melting temperatures (T_m) for 2623 small RNA sequences (including 623 esmiRNAs and 2000 negative control sequences). Next, we seek a QSAR classification model: esmiRNA = $0.035 \times T_m - 0.078 \times \Delta S - 8.748$. The model correctly recognized 543 (87.2%) of esmiRNAs and 935 (93.5%) of non-esmiRNAs divided into both training and validation series. The model also recognized 908 out of 1000 additional negative control sequences. ROC curve analysis (area = 0.93) demonstrated that the present model significantly differentiates from a random classifier. In addition, we map the influence of thermodynamic parameters over esmiRNA activity. Last, a double ordinate Cartesian plot of cross-validated residuals (first ordinate), standard residuals (second ordinate), and leverages (abscissa) defined the domain of applicability of the model as a squared area within ±2 band for residuals and a leverage threshold of $h = 0.0074$. The present is the first QSAR model for quickly accurate selection of new esmiRNAs with potential use in bioorganic and medicinal chemistry.

1. Introduction

MicroRNAs (miRNAs) are small non-coding bioorganic molecules with a broad spectrum of functions described mostly in invertebrates. They act as post-transcriptional regulators of gene expression, miRNAs trigger target mRNA degradation or translational repression. A role of miRNA has been described in hematopoietic, adipocytic, neurogenesis, muscle differentiation, regulation of insulin secretion, plant and fungi biology, and potentially regulation of cancer growth.[1–3] A key step toward understanding the function of the hundreds of miRNAs identified in animals is to determine their expression during early stages of animal development, which has been barely studied.[4] In this sense, it is important to identify new species of miRNA and obtain a comprehensive quantitative profile of small RNA expression in developed and stem embryo cells. With this aim, other authors used experimental methods that potentially identify virtually all of the small RNAs in a sample. This approach allowed them to detect 390 miRNAs, including 195 known miRNAs covering approximately 80% of previously registered mouse miRNAs as well as 195 new miRNAs, which are so far unknown in mouse embryos. Some of these miRNAs showed temporal expression profiles during prenatal development. These results indicate existence of a

Keywords: QSAR; RNA structure; miRNA prediction; RNA folding thermodynamics; Mouse embryo transcripts; Stem cells; Cancer.

* Corresponding author. Tel.: +34 981 563100; fax: +34 981 594912; e-mail addresses: gonzalezdiazh@yahoo.es; qohumbe@usc.es

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.01.050

significant number of new miRNAs expressed at specific stages of mammalian embryonic development and which were not detected by earlier methods.[5] Other authors have studied the expression of miRNAs in embryo stem cells, which constitute a previous step in cell development (stem cells are undifferentiated cells). These authors developed a new experimental method to determine more than 200 different miRNAs in embryo stem cells.[6] In these cells, miRNAs play an important role in the maintenance of stemness in addition to cell cycle regulation, apoptosis, cell differentiation, and imprinting. Concordant with this, aberrant expression of miRNA genes could lead to human disease, including cancer. Although the connection of miRNAs with cancer has been suspected for several years, recent studies have confirmed the suspicion that miRNAs regulate cell proliferation and apoptosis, and play a role in cancer.[7–14] Consequently, it is of major importance developing new timely methods for the prediction of miRNAs in general and, due to its great importance, becomes of special relevance the development of quick methods for the accurate prediction of miRNAs at early stages of organism and cell development (coined herein as esmiRNAs).[15–18]

In our opinion, we can apply QSAR techniques[19] classically used for small-to-medium sized bioorganic molecules for prediction of miRNAs. In fact, other authors have previously used QSAR to predict the properties of biopolymers including proteins and RNAs.[20–23] To fulfill this aim, we have first to calculate miRNA structural parameters (molecular descriptors) expected to be correlated with miRNA action and later apply statistical analysis to seek the miRNAs-QSAR model as classic researchers used to do.[24–26] In principle, there are many possible molecular descriptors used in classic QSAR, that we could select for this work.[27–32] The different classes of molecular descriptors include topological indices, quantum descriptors, constitutional parameters, and physicochemical magnitudes (including thermodynamic parameters).[33]

On the other hand, different methods to predict RNA bioorganic secondary structure have been reported, which deal with stochastic context free grammars,[34] dynamic treatment of constraints,[35] sequence alignment, and k-nearest neighbor networks.[36] However, still many researchers prefer theoretic physicochemical methods based on thermodynamic calculations.[37,38] We can state with certainty that biological process took place under the direct influence of natural laws such as the second law of thermodynamics. Consequently, thermodynamics becomes a biophysical method of wide application in biochemistry. Applications cover, for instance: drug–target interactions,[39] protein domain stability,[40] membrane proteins' folding,[41] energy landscape analysis of the genetic code,[42] and many others. Then, the propensity of a small RNA sequence to act as miRNA, which depends on its bioorganic secondary structure, follows also thermodynamics laws. Because of this, we can explain the folding of a RNA sequence into a bioorganic secondary structure in terms of thermodynamic parameters such as folding free energies (ΔG), enthalpies (ΔH), and entropies (ΔS) as well as melting temperatures (T_m). These parameters tell us about how hardly or not an RNA sequence manages to fold into a bioorganic secondary structure.[43,44] Thence, we can expect that among the many possible parameters to be selected thermodynamic molecular descriptors should be selected to seek a Quantitative-structure–activity-relationship (QSAR) for fast identification of miRNAs.

In this work, we report by the first time a thermodynamic QSAR study for two recently reported esmiRNA databases (one for mouse embryo mature cells and other for embryo stem cells).[5,6] First, we removed all the coincident miRNA sequences. Second, we calculated all the above-mentioned thermodynamic parameters for 623 non-coincident esmiRNA sequences listed by these authors and 2000 negative control sequences. Control sequences were above 22 bp-length and were generated at random avoiding folding similarity or clustering among them or with miRNAs. The use of random sequences as negative control group has been validated in the literature.[45,46] Next, we carried out a forward stepwise Linear Discriminant Analysis (LDA)[47,48] to find a linear QSAR based on thermodynamic parameters that discriminate between RNAs and other sequences. Previously, the data set was divided at random into training and validation series in order to train and validate the model, respectively. Next, we perform a ROC[49] curve analysis testing how significantly the model differentiates from a random classifier. In addition, we performed a 2D map desirability analysis[50] of the QSAR model to determine the influence of the thermodynamic folding landscapes over the probability to act as esmiRNA. However that our model focused only on esmiRNA from mouse mature embryos and stem cells we finished the work with a leverage-based analysis of the domain of applicability of the model.[51] The present result offers a QSAR thermodynamic basis for further design and selection of new esmiRNAs a gain in clarity on the role of the secondary bioorganic structure of miRNAs at early stages of organism development.

2. Results and discussion

2.1. Train and validation of the QSAR thermodynamic model for esmiRNAs prediction

Tremendous progress in DNA sequencing has yielded the genomes of a host of important organisms. The utilization of these resources requires understanding of the function of each gene. Standard methods of functional assignment involve sequence alignment to a gene of known function; however, such methods often fail to find any significant matches. Some authors have recently reviewed a number of recent alignment methods[52] and alternative methods that may be of use when sequence alignment fails.[53]

In special the study and function annotation of mouse gene transcripts has covered a high importance now a days with special emphasis on small transcripts.[54,55] Unfortunately, in the case of RNAs there are very less studied alternative parameters to seek QSAR models instead of using alignment procedures. Our group has reported some parameters such as spectral moments and entropies to connect RNA secondary structure with function.[48,56–58] Here, we describe possibly the first QSAR model that can assign esmiRNA function from secondary bioorganic structure thermodynamic parameters and without being reliant upon alignments.

This method follows the general strategy recently outlined by Bentwich for miRNA prediction laying emphasis on the use of thermodynamic parameters QSAR approach.[18] The method combines RNA folding thermodynamic parameters with statistical analysis to seek models like Eq. 1. One advantage of the method is the fast calculation of the parameters. We can use this QSAR equation to calculate the probability of large databases of small RNA sequences to act as esmiRNAs without carrying out any experimental measurements. Afterwards, only the sequences with the higher predicted probability can be experimentally validated to confirm the esmiRNA action. The data set was explored with the LDA method.[59–68] The best model found was the following:

$$\text{esmiRNA} = 0.035 \times T_m - 0.078 \times \Delta S - 8.748 \qquad (1)$$

$$R_c = 0.72; \quad \lambda = 0.48; \quad \chi^2 = 890.49; \quad p < 0.001.$$

where R_c is the canonical regression coefficient, λ is the Wilk's statistics, χ^2 is the Chi-sqr statistic, and p is the level of error. The procedure as any QSAR may represent important saving of time, material resources, and laboratory animals.[24,28,69,70] This model presents an overall accuracy of approximately 94% in both training and validation series. We depict detailed information for classification results in Table 1. This is exceptionally good value for this kind of simpler linear classifier (LDA) and using only two variables, for which values higher than 85% are accepted.[67,70,71]

A ROC curve analysis[72] developed to compare the present QSAR model with a random classifier showed an area for the ROC curve of 0.98 (see Fig. 1). Finally, Figure 2 depicts the results of a desirability analysis developed to determine the levels of the independent variables that ensure the more desirable values of the dependent variable (esmiRNA action).

Table 1. Training, validation, and overall (both) results for the QSAR classification model

Observed RNA class	Accuracy or predictability[a] (%)	Non-esmiRNA[a]	esmiRNA[a]
Training (total accuracy 91.54)			
non-esmiRNA[b]	93.9	**704**	46
esmiRNA[b]	87.8	57	**411**
Validation (total predictability 89.63)			
non-esmiRNA[b]	92.4	**231**	19
esmiRNA[b]	85.2	23	**132**
Both (total 91.07)			
non-esmiRNA[b]	93.5	**935**	65
esmiRNA[b]	87.2	80	**543**

Classification matrices: [b]observed values in rows and [a]predicted ones in columns.

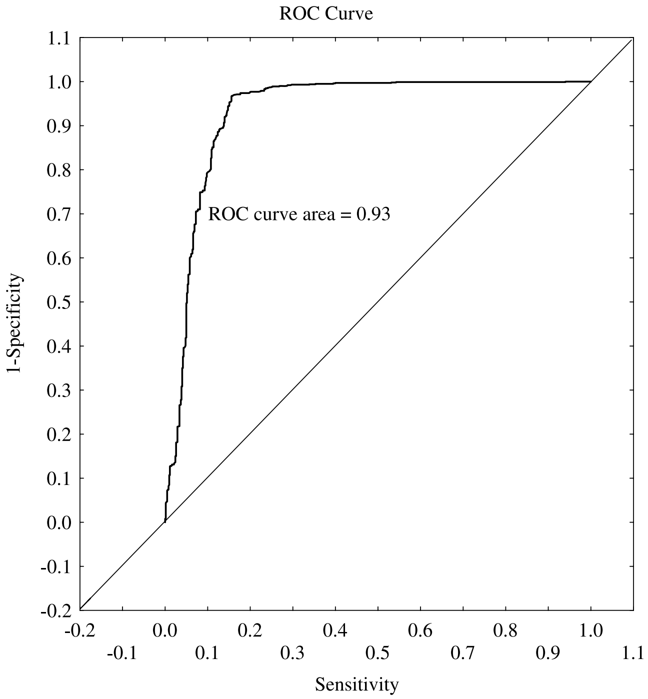

Figure 1. ROC curve analysis for the model developed for esmiRNA recognition.

2.2. Desirability analysis for thermodynamic QSAR of the esmiRNA action

Due to the large amount of information to be processed in biopolymer QSAR studies, in previous works we have used parameters with complicated interpretation in physicochemical terms.[73,74] An advantage of the present procedure is that the parameters investigated: ΔG, ΔH, ΔS, and T_m are well known thermodynamic magnitudes. These parameters are supported on physicochemical basis increasing model rigor in theoretical terms.[38,75] In particular, in the model (1) reported herein for esmiRNA prediction only ΔS and T_m presented a significant contribution. In general, miRNA activity including esmiRNA activity is intimately linked to RNA folding. Consequently, it sounds strange that ΔG and ΔH, which are very important RNA folding parameters' apparently do not contribute to the esmiRNA action. Nevertheless, we should not forget that by definition $T_m = 1000 \times (\Delta H/\Delta S)$. Consequently, rewriting Eq. 1 one can estimate a negative direct effect of ΔH increasing over esmiRNA action:

$$\text{esmiRNA} = 35.0 \times \frac{\Delta H}{\Delta S} - 0.078 \times \Delta S - 8.748 \qquad (2)$$

Notably, the high contribution of T_m presupposes a $\Delta G = 0$, which is the condition to calculate T_m from $\Delta G = \Delta H - T \times \Delta S$. It points to an interesting conclusion; esmiRNA activity may be increased for miRNAs with high values of T_m and consequently ΔH (note positive contribution in the equation) when miRNA struc-

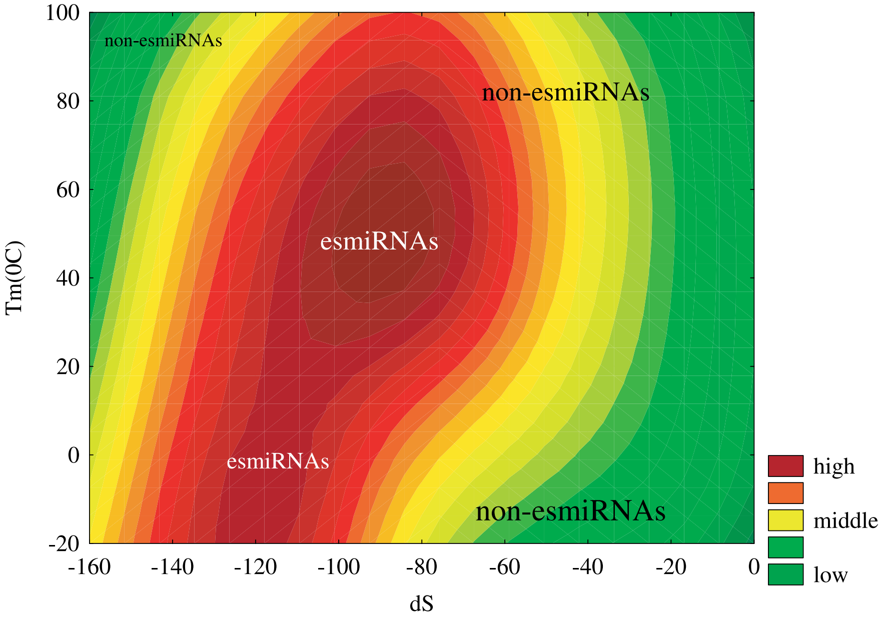

Figure 2. Thermodynamic folding landscape desirability analysis for esmiRNAs.

tures are in a folding state of thermodynamic equilibrium ($\Delta G = 0$). This result coincides with previous findings on the importance of thermodynamics to detect miRNAs in other species.[76] In contrast with transfer RNAs and ribosomal RNAs, the majority of the miRNAs clearly exhibit a high tendency in the sequence toward a stable secondary structure.[46]

Certainly, detailed interpretation of the relationship between folding thermodynamic and esmiRNA action may be complicated and somehow outside the scope of this work. In any case, we can answer one of the more important questions connected with the practical use of this thermodynamic model. What levels of the independent thermodynamic parameters (T_m and ΔS) ensure the esmiRNA activity for a small RNA molecule? This question is of major importance for the future application of the method in the design of esmiRNAs. We reach this aim using a desirability analysis. Desirability analysis has been successfully used recently with the similar aim for protein sequences–function relationships.[73] As depicted in Figure 2, small RNAs that lie in the desirability 2D map within the dark red area have a high propensity to act as esmiRNAs. Other areas of the 2D map for thermodynamic parameters' landscape have presupposed weak or negative contribution to esmiRNA activity.

2.3. Brief note on the domain of applicability of the model

The interest in QSAR has steadily increased in recent decades. It is generally acknowledged that these empirical relationships are valid only within the same domain for which they were developed. However, model validation is sometimes neglected, and the application domain is not always well-defined.[77] The purpose of this section is to outline how validation and domain definition determines in which situation it is correct to use the model. The aim of the present work was to develop a model for predicting miRNA at early stages of organism and cell development due to the high importance of these miRNAs for different aspects including cell differentiation, stemness, and cancer. In consonance, we selected mouse embryo miRNA and embryo stem cell miRNAs. Consequently, one may not pretend to extrapolate the use of this model to other species or tissues making uncertain predictions in conditions very different to those fixed to derive the model.[31] More in detail, a double ordinate Cartesian plot of cross-validated residuals (first ordinate), standard residuals (second ordinate), and leverages (abscissa) defined the domain of applicability of the model as a squared area within ±2 band for residuals and a leverage threshold of $h = 0.0074$. As can be noted in Figure 3, almost all miRNAs and negative control group sequences used in training and validation lie within this area. Actually, some sequences have a leverage higher than the threshold but show leave-one out (LOO) residuals, cross-validation, and standard residuals within the limits. In closing, no apparent outliers were detected and the model can be used with high accuracy in this applicability domain.[31,78,79]

3. Conclusions

We demonstrate that the annotation of and small RNA sequence as esmiRNAs is a linear QSAR function of the thermodynamic parameters ΔS and T_m. The values of the input variables that ensure the desire property were given. The model presents a high accuracy, clearly differentiates from a random classifier, and a very well-defined domain of applicability. The present work reports the first QSAR model to predict esmiRNAs given a predicted RNA secondary structure.

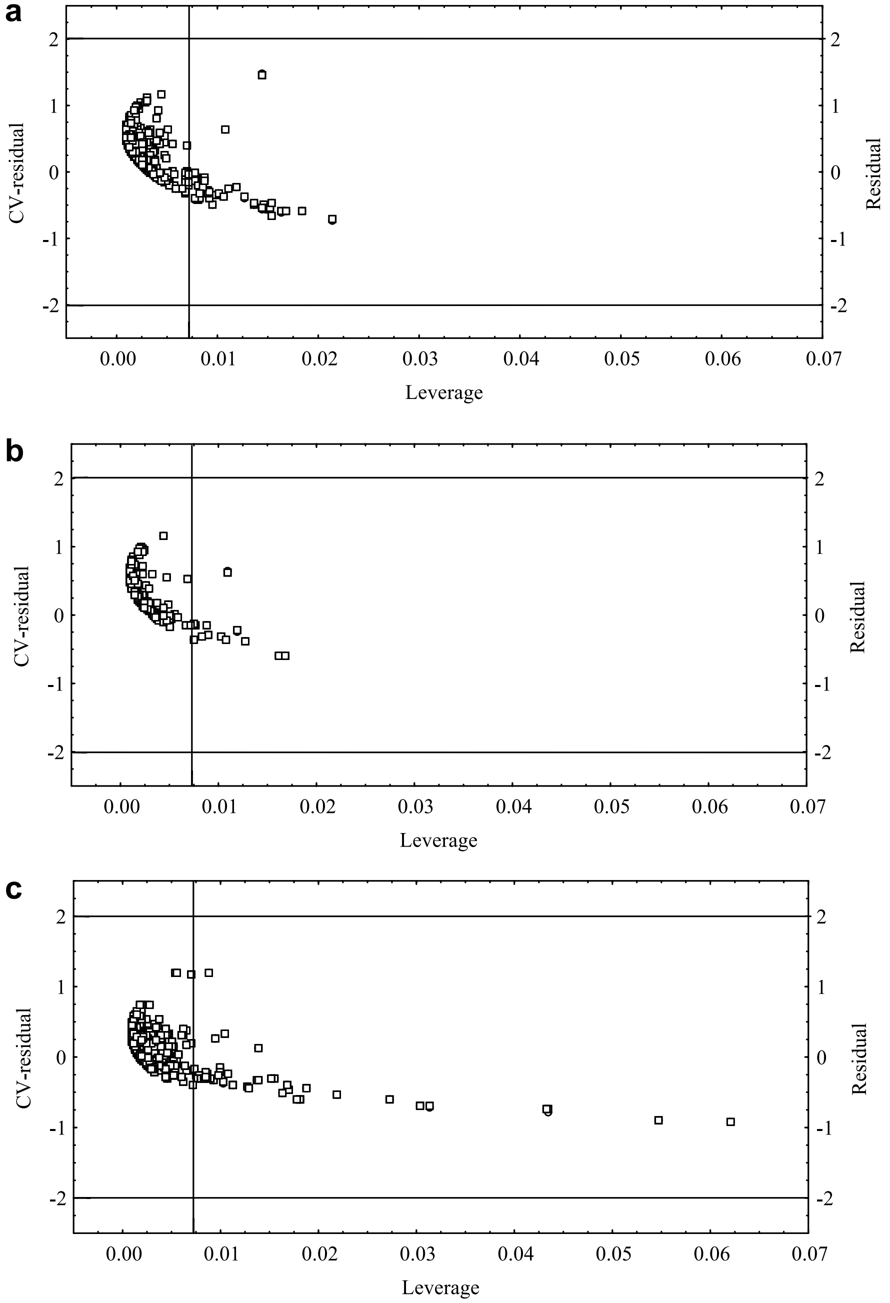

Figure 3. (a) esmiRNA embryo (b) esmiRNA stem cell (c) RNA negative control group, model leverage threshold was $h = 0.0074$.

4. Methods

4.1. Calculation of thermodynamic parameters (ΔG, ΔH, ΔS and T_m) used to seek the QSAR model

First, the sequences of the 623 esmiRNAs were withdrawn from two different public sources.[5,6] Later, we removed all the coincident miRNA sequences. Second, we generated at random negative control sequences with above 22 bp-length avoiding folding similarity or clustering among them or with miRNAs. Next, the list of all sequences was introduced as input of the RNA secondary structure prediction server DINAMelt. Specifically, we used the fast fold option named Quickfold (http://www.bioinfo.rpi.edu/applications/hybrid/quikfold.php). This server calculated the secondary structure of all the sequences as well as the folding thermodynamic parameters ΔG, ΔH, ΔS, and T_m. We run the job 060429_194550. Job parameters were: energy rule RNA (2:3), at 37 °C, $[Na^+] = 1$ M, $[Mg^{2+}] = 0$ M, sequence type linear, structures 5% suboptimal, window size as default, maximum of 1 folding, and no limit to maximum distance between paired bases. At these conditions the computation took only a few seconds.[80]

4.2. Derivation of the QSAR model

The database containing the ΔG, ΔH, ΔS, and T_m values for all the sequences was the starting point for the analysis. Prior to analysis, we inserted a dummy variable esmiRNA, being esmiRNA = 1 for an esmiRNA sequence and esmiRNA = 0 for a non-esmiRNA sequence. We used a LDA to derive the classifier. We split the whole data set into two series selected at random. The larger series named training series contains the 75% of all the sequences, the remnant sequences conforming to validation series. A third additional series not used for training or validation and containing 1000 additional negative control sequence was presented too. We used the LDA option of the software STATISTICA[81] for the statistical analysis, selecting forward stepwise as the variable selection technique, and estimating the ROC curve point by point. The desirability analysis was derived with the profiler section of the LDA option. A double ordinate Cartesian plot of cross-validated residuals (first ordinate), standard residuals (second ordinate), and leverages (abscissa) defined the domain of applicability of the model.[31,78,79]

Acknowledgments

The authors thank the Xunta de Galicia (projects PXIB20304PR and BTF20302PR) and the Ministerio de Sanidad y Consumo (project PI061457) for partial financial support. Both, González-Díaz, H. and Vilar, S. acknowledge scholarship funding from the Xunta de Galicia for a one-year post-doctoral position (2007) in the Dipartimento Farmaco Chimico Tecnologico of the University of Cagliari and Department of Pharmaceutical Sciences, University of Padova, Italy, respectively. This author also acknowledges two contracts as guest professor in the Department of Organic Chemistry, Faculty of Pharmacy, University of Santiago de Compostela, Spain, in 2006 and 2007. Thanks are also given to the editor Prof. H. Waldman and two unknown referees.

Supplementary data

The supplementary material contains one table, which depicts the code, sequence, observed class, residual, deleted or LOO residual, leverage, and predicted probability as esmiRNA, and training or validation designation for each of the 2623 sequences studied. This table also contains the folding entropies (ΔS) as well as melting temperatures calculated for each of these sequences. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2007.01.050.

References and notes

1. Krichevsky, A. M.; Sonntag, K. C.; Isacson, O.; Kosik, K. S. *Stem Cells* **2006**, *24*, 857.
2. Chen, X. *FEBS Lett.* **2005**, *579*, 5923.
3. Nakayashiki, H. *FEBS Lett.* **2005**, *579*, 5950.
4. Kloosterman, W. P.; Wienholds, E.; de Bruijn, E.; Kauppinen, S.; Plasterk, R. H. *Nat. Methods* **2006**, *3*, 27.
5. Mineno, J.; Okamoto, S.; Ando, T.; Sato, M.; Chono, H.; Izu, H.; Takayama, M.; Asada, K.; Mirochnitchenko, O.; Inouye, M.; Kato, I. *Nucleic Acids Res.* **2006**, *34*, 1765.
6. Tang, F.; Hajkova, P.; Barton, S. C.; Lao, K.; Surani, M. A. *Nucleic Acids Res.* **2006**, *34*, e9.
7. Yanaihara, N.; Caplen, N.; Bowman, E.; Seike, M.; Kumamoto, K.; Yi, M.; Stephens, R. M.; Okamoto, A.; Yokota, J.; Tanaka, T.; Calin, G. A.; Liu, C. G.; Croce, C. M.; Harris, C. C. *Cancer Cell* **2006**, *9*, 189.
8. Hammond, S. M. *Curr. Opin. Genet. Dev.* **2006**, *16*, 4.
9. Croce, C. M.; Calin, G. A. *Cell* **2005**, *122*, 6.
10. Hatfield, S. D.; Shcherbata, H. R.; Fischer, K. A.; Nakahara, K.; Carthew, R. W.; Ruohola-Baker, H. *Nature* **2005**, *435*, 974.
11. Houbaviy, H. B.; Murray, M. F.; Sharp, P. A. *Dev. Cell* **2003**, *5*, 351.
12. Zhang, B.; Pan, X.; Anderson, T. A. *J. Cell Physiol.* **2006**, *209*, 266.
13. Cummins, J. M.; He, Y.; Leary, R. J.; Pagliarini, R.; Diaz, L. A., Jr.; Sjoblom, T.; Barad, O.; Bentwich, Z.; Szafranska, A. E.; Labourier, E.; Raymond, C. K.; Roberts, B. S.; Juhl, H.; Kinzler, K. W.; Vogelstein, B.; Velculescu, V. E. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 3687.
14. Volinia, S.; Calin, G. A.; Liu, C. G.; Ambs, S.; Cimmino, A.; Petrocca, F.; Visone, R.; Iorio, M.; Roldo, C.; Ferracin, M.; Prueitt, R. L.; Yanaihara, N.; Lanza, G.; Scarpa, A.; Vecchione, A.; Negrini, M.; Harris, C. C.; Croce, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 2257.
15. Wheeler, G.; Ntounia-Fousara, S.; Granda, B.; Rathjen, T.; Dalmay, T. *FEBS Lett.* **2006**, *580*, 2195.
16. Enright, A. J.; John, B.; Gaul, U.; Tuschl, T.; Sander, C.; Marks, D. S. *Genome Biol.* **2003**, *5*, R1.
17. Adai, A.; Johnson, C.; Mlotshwa, S.; Archer-Evans, S.; Manocha, V.; Vance, V.; Sundaresan, V. *Genome Res.* **2005**, *15*, 78.
18. Bentwich, I. *FEBS Lett.* **2005**, *579*, 5904.
19. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F. *Bioorg. Med. Chem.* **2006**, *14*, 2398.
20. Caballero, J.; Fernñandez, L.; Abreu, J. I.; Fernández, M *J. Chem. Inf. Model.* **2006**, *46*, 1255.
21. Marrero Ponce, Y.; Castillo Garit, J. A.; Nodarse, D. *Bioorg. Med. Chem.* **2005**, *13*, 3397.
22. Marrero-Ponce, Y.; Nodarse, D.; González-Díaz, H.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Int. J. Mol. Sci.* **2004**, *5*, 276.
23. Marrero-Ponce, Y.; Medina-Marrero, R.; Castro, A. E.; Ramos de Armas, R.; González-Díaz, H.; Romero-Zaldivar, V.; Torrens, F. *Molecules* **2004**, *9*, 1124.
24. Caballero, J.; Fernandez, M. *J. Mol. Model.* **2006**, *12*, 168.
25. Fernandez, M.; Caballero, J.; Helguera, A. M.; Castro, E. A.; Gonzalez, M. P. *Bioorg. Med. Chem.* **2005**, *13*, 3269.
26. Fernandez, M.; Caballero, J.; Tundidor-Camba, A. *Bioorg. Med. Chem.* **2006**, *14*, 4137.
27. Duchowicz, P. R.; Fernandez, M.; Caballero, J.; Castro, E. A.; Fernandez, F. M. *Bioorg. Med. Chem.* **2006**, *14*, 5876.
28. Perez Gonzalez, M.; Morales Helguera, A. *J. Comput. Aided Mol. Des.* **2003**, *17*, 665.
29. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Besada, P.; Gonzalez-Moa, M. J. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3491.
30. Prabhakar, Y. S.; Rawal, R. K.; Gupta, M. K.; Solomon, V. R.; Katti, S. B. *Comb. Chem. High Throughput Screen.* **2005**, *8*, 431.
31. Papa, E.; Villa, F.; Gramatica, P. *J. Chem. Inf. Model.* **2005**, *45*, 1256.
32. Caballero, J.; Fernandez, M. *J. Mol. Model. (Online)* **2006**, *12*, 168.
33. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH, 2002.

34. Knudsen, B.; Hein, J. *Bioinformatics* **1999**, *15*, 446.
35. Gaspin, C.; Westhof, E. *J. Mol. Biol.* **1995**, *254*, 163.
36. Bindewald, E.; Shapiro, B. A. *RNA* **2006**, *12*, 342.
37. Mathews, D. H. *J. Mol. Biol.* **2006**, *359*, 526.
38. Mathews, D. H.; Sabina, J.; Zuker, M.; Turner, D. H. *J. Mol. Biol.* **1999**, *288*, 911.
39. Chaires, J. B. *Arch. Biochem. Biophys.* **2006**, *453*, 26.
40. Nagy, M.; Akoev, V.; Zolkiewski, M. *Arch. Biochem. Biophys.* **2006**, *453*, 61.
41. Minetti, C.; Remeta, D. P. *Arch. Biochem. Biophys.* **2006**, *453*, 30.
42. Klump, H. H. *Arch. Biochem. Biophys.* **2006**, *453*, 85.
43. Freier, S. M.; Kierzek, R.; Jaeger, J. A.; Sugimoto, N.; Caruthers, M. H.; Neilson, T.; Turner, D. H. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 9373.
44. Lu, G.; Hallett, M.; Pollock, S.; Thomas, D. *Nucleic Acids Res.* **2003**, *31*, 3755.
45. Altuvia, Y.; Landgraf, P.; Lithwick, G.; Elefant, N.; Pfeffer, S.; Aravin, A.; Brownstein, M. J.; Tuschl, T.; Margalit, H. *Nucleic Acids Res.* **2005**, *33*, 2697.
46. Bonnet, E.; Wuyts, J.; Rouze, P.; Van de Peer, Y. *Bioinformatics* **2004**, *20*, 2911.
47. Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. *FEBS Lett.* **2005**, *579*, 4297.
48. Gonzalez-Diaz, H.; de Armas, R. R.; Molina, R. *Bioinformatics* **2003**, *19*, 2079.
49. Madazli, R.; Kuseyrioglu, B.; Uzun, H.; Uludag, S.; Ocak, V. *Int. J. Gynaecol. Obstet.* **2005**, *89*, 251.
50. Gonzalez-Diaz, H.; Sanchez-Gonzalez, A.; Gonzalez-Diaz, Y. *J. Inorg. Biochem.* **2006**, *100*, 1290.
51. Hill, T.; Lewicki, P. *STATISTICS Methods and Applications*; Tulsa: StatSoft, 2006.
52. Dobson, P. D.; Cai, Y. D.; Stapley, B. J.; Doig, A. J. *Curr. Med. Chem.* **2004**, *11*, 2135.
53. Dobson, P. D.; Doig, A. J. *J. Mol. Biol.* **2005**, *345*, 187.
54. Okazaki, Y.; Furuno, M.; Kasukawa, T.; Adachi, J.; Bono, H.; Kondo, S.; Nikaido, I.; Osato, N.; Saito, R.; Suzuki, H.; Yamanaka, I.; Kiyosawa, H.; Yagi, K.; Tomaru, Y.; Hasegawa, Y.; Nogami, A.; Schonbach, C.; Gojobori, T.; Baldarelli, R.; Hill, D. P.; Bult, C.; Hume, D. A.; Quackenbush, J.; Schriml, L. M.; Kanapin, A.; Matsuda, H.; Batalov, S.; Beisel, K. W.; Blake, J. A.; Bradt, D.; Brusic, V.; Chothia, C.; Corbani, L. E.; Cousins, S.; Dalla, E.; Dragani, T. A.; Fletcher, C. F.; Forrest, A.; Frazer, K. S.; Gaasterland, T.; Gariboldi, M.; Gissi, C.; Godzik, A.; Gough, J.; Grimmond, S.; Gustincich, S.; Hirokawa, N.; Jackson, I. J.; Jarvis, E. D.; Kanai, A.; Kawaji, H.; Kawasawa, Y.; Kedzierski, R. M.; King, B. L.; Konagaya, A.; Kurochkin, I. V.; Lee, Y.; Lenhard, B.; Lyons, P. A.; Maglott, D. R.; Maltais, L.; Marchionni, L.; McKenzie, L.; Miki, H.; Nagashima, T.; Numata, K.; Okido, T.; Pavan, W. J.; Pertea, G.; Pesole, G.; Petrovsky, N.; Pillai, R.; Pontius, J. U.; Qi, D.; Ramachandran, S.; Ravasi, T.; Reed, J. C.; Reed, D. J.; Reid, J.; Ring, B. Z.; Ringwald, M.; Sandelin, A.; Schneider, C.; Semple, C. A.; Setou, M.; Shimada, K.; Sultana, R.; Takenaka, Y.; Taylor, M. S.; Teasdale, R. D.; Tomita, M.; Verardo, R.; Wagner, L.; Wahlestedt, C.; Wang, Y.; Watanabe, Y.; Wells, C.; Wilming, L. G.; Wynshaw-Boris, A.; Yanagisawa, M.; Yang, I.; Yang, L.; Yuan, Z.; Zavolan, M.; Zhu, Y.; Zimmer, A.; Carninci, P.; Hayatsu, N.; Hirozane-Kishikawa, T.; Konno, H.; Nakamura, M.; Sakazume, N.; Sato, K.; Shiraki, T.; Waki, K.; Kawai, J.; Aizawa, K.; Arakawa, T.; Fukuda, S.; Hara, A.; Hashizume, W.; Imotani, K.; Ishii, Y.; Itoh, M.; Kagawa, I.; Miyazaki, A.; Sakai, K.; Sasaki, D.; Shibata, K.; Shinagawa, A.; Yasunishi, A.; Yoshino, M.; Waterston, R.; Lander, E. S.; Rogers, J.; Birney, E.; Hayashizaki, Y. *Nature* **2002**, *420*, 563.
55. Gojobori, T.; Nei, M. *J. Mol. Evol.* **1981**, *17*, 245.
56. Gonzalez-Diaz, H.; Aguero-Chapin, G.; Varona-Santos, J.; Molina, R.; de la Riva, G.; Uriarte, E. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2932.
57. Gonzalez-Diaz, H.; de Armas, R. R.; Molina, R. *Bull. Math. Biol.* **2003**, *65*, 991.
58. Gonzalez-Diaz, H.; Perez-Bello, A.; Uriarte, E.; Gonzalez-Diaz, Y. *Bioorg. Med. Chem. Lett.* **2005**.
59. Marrero-Ponce, Y.; Medina-Marrero, R.; Castillo-Garit, J. A.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 3003.
60. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castanedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sanchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005.
61. Mazzatorta, P.; Benfenati, E.; Lorenzini, P.; Vighi, M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 105.
62. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. *J. Antimicrob. Chemother.* **2004**, *53*, 65.
63. Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert-Salvador, M. T.; Diaz-Villanueva, W.; Medina-Casamayor, P. *J. Mol. Graphics Model.* **2003**, *21*, 375.
64. Murcia-Soler, M.; Perez-Gimenez, F.; Nalda-Molina, R.; Salabert-Salvador, M. T.; Garcia-March, F. J.; Cercos-del-Pozo, R. A.; Garrigues, T. M. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1345.
65. Garcia-Garcia, A.; Galvez, J.; de Julian-Ortiz, J. V.; Garcia-Domenech, R.; Munoz, C.; Guna, R.; Borras, R. *J. Biomol. Screen.* **2005**, *10*, 206.
66. Duart, M. J.; Anton-Fos, G. M.; Aleman, P. A.; Gay-Roig, J. B.; Gonzalez-Rosende, M. E.; Galvez, J.; Garcia-Domenech, R. *J. Med. Chem.* **2005**, *48*, 1260.
67. Bruno-Blanch, L.; Galvez, J.; Garcia-Domenech, R. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2749.
68. Gozalbes, R.; Brun-Pascaud, M.; Garcia-Domenech, R.; Galvez, J.; Pierre-Marie, G.; Jean-Pierre, D.; Derouin, F. *Antimicrob. Agents Chemother.* **2000**, *44*, 2771.
69. Santana, L.; Uriarte, E.; González-Díaz, H.; Zagotto, G.; Soto-Otero, R.; Méndez-Alvarez, E. *J. Med. Chem.* **2006**, *49*, 1149.
70. Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Pereira, D. M.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Aran, V. J.; Martinez-Fernandez, A. R.; Garcia Sanchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838.
71. Duart, M. J.; Garcia-Domenech, R.; Anton-Fos, G. M.; Galvez, J. *J. Comput. Aided Mol. Des.* **2001**, *15*, 561.
72. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J. P.; Bertrand, H. O. *J. Med. Chem.* **2005**, *48*, 2534.
73. Agüero-Chapin, G.; Gonzalez-Diaz, H.; Molina, R.; Varona-Santos, J.; Uriarte, E.; Gonzalez-Diaz, Y. *FEBS Lett.* **2006**, *580*, 723.
74. Ramos de Armas, R.; Gonzalez-Diaz, H.; Molina, R.; Uriarte, E. *Proteins* **2004**, *56*, 715.
75. SantaLucia, J. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1460.
76. Han, J.; Lee, Y.; Yeom, K.-H.; Nam, J.-W.; Heo, I.; Rhee, J.-K.; Sohn, S. Y.; Cho, Y.; Zhang, B.-T.; Kim, V. N. *Cell* **2006**, *125*, 887.
77. Oberg, T. *Chem. Res. Toxicol.* **2004**, *17*, 1630.
78. Liu, H.; Papa, E.; Gramatica, P. *Chem. Res. Toxicol.* **2006**, *19*, 1540.
79. Gramatica, P.; Giani, E.; Papa, E. *J. Mol. Graphics Model.* **2006**.
80. Markham, N. R.; Zuker, M. *Nucleic Acids Res.* **2005**, *33*, W577.
81. StatSoft. 2002.