On the normalization of RNA equilibrium free energy to the length of the sequence

Dmitri D. Pervouchine, Joel H. Graber, and Simon Kasif

Bioinformatics Program, Boston University, Boston MA 02215

Keywords: RNA, secondary structure, dynamic programming, micro-RNA

Author for correspondence:

D. D. Pervouchine
Bioinformatics Program
Boston University
Boston
MA 02215
tel.: (617) 353-5463
fax: (617) 353-5462
e-mail: dp@bu.edu

ABSTRACT

There is no universal definition of stability for RNA secondary structures. In this paper we present an approach that is based on normalization of the equilibrium free energy to the length of the sequence: a segment of RNA is said to be stable if the ratio of the equilibrium free energy to the length of the segment is greater than a certain threshold value. Discarding the segments whose normalized equilibrium free energies are smaller than the threshold allows us to view the secondary structure at different levels of stability. Confined to only highly stable structures, the algorithm for secondary structure prediction admits a number of simplifications that make it computationally tractable for large sequences and advantageous over most other methods on genome-wide scale. This method was applied to *Caenorhabditis elegans* genome to localize the regions that encode for stable secondary structures. In particular, 36 of 56 previously reported micro-RNAs were localized to 4% of the genome. A fraction of long (400 nt and longer) stable inverted repeats in the genomic sequence of *C. elegans* was found. Their distribution is very uneven and skewed towards the ends of chromosomes. This method can be used for genome-wide detection of transcription termination signals, putative micro-RNAs, and other regulatory elements that involve stable RNA secondary structures.

Introduction

Existing RNA folding algorithms fall into two major classes: MFOLD type algorithms and covariance models. MFOLD type algorithms find structures with the lowest equilibrium free energy and share a common dynamic programming core, while covariance models recognize positions that are covarying to maintain base complementarity and use the strategy of Hidden Markov Models [1]. The dynamic programming approach to RNA secondary structure prediction was pioneered about 30 years ago. Nussinov presented a simple free energy function that is minimized when the secondary structure contains maximum number of complementary base pairs [2]. Tinoco calculated the free energy as a sum of independent energies for each of the loops in the structure, rather than for each of the complementary base pairs [3]. Zuker introduced more realistic free energy function that is based on experimentally determined thermodynamic parameters and takes into account coaxial stacking energies [4, 5].

Zuker's algorithm requires $O(n^3)$ time and $O(n^2)$ memory for a sequence of length *n* but doesn't allow for pseudoknots.

The output of an MFOLD type algorithm is a list of intramolecular base pairings (that is, the secondary structure) that corresponds to the lowest equilibrium free energy. The structure is derived from the recursively calculated dynamic programming matrix, whose entries are the minimal folding energies of all segments of the sequence. This matrix also contains information about all suboptimal structures. Suboptimal structures have energies that are nearly equal to the lowest equilibrium free energy and, therefore, are almost as thermodynamically stable as the optimal structure. Some of the base pairs are present in all or in the majority of suboptimal structures, while other pairings are specific for only few of them. This suggests that the secondary structure consists of core and variable parts.

The definition of the core part, as being the segment that has the largest negative equilibrium free energy, is confronted with the following difficulty: longer segments have lower free energy because they have more bases to pair, and the segment with the largest negative free energy turns out to be the whole sequence. This leads us to the idea of normalization of the equilibrium free energy to the length of the segment. That is, in order to treat all parts of the sequence equitably we need to divide the equilibrium free energy of each segment by an appropriate quantity that is a function of length. In this work we explore only the case when this function is linear, since it is the most appropriate normalization in context of stability: the minimum equilibrium free energy over all possible sequences of length n grows linearly with n. A linear normalization factor also has an advantage of scaling the free energies from zero to one and allows analytic development of statistical frameworks. The maximum number of base pairs approximation, which is discussed in the next section, makes this method computationally tractable for large sequences and advantageous over most other methods on genome-wide scale.

Method

The free energies don't need to be calculated separately for each segment of the sequence, as they are already encrypted in the dynamic programming matrix provided by a secondary structure prediction tool. Denote the entries of this matrix by E_{ij}. Once E_{ij} are known, we divide them by the lengths of corresponding segments. This transforms E_{ij} to ε_{ij}, where

$$\varepsilon_{ij} = \frac{E_{ij}}{|j - i + 1|}. \quad (1)$$

The quantity of ε_{ij} has units of linear density of free energy and is referred to as the *stability factor*. If we now chose a threshold ε for the stability factor and drop all segments that have values of ε_{ij} smaller than ε then the rest of the sequence will correspond to the stable part of the secondary structure. This stable part depends on ε and is called ε-stable. Stable structures with different levels of stability are related to each other monotonically: if $\varepsilon_1 > \varepsilon_2$, then the ε_1-stable structure (as a set of base pairs) is a subset of the ε_2-stable structure.

Before elaborating on these ideas, let us make a number of simplifications. First, if we are interested in highly stable secondary structures, that is, ones with only few bases unpaired, then we don't need much accuracy in the prediction of the equilibrium free energy and can roughly approximate it (up to a constant factor) with the maximum number of paired bases. This facilitates an increase in speed and the ability to manipulate the matrix E_{ij} analytically. Once the regions that have high density of equilibrium free energy are found, their secondary structures can be refined with MFOLD. Later we show that stem-loops are, in general, more stable than the other types of unknotted secondary structures. Branching structures usually have large unpaired regions at the points where they branch and, therefore, have smaller percentage of paired bases per unit length. By the same reason, the most stable stem-loops are ones that have long stems and short loops. We make use of these observations by considering only non-branching secondary structures. This allows us to implement the calculations of E_{ij} in quadratic time and space using simplified version of the original Nussinov algorithm [2]. If we confine ourselves to the stem-loops whose lengths are uniformly bounded by a constant k then we won't need E_{ij} for $|j-i|>k$. In this case calculation of E_{ij} can be done in linear time and space.

From now on we consider only non-branching secondary structures and assume that E_{ij} is the maximum number of complementary base pairs in the segment between bases i and j. It is clear that $0 \le E_{ij} \le j-i+1$ and, therefore, $0 \le \varepsilon_{ij} \le 1$ for all i and j. This shows that the natural range of the stability threshold is from zero to one. As we will see later, the typical values of stability threshold for highly stable structures (that is, ones of interest to us) vary from 0.6 to 0.9.

Approximation of the free energy with the maximum number of complementary bases allows us to answer important statistical questions. First, we estimate the probability distribution function of equilibrium free energies. Then we calculate the probability of occurrence of an n-long and ε-stable subsequence in a random sequence of a given length. From these results we infer the critical length of a stem-loop, whose presence in a sequence of given length is essentially non-random. Namely, the critical length (in bases) of a stem-loop n_0 at the significance level α is given by formula

$$n_0 = \frac{\ln\alpha - \ln N}{\ln\lambda(\varepsilon)}, \tag{2}$$

where

$$\lambda(\varepsilon) = \left(e\sqrt{p}(\varepsilon^{-1}-1)\right)^{\varepsilon}. \tag{3}$$

Here N is the sequence length, ε is the stability threshold, e is the base of the natural logarithm, and p is the parameter that describes the nucleotide composition of the sequence ($p = 0.25$ for uniform distribution of bases). The reader is referred to appendix for detailed derivation of these formulae. Table 1 gives an example of calculations of critical stem-loop lengths for $p = 0.25$ and $N = 10^8$.

Results

Based on the method described in the previous section we developed STLS — a program for prediction of stable RNA secondary structures. The input of STLS consists of three components: nucleotide sequence, stability threshold, and other parameters. The first two arguments have obvious meaning. Description of other parameters is found in the STLS manual. The output of STLS tabulates stability factors, positions and secondary structures of all ε-stable segments. A copy of STLS can be obtained from our web site http://genomics10.bu.edu/dp/rna/stls/ or from the authors by request. In this section we give a number of tests for STLS and also list some of its applications.

To test the accuracy of STLS predictions we generated 700 random 400-long nucleotide sequences with uniform distribution of bases. For each sequence we took the secondary structure S_s predicted by STLS and compared it to the secondary structure S_v predicted by VIENNA algorithm [6]. As a quantitative measure of accuracy we used σ, the ratio of the number of base pairs that belong to both S_s and S_v to the number of base pairs that belong to S_s, that is $\sigma = |S_s \cap S_v| / |S_s|$, where $|S|$ denotes the cardinality of a set S. In other words, σ is the specificity of the prediction of STLS with respect to the prediction of VIENNA. For ε=0.68, 0.72, 0.78, and 0.80 we obtained the following values of σ: 0.88, 0.91, 0.98, and 0.99 respectively. This indicates that the predictions of STLS are consistent with the predictions of the traditional method if the stability threshold is high enough.

Another accuracy test was performed on the set of 56 micro-RNAs in *Caenorhabditis elegans*. Recall that micro-RNAs are small (about 22nt) regulatory RNAs that are generated from the common stem-loop precursors by the process requiring Dicer, a protein that cleaves the stem-loop [7]. These precursors are structures with short loops and long stems, which have only few small bulges or internal loops. This study was motivated by the challenge to determine whether STLS can find micro-RNA sequences in *C. elegans* genome. The complete sequence of *C. elegans* was obtained from the WORMBASE website [8] and then scanned with STLS using stability thresholds varying from 0.68 to 0.82. For ε=0.68 it yielded approximately 44000 stable stem-loops, which together correspond to approximately 4% of the genome. It turned out that 36 of 56 micro-RNAs (64.3%) were in our list.

However, micro-RNAs correspond to only a small fraction of the stem-loops found by STLS. It is interesting to explore the rest of the list. We examined all 44000 stem-loops for their lengths, locations in chromosomes and membership in annotated functional parts of the genome. The distribution of length (figure 1) revealed that the most of them are short (less than 100nt). However, there is a large fraction of very long stable stem-loops (400 nt and longer), which can be seen at higher thresholds (ε=0.82). In the next experiment we investigated the spatial density of stable stem-loops. Figure 2 shows that this distribution is very uneven and biased towards the ends of chromosomes. To verify this observation numerically, we subdivided all chromosomes into 250 bins of equal length and calculated the number of stem-loops n_i, and GC content β_i for each bin. The values of Pearson correlation coefficient r (for n_i versus $2 \cdot |\beta_i - 0.5|$), Spearman rank correlation r', and χ^2 statistics (for n_i) are given in table 3. The P-values for χ^2 test with $n = 249$ degrees of freedom were computed using approximation by normal distribution with mean n and standard deviation $\sqrt{2n}$.

In an attempt to elucidate possible functions of the stem-loops found, we mapped them to annotated introns, exons, intergenic regions, 5'- and 3'-untranslated regions (5'UTRs and 3'UTRs, respectively). The quality of this analysis, however, is very dependent on the quality of the predictions of intron-exon boundaries and untranslated regions. As the actual length of the UTRs was not known, we used 180-long sequences upstream and downstream of the corresponding genes. The corresponding densities of the stable stem-loops in introns, exons, genes (introns + exons), intergenic regions, 5'UTRs, and 3'UTRs were 23.4, 11.6, 16.3, 17.4, 7.9, and 10.8 bases of stem-loops per 1000 of bases of sequence, respectively. Note that introns have higher, while untranslated regions have lower values of stem-loop density compared to exons and intergenic regions.

Discussion

A common argument against the maximum number of base pairs approach to the RNA secondary structure prediction problem is that the strands of a helix are held together by coaxial stacking interactions of paired bases rather than by hydrogen bonds. In the VIENNA package the energies of helices are calculated by adding stacking energies for each pair of neighboring base pairs. However, if we set the stability threshold at 0.78 or higher then the predictions of both VIENNA and STLS are essentially the same (specificity $\sigma = 98\%$), although STLS doesn't take stacking energies into account. This happens because at high value of threshold we *a priori* confine ourselves to very long helices. In any double-stranded structure of length n there are n base pairings, and $n-1$ stacking interactions. Certainly, as n increases, the differences between stacking energies of different base pairs average out yielding a quantity that is proportional to the number of stacking interactions. So, when the normalization is performed, the discrepancy between a purely base pairing energy function and the energy function that also takes stacking energies into account decreases as n gets larger and $(n-1)/n$ becomes closer to one. This explains why lengthy stem-loops were correctly predicted by STLS at high threshold levels. Of course, the predictions of STLS and VIENNA differ more significantly when ε is smaller than 0.78, which has to do with the fact that stacking doesn't treat all combinations of bases equally, while STLS scores them the same. Note that one of the advantages of STLS is increase in calculation speed (about 5000 fold for a 500-long sequence). Thus, in the cases when lower thresholds are needed, one can use STLS for preliminary identification of the stable regions, and then refine their secondary structures with VIENNA or MFOLD. We also recall that STLS can use any matrix of free energies E_{ij} (for instance, the matrix generated by VIENNA) as an input.

In the experiment with micro-RNAs we were able to localize about 64% of micro-RNAs [7] to a list of 44,000 segments (approximately 4% of the genome) without any prior knowledge about their positions and relying solely on the hypothesis that they belong to stable stem-loops. This result, of course, is a usual interplay between sensitivity and specificity: the sensitivity drops dramatically when the specificity increases (table 2).

There was no micro-RNAs found for $\varepsilon = 0.76$. This fact might indicate that the stem-loop precursors of micro-RNAs have to be stable but slightly imperfect in order to be recognized by the cellular machinery and bypass degradation pathways.

While micro-RNAs are needed to repress the translation of a target gene, the function of the other stem-loops, as well as their origin, remains unclear. Figure 1 shows that the most of them have length about 100nt, but there is also a fraction of longer (400 nt and above) stem-loops. According to table 1, the presence of these long stem-loops is qualified as essentially non-random. It is remarkable that they were seen at all four thresholds, separating from the short fractions when the threshold increases. The similarity of the peaks' structures in figure 1 (b), (c), and (d) suggests that these very long stem-loops are threshold independent.

From now on we set $\varepsilon = 0.82$. The regularity of the peaks in figure 1 (d) suggests that such family of stem-loops might appear as a result of gene duplication or be developed by exogenous factors such as multiple incorporation of double-stranded genetic material into the same spot on DNA. The latter hypothesis is supported by the fact that the abscissas of peaks on figure 1 are almost multiple of each other. However, the spatial correlations between stem-loops' locations (figure 2) is indicative of generation through tandem duplication. Indeed, repetitive elements often have inverted repeats associated with them (in the form of LTRs), and as such would be expected to score highly in STLS.

It is very remarkable that the distribution of stable stem-loops in *C. elegans* chromosomes is very uneven and biased towards their ends (see figure 2 and P-values in table 3). This finding is in good agreement with the results of Surzycki and Belknap on the distribution of MITE-like repeats in *C. elegans* [9] and with the fact that central regions of autosomes (chromosomes I-V) have higher density of genes than their arms [10]. One may expect a greater probability of paired bases in sequences that are either GC-rich or GC-poor, since the probability of any two bases being complementary is higher in such sequences. Our analysis (Pearson and Spearman statistics in table 3) showed that this was not contributing significantly.

Also, it is not entirely clear whether the clusters of stable stem-loops carry out any biological function. Recall that the density of the stable stem-loops in introns and intergenic regions is at least two times higher than in untranslated regions. This observation is not surprising because the untranslated regions usually contain important cis-elements that are responsible for protein-RNA interactions; secondary structures would potentially compete with or disrupt such interactions and therefore might be expected to be selected against in the UTR sequence.

Conclusions

We presented an efficient method for prediction of stable RNA secondary structure. The key part of the method relies on the normalization of the equilibrium free energy to the length of the sequence. In the class of stable secondary structures the algorithm was shown to have good performance for long sequences and the results are consistent with the other RNA secondary structure prediction methods. Using this method we located the regions in *C. elegans* genomic sequence that encode for stable secondary structures and characterized their distributions. In particular, we localized 64% of micro-RNAs previously reported by Lau in approximately 4% of the genome relying solely on the property of micro-RNAs to belong to the stable stem-loops. We report that there is a fraction of long (400nt and above) stable stem-loops in *C. elegans* genome; their distribution is very uneven and skewed towards the ends of chromosomes. The method we developed can be used for the detection of transcription termination signals and putative micro-RNAs, as well as many other regulatory elements that correspond to stable secondary structure.

Acknowledgements

This work is supported by Bioinformatics Program at Boston University. The first author wishes to thank Chris Sander, Debbie Marks, Charles DeLisi and Robert Berwick for their valuable comments and insightful discussions.

Appendix

We recall the formal language of secondary structures. Let $X=(x_1,\ldots x_n)$ be a sequence of letters from the alphabet $\Omega=\{A,C,T,G\}$. A *secondary structure* is a set S of pairs (i,j) with $1\le i<j\le n$ such that for all $(i,j),(i',j')\in S$ the condition $i=i'$ implies $j=j'$ and vice versa. The relation "$\prec$", where $(i',j')\prec(i,j)$ if $i<i'<j'<j$, defines a partial order on S. A secondary structure is said to be *non-branching* if "$\prec$" is a linear order, that is, if $\pi\prec\pi'$ or $\pi'\prec\pi$ for all $\pi,\pi'\in S$. Consider the following additive energy function

$$E(X,S)=\sum_{(i,j)\in S}e(x_i,x_j),$$

where $e(\,\cdot\,,\cdot\,)$ is a scoring matrix. For simplicity, now we assume that $e(x,y)=2$ if x and y are Watson-Crick complementary bases and $e(x,y)=0$ otherwise. We define

$$E(X)=\max_S\{E(X,S)\mid S\ is\ non-branching\},$$

that is, $E(X)$ is the maximum number of complementary bases in the sequence X over all possible non-branching secondary structures.

For a given sequence X, the value of $E(X)$ is calculated recursively by formula

$$E_{ij}=\max\{E_{i+1j},E_{ij-1},E_{i+1j-1}+e(x_i,x_j)\},$$

where $E_{ij}=E((x_i,\ldots,x_j))$, $i,j=1\ldots n$. Then the matrix E_{ij} is transformed to the matrix ε_{ij} using (1), and then all regions that have ε_{ij} greater than a certain threshold value are identified. If the length of a stem-loop has a prior upper bound d, then in place of E_{ij} we consider the matrix $E'_{ik}=E_{i\,i+k}$, which is also expressed recursively as

$$E'_{ik}=\max\{E'_{i+1k-1},E'_{ik-1},E'_{i+1k-2}+e(x_i,x_{i+k})\},$$

where $i=1\ldots n-d$ and $k=1\ldots d$, and then it is transformed to the matrix ε_{ij}.

Now we want to calculate the probability of observing a stable stem-loop in a random sequence. Suppose that X is a sequence of independent random letters $x_1,\ldots,x_n$ that came from the same (but not necessarily uniform) distribution. Let p denote the probability that two independent random letters from this distribution are complementary. Then $E_n = E(X)$ is a random number that depends only on n and p.

We are ready to estimate the probability distribution of E_n. Define $P_n(k)$ as the probability that $E_n \geq k$. We may assume that k is an even integer. For non-branching structures the event $\{E_n \geq k\}$ can be decomposed into the sum of smaller mutually exclusive. Namely, the outmost arc of a non-branching secondary structure that connects bases i and j can be placed in $n-1+l$ different ways, where $l = j-i+1$, and there are n possibilities for choosing the value of l. Since the probability that the outmost arc has length l is smaller than or equal to p, we get

$$P_n(k) \leq \sum_{l=k}^{n} (n-l) P_{l-2}(k-2) \tag{4}$$

Applying (4) to itself recursively s times, we get

$$P_n(k) \leq \sum_{l_1=k}^{n} (n-l_1) \sum_{l_2=k}^{l_1} (l_1-l_2) \ldots \sum_{l_s=k}^{l_{s-1}} (l_{s-1}-l_s) P_{l_s-2s}(k-2s).$$

This process will stop when $k=2s$. To estimate $P_n(k)$, we use the following continuous approximation

$$\sum_{i=0}^{n} (n-i)\, i^k \approx \int_{o}^{n} (n-x)\, x^k dx = \frac{n^{k+2}}{(k+2)(k+1)}.$$

Then we can estimate $P_n(k)$ as $P_n(k) \leq \frac{(n-k)^k}{k!\, p^{k/2}}$. Using the Stirling formula, we get

$$P_n(k) \leq \left(\frac{n-k}{k} \right)^k e^k p^{k/2},$$

where e is the base of the natural logarithm. Equivalently, if we denote k/n by ε then

$$P_n(\varepsilon) \leq \left(\lambda(\varepsilon)\right)^n, \tag{5}$$

where $\lambda(\varepsilon)$ is given by (3). Note that $P_n(\varepsilon)$ is the probability that a random sequence of length n has stability factor greater than or equal to ε. Now we fix ε and consider $P_n(\varepsilon)$ as a function of n. The inequality (5) gives us an upper limit estimate for $P_n(\varepsilon)$. We are interested in a range of ε such that $(\lambda(\varepsilon))^n$ decays when n increases. This condition holds only if $\varepsilon \geq \varepsilon_0 = e\sqrt{p}/(1+e\sqrt{p})$. Particularly, for the uniform distribution we have $p = 0.25$ and $\varepsilon_0 = 0.57$. The approximation for $k!$ assumes that the values of k and, therefore, of n are large enough. Thus, formula (5) effectively estimates only the "tail"' of the function $P_n(\varepsilon)$.

It trivially follows from (5) that if $N >> n$ then the probability that a random sequence of length N contains a n-long and ε-stable subsequence is $N \cdot P_n(\varepsilon)$. Now we are interested in the critical value of n, at which the presence of a n-long and ε-stable subsequence in a sequence of length N can be considered as non-random. Simple algebra proves that if $N >> n$, $N(\lambda(\varepsilon))^n << 1$ and n is large enough, then the critical value of n at significance level α is given by (2).

References

1. Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological sequence analysis*. University Press, Cambridge, 1998.
2. R. Nussinov, G. Pieczenik, J.R. Griggs, and D. J. Kleitman. “Algorithms for Loop Matching” *Journal of Applied Mathematics* **35**(1):68-82, 1978.
3. Tinoco, P.N. Borer, B. Dengler, M.D. Levine. “Estimation of Secondary Structures of Ribonucleic Acids” *Nature* **230**:362-367, 1971.
4. Amy E. Walter, Douglas H. Turner, James Kim, Matthew H. Lyttle, Peter Muller, David H. Mathews, and Michael Zuker. “Coaxial Stacking of Helixes Enhances Binding of Oligoribonucleotides and Improves Predictions of RNA Folding” *Proceedings of National Academy of Sciences* **91**:9218—9222, 1994.
5. D.H. Mathews, J. Sabina, M. Zucker, and H. Turner. “Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure” *Journal of Molecular Biology* 288:911-940, 1999.
6. Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, L. Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. “Fast Folding and Comparison of RNA Secondary Structures” *Monatshefte Fur Chemie* **125**(1):167-188, 1994.
7. Nelson C. Lau, Lee P. Lim, Earl G. Weinstein, and David P. Bartel. “An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis Elegans” *Science* **294**:858-862, 2001.
8. Lincoln Stein, Paul Sternberg, Richard Durbin, Jean Thierry-Mieg, John Spieth. “WormBase: Network Access to the Genome and Biology of Caenorhabditis Elegans” *Nucleic Acids Research* **29**(1):82-86, 2001.
9. Stefan A. Surzycki and William R. Belknap. “Repetitive-DNA elements are similarly distributed on *Caenorhabditis elegans* autosomes” *Genetics* **97**(1):245-249, 2000.
10. Laurent Duret, Gabriel Marais, and Christian Biémont. “Transposons but Not Retrotransposons Are Located Preferentially in Regions of High Recombination Rate in Caenorhabditis elegans” *Genetics* **156**:1661-1669, 2000.

Figure 1. The molar (blue) and mass (green) distributions of lengths of stable stem-loops at the thresholds 0.72 (a), 0.78 (b), 0.80 (c), and 0.82 (d). The numerical value of the molar (respectively, mass) distribution function is the percentage of ε-stable stem-loops (respectively, nucleotides in ε-stable stem-loops), whose lengths range from x to dx (here dx is 20 nucleotides).

Figure 2. Distributions of stable stem-loops in the genome of *C. elegans*. Darker regions correspond to higher density of stem-loops. The same intensity scale is used for all six chromosomes.

Table 1. Critical lengths of stem-loops at significance level α, stability thresholds ε, sequence length $N = 10^8$ (genome size of *C. elegans*), and probability of complementary pairing $p = 0.25$.

Table 2. Percentage of micro-RNAs found in ε-stable stem-loops. Here ε and "% of genome" denote the stability threshold and the percentage of bases that belong to ε-stable stem-loops relative to genome size, respectively.

Table 3. Pearson correlation coefficient r, Spearman rank correlation r', and χ^2 for the number of stem-loops in the i-th bin ($i = 1\ldots 250$) as a function of deviation of GC content from the uniform distribution. The values of χ^2 are converted to z-score and P-values are shown.

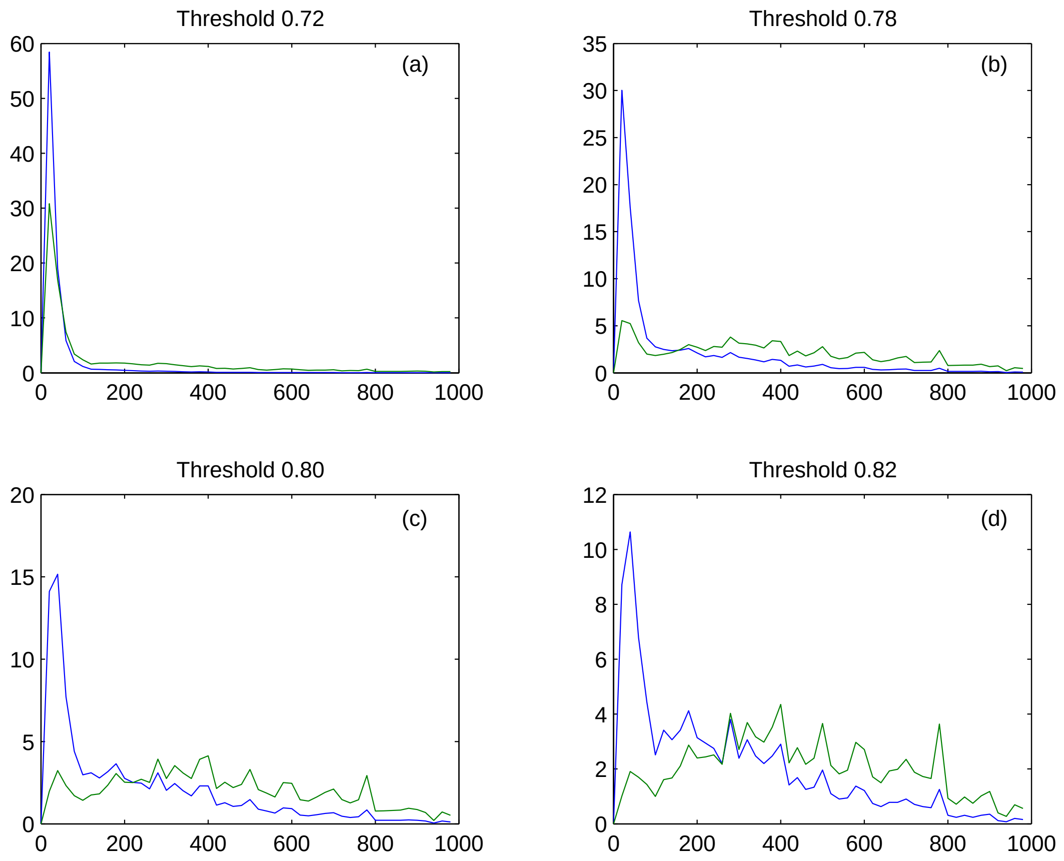

Figure 1.

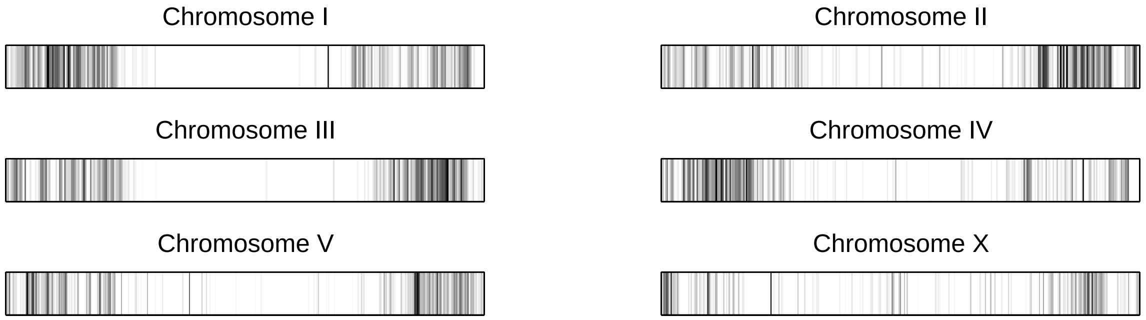

Figure 2.

α	ε				
	0.68	0.72	0.78	0.80	0.82
0.05	70	46	28	24	21
0.001	83	55	33	29	25
0.00001	98	65	40	34	30

Table 1.

ε	% of micro-RNA	% of the genome
0.65	64.3	17.5
0.68	64.3	4.0
0.71	33.9	2.2
0.74	7.1	1.1
0.76	0	0.4

Table 2.

Chromosome	I	II	III	IV	V	X
$r \cdot 10^2$	1.95	-0.62	-5.78	0.63	-9.10	2.82
$r' \cdot 10^2$	-0.51	-14.12	-0.47	8.87	-2.95	-3.11
χ^2	1020	446	831	674	795	368
z	34.45	8.77	25.98	18.97	24.41	5.51
P-value	$7 \cdot 10^{-259}$	$7 \cdot 10^{-18}$	$8 \cdot 10^{-148}$	$2 \cdot 10^{-79}$	$1 \cdot 10^{-130}$	$2 \cdot 10^{-10}$

Table 3.