BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 4 2014, pages 457–463
doi:10.1093/bioinformatics/btt701

Sequence analysis

Advance Access publication November 28, 2013

RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments

Ruslan A. Soldatov[1,2,*], Svetlana V. Vinogradova[1,2] and Andrey A. Mironov[1,2]

[1]Institute for Information Transmission Problems (the Kharkevich Institute), Russian Academy of Sciences, 19 Bolshoy Karetny per., Moscow 127994 and [2]Department of Bioengineering and Bioinformatics, Moscow State University, 1-73 Vorobyevy Gory, Moscow 119991, Russia

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: During the past decade, new classes of non-coding RNAs (ncRNAs) and their unexpected functions were discovered. Stable secondary structure is the key feature of many non-coding RNAs. Taking into account huge amounts of genomic data, development of computational methods to survey genomes for structured RNAs remains an actual problem, especially when homologous sequences are not available for comparative analysis. Existing programs scan genomes with a fixed window by efficiently constructing a matrix of RNA minimum free energies. A wide range of lengths of structured RNAs necessitates the use of many different window lengths that substantially increases the output size and computational efforts.

Results: In this article, we present an algorithm RNASurface to efficiently scan genomes by constructing a matrix of significance of RNA secondary structures and to identify all locally optimal structured RNA segments up to a predefined size. RNASurface significantly improves precision of identification of known ncRNA in *Bacillus subtilis*.

Availability and implementation: RNASurface C source code is available from http://bioinf.fbb.msu.ru/RNASurface/downloads.html.

Contact: ruslansoldatov@gmail.com

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on September 13, 2013; revised on November 8, 2013; accepted on November 25, 2013

1 INTRODUCTION

RNAs perform diverse enzymatic, regulatory and structural functions in living cells and are involved in nearly all housekeeping cellular processes. Numerous new classes of non-coding and regulatory RNAs were discovered during past decades. For example, microRNAs regulate gene expression through post-transcriptional repression of messenger RNA (Bartel, 2009). Other prominent examples include the RNA role in translation (Kozak, 2005), messenger RNA localization (Martin and Ephrussi, 2009), alternative splicing (Pervouchine *et al.*, 2012), epigenetic states (Lee, 2012), virus replication (Gultyaev *et al.*, 2010) and so forth. Moreover, according to recent studies of the human genome, up to 90% of the genome may be transcribed, and a substantial fraction of transcripts may represent functional non-coding RNAs (Djebali *et al.*, 2012).

Biological function of the majority of functional RNAs crucially depends on its tertiary structure. RNA molecules fold hierarchically and sequentially, therefore the tertiary structure is largely determined by the secondary structure scaffold (Tinoco and Bustamante, 1999). For example, riboswitches regulate gene expression by switching between two specific secondary structures (Vitreschak *et al.*, 2004). There exist numerous methods to predict secondary structure computationally (for a review see Mathews *et al.*, 2010), among which free energy minimization (Zuker and Stiegler, 1981) based on nearest neighbor thermodynamic parameters (Mathews *et al.*, 2004) is the most prevalent approach. Generally, RNA sequences have numerous suboptimal structures, which have energies close to the optimal one (Giegerich *et al.*, 2004). In addition to this, inaccuracies of the energy parameters, base modifications and dependence on cell conditions yield a moderate quality of the structure prediction by energy minimization with the average sensitivity of 75% and selectivity of 79% (Lorenz *et al.*, 2011). However, the optimal energy could be a good measure to define propensity of a sequence to fold in a stable structure, the minimum free energy (*MFE*) of functional RNAs can be distinguished from that of random sequences (Clote *et al.*, 2005). On the other hand, there exist RNA regions actively bound by protein factors. Relatively low *MFE*, due to selection against formation of stable secondary structure, may represent a signal of an accessible site (Keller *et al.*, 2012). Structural potential of an RNA sequence reflects significance of its secondary structure *MFE*.

Next-generation sequencing technologies produce vast amounts of genomic data. Hence, identification in genomes of segments with unusual structural potential is of special interest. This problem has two sides: to efficiently scan genomes and to assess significance of output hits. There are several programs that efficiently scan genomes. RNALL (Wan *et al.*, 2006) and RNAslider (Horesh *et al.*, 2009) use a window-based approach to calculate the MFE and structure of each local sequence of a predefined size. RNALfold (Hofacker *et al.*, 2004) slightly diminishes the output size by considering only structures with paired first and last nucleotides, so-called closed structures. These programs focus on efficient computing and ignore significance of output hits. The only discriminator in these programs is *MFE*. However, *MFE* strongly depends on the length and dinucleotide composition of a sequence (Bonnet *et al.*, 2004; Workman and Krogh, 1999). Comprehensive comparison of RNA structure scores (Freyhult *et al.*, 2005) shows that

*To whom correspondence should be addressed.

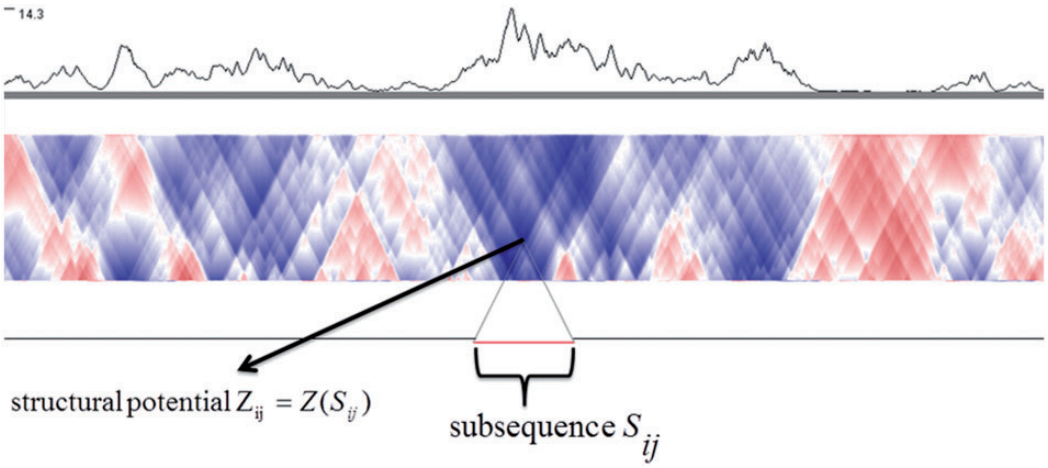


Fig. 1. The surface of the structural potential of 1000 nt region in the *B.subtilis* genome (1180240–1181240) that contains an SAM riboswitch. Points on the surface correspond to subsequences. Dedicated red subsequence has a size of ~100 nt. Colors represent significance of the RNA secondary structure: blue points correspond to highly structured regions and red points to unstructured ones. The slanted line from a highly structured peak to the respective region in the sequence is shown. The top plot is a one-dimensional measure of the RNA secondary structure significance (the maximum squared *Z*-score) introduced in Section 2.3.2

Z-score (Washietl *et al.*, 2005) can be used as a suitable measure of the structural potential. The *Z*-score of a sequence is calculated as

$$Z = \frac{E - \mu}{\sigma} \quad (1)$$

where E is the MFE of a biological sequence, μ and σ are the average and standard deviation of the energy distribution of shuffled sequences with preserved length and average dinucleotide composition. There also exists a less common approach, which uses uniqueness of secondary structure instead of thermodynamic stability to assess its significance (Le *et al.*, 2003).

An early attempt to detect significant structures was based on a Monte Carlo simulation (Le *et al.*, 1988). This approach becomes impractical for sequences of moderate (several thousand of nucleotides) size. Further studies introduced look-up tables to eliminate bias of nucleotide composition, combined with regression on sequence length (Chen *et al.*, 1990) or the notion of asymptotic *Z*-score (Clote *et al.*, 2005). The current state-of-the-art program RNALfoldz (Gruber *et al.*, 2010a) extends RNALfold by calculating *Z*-score using SVR. To our knowledge, the implementation of RNALfoldz is now embedded in RNALfold.

Approaches, that analyze RNA segments of a fixed size, suffer from an inability to detect structured elements of varying sizes, whereas functional RNAs fall into a wide range of sizes. Here, we circumvent this limitation. To that end, we define the surface of the structural potential as the surface of *Z*-scores for all genome subsequences shorter than a predefined size. Figure 1 visualizes this concept. The line at the bottom represents a segment of the genome, each point on the heatmap represents a subsequence and color represents structural potential of this subsequence (from highly structured dark blue to highly non-structured dark red). We introduce the program RNASurface that efficiently reconstructs the surface of structural potential using a new fast algorithm for *Z*-score evaluation. This approach introduces an intuitively clear definition of locally optimal segments as peaks in the surface of structural potential. Besides identification of structured regions, the surface usage allows one to identify structured domains comprising the region. We apply our method to the genome of *Bacillus subtilis* and identify genome regions with high structural potential.

A highly accurate determination of structured intervals or regions of accessibility is useful for several purposes:

- preprocessing during *de novo* search for regulatory and non-coding RNAs;
- accurate definition of ncRNA boundaries; and
- correlation of genome-scale structural potential with other genomic features (gene boundaries, ribosome profiling, transcriptome data, binding sites of protein factors etc.).

2 METHODS

2.1 Surface of structural potential and definition of locally optimal sequence

Given a sequence S, S_{ij} denotes a subsequence from position i to position j and $E_{ij} = E(S_{ij})$ denotes its *MFE*. *MFE* calculation of S uses a dynamic programming algorithm. The optimal score E_{ij} is defined recursively through optimal scores of shorter subsequences by the Zuker algorithm (Zuker and P Stiegler, 1981) and is stored in the *MFE* matrix. This property of dynamic programming allows the algorithm to obtain and store optimal energies of each subsequence of S. Further, *Z*-score Z_{ij} of each subsequence S_{ij} is estimated based on a fast *Z*-score evaluation procedure and is stored in the *Z*-score matrix. Given a genome of size N, the upper bound L and the lower bound m of possible RNA sequence length, the surface of structural potential is defined as a trapezoid-shaped part of the *Z*-score matrix

$$\{Z_{ij} = \frac{E_{ij} - \mu(l)}{\sigma(l)} \mid l = j - i + 1, m \le l \le L\} \quad (2)$$

Parameters $\mu(l)$ and $\sigma(l)$ also depend on the dinucleotide content, this point is addressed later in the text (Sections 2.2.1 and 2.2.3). The MFE is a negative value, therefore more structured sequences have smaller Z-scores. A sequence is called locally optimal if small changes in its boundary coordinates only increase Z-score. Such sequences correspond to local surface peaks. Formally, S is called k-locally optimal, if and only if

$$\forall S' : |S' \cup S| - |S' \cap S| \leq k \Rightarrow Z(S') > Z(S)$$

Parameter k is interpreted as the amplitude of boundary changes. Increasing k leads to more global surface peaks and substantially reduces output size, but also causes skipping of subtle local peaks. Thus, k controls the relation between boundary accuracy and output size.

Efficient construction of the surface is discussed in Section 2.4, whereas Sections 2.2 and 2.3 describe preprocessing required for Z-score estimation and inference of surface properties, respectively.

Minor changes in the above definitions convert the problem to a form useful for the analysis of accessibility regions instead of structured regions. However, the methods described later in the text concern the analysis of structured regions.

2.2 Z-score estimation

Reconstruction of the structural potential surface requires developing a new strategy to Z-score estimation, as previous approaches are time-expensive. Among various measures, Z-score best reflects the structural potential (Freyhult *et al.*, 2005). The Z-score of an RNA sequence is calculated for energies of simulated sequences of the same length and average dinucleotide frequencies. Stacking interactions contribute most to the free energy formation; therefore simulations that preserve dinucleotide composition and possible stacking are of major importance (Workman and Krogh, 1999). Among various approaches to efficient evaluation of parameters μ and σ required for calculation of Z-score, RNAz 2.0 (Gruber *et al.*, 2010b) support vector regression (SVR) is considered to be the most efficient one. It uses several thousand support vectors with a radial kernel.

In our case, surface construction requires $\sim NL^2$ evaluations of Z-scores, where N is the genome length and L is the upper bound of a possible RNA length. SVR requires operations on thousands of support vectors to estimate each point. Therefore, it becomes a bottleneck of the algorithm. Instead, we use quadratic regression to fit the dependence of μ and σ on the dinucleotide composition. While preserving the quality of the approximation, this regression allows for fast recursive recalculation of parameters. Thus, although previous approaches focused on short L up to several hundred nucleotides, we can expand the segment length up to thousands of nucleotides.

2.2.1 Generalization of the segment dinucleotide content To estimate parameters of the MFE background distribution (μ and σ) of a target RNA segment, common practice is to use random sequences having the same average dinucleotide content as the target segment. In case of a functional ncRNA, structural constrains may influence its dinucleotide content (Schattner, 2002). Hence, bias of a dinucleotide content of a segment compared with its genomic background may be a consequence of selection on the secondary structure and represents an additional genomic signature (Smit *et al.*, 2009). We simulate sequences having the same dinucleotide content as the local genomic context of the target segment using the 1-order Markov model. We define genomic context of a segment S_{ij} as a region $S_{i_0j_0}$ of size d containing the segment with its flanking sequences $S_{i_0 i}$ and S_{jj_0}. The size d of a genomic context is recommended to be sufficiently large compared with the window size L to obtain a robust estimation of background dinucleotide frequencies. Z-score of a segment is estimated using the dinucleotide content of its genomic context.

If a segment size exceeds d, then its genomic context is the segment itself. Thus, option $d=0$ permits one to return to common practice without considering genomic background.

2.2.2 Approximation by sequence length The Z-score formula (2) contains the mean and standard deviation, which are dependent on sequence length l, thus estimation of these dependencies becomes central for the algorithmic complexity. Functions $\mu(l)$ and $\sigma(l)$ denote the mean and standard deviation of the energy distribution for a fixed average dinucleotide composition and variable length l. Earlier it has been observed (Chen *et al.*, 1990) and then proved (Clote *et al.*, 2005) that $\mu(l)$ is asymptotically linear.

Here, we performed extensive simulations up to sequence length of 4000 nt and found that at moderate length intervals, $\mu(l)$ is linear and $\sigma(l)$ perfectly fits as the root of length (Fig. 2).

To overcome approximation inaccuracy for large lengths, we fix 11 control lengths in the range from 60 to 1200 nt and linearly approximate $\mu(l)$ and $\sigma^2(l)$ for intermediate control lengths (for details see Supplementary Material S1).

2.2.3 Quadratic regression of the dinucleotide space For each of 11 control lengths, we apply quadratic regression to estimate the dependence of parameters on the dinucleotide composition. Following RNAz 2.0 (Gruber *et al.*, 2010), we constrain analysis to sequences of $G+C$ content and $G/(G+C)$, $A/(A+T)$ ratios ranging from 0.2 to 0.8 (the set of such sequences is denoted by Γ). Gruber *et al.* found that ~99% of the human ENCODE genome belongs to Γ and thus Γ covers nearly all functional elements. We further grid sequence region Γ on 27 subregions $(\Gamma_1, \ldots, \Gamma_{27})$, by dividing each direction $(G+C, G/(G+C), A/(A+T))$ into three parts: 0.2–0.4, 0.4–0.6 and 0.6–0.8. For each subregion Γ_i, we construct quadratic regressions of parameters on dinucleotide frequencies by generating 20 000 values of parameters μ and σ to fit the coefficients of quadratic regressions.

Summing up, we construct 27 quadratic regressions on dinucleotide frequencies for each of the 11 control lengths.

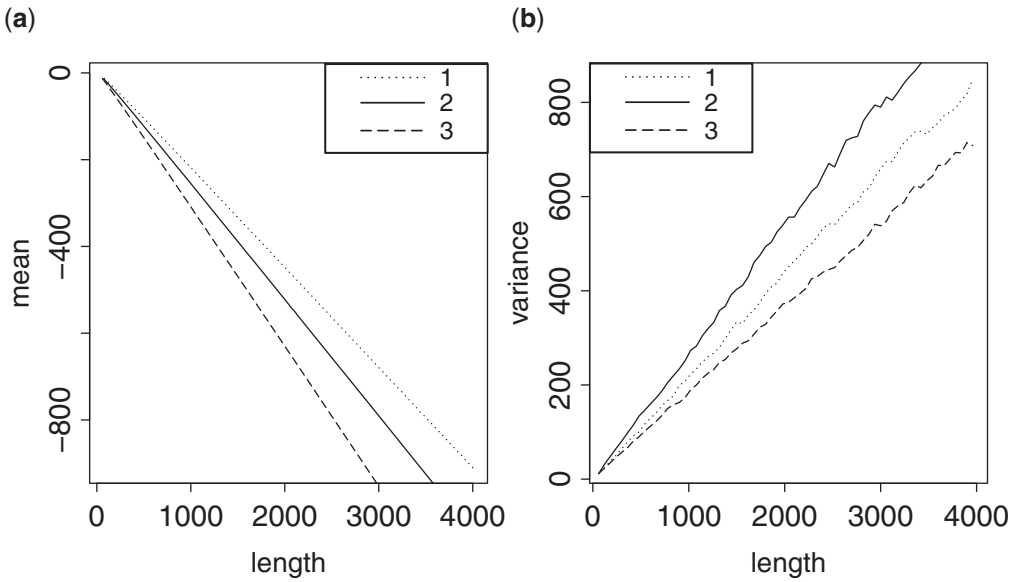
2.2.4 Quality of approximation We chose 24 000 random sequences with length ranging from 60 to 2400 nt from the human and *Drosophila* genomes (Supplementary Fig. S2). We compared Z-scores of these sequences obtained by our approximation and by sampling (2000 sequences with preserved average dinucleotide frequencies were sampled to estimate Z-score). Approximation and direct calculation show excellent concordance with $R^2=0.997$ (Supplementary Fig. S1a). The concordance remains high for lengths up to 2400 nt (Supplementary Fig. S1b). The average error of Z-score is 0.1. Moreover, the average error for small sequences (up to 250 nt) falls to 0.041, comparable with 0.076 of the SVR approach.

2.3 Inference surface properties

2.3.1 Smoothing the surface The surface of the structural potential is rugged in particular due to errors in Z-score estimation and experimental errors in measurement of thermodynamics parameters (Layton and Bundschuh, 2005) and intrinsic properties of the RNA folding landscape (Fontana *et al.*, 1993). Thus, significant secondary structure in a genome corresponds to many adjacent surface peaks, although only one peak is informative. We smooth the surface with an exponential kernel to emphasize one or few peaks among clusters of adjacent peaks and as a result to reduce the output size. The distance between two genome intervals (i, j) and (k, l) is defined as follows:

$$d[(i,j),(k,l)] = |i-k| + |j-l|$$

Each surface position Z_{ij} is smoothed by taking into account only positions Z_{kl} at the distance $d[(i,j),(k,l)]$ not exceeding a predefined parameter s with the weight $e^{-d[(i,j),(k,l)]\cdot\lambda}$.



Dinucleotide content, %

	1	2	3
aa	5.5	11	6.6
at	7.4	1.4	8.2
ag	5.5	11	1.6
ac	8.9	6.9	8.2
ta	6	5.5	1.6
tt	5.5	4.1	4.9
tg	6.4	2.7	9.8
tc	7.8	4.1	8.2
ga	9.2	8.2	14.8
gt	5.5	8.2	9.8
gg	3.3	11	6.6
gc	3.2	4.1	3.3
ca	6.4	5.5	3.3
ct	7.4	2.7	1.6
cg	6	6.9	9.9
cc	6	6.9	1.6

Fig. 2. Dependence of mean and variance of the energy distribution on sequence length. Line types denote dinucleotide content of sequences (see inset). (**a**) Mean energy nearly linearly decreases with length. (**b**) Variance (squared standard deviation) nearly linearly increases with length

The surface smoothed at s nucleotides Z_{ij}^s is derived from Z_{ij} as follows:

$$Z_{ij}^s = \frac{1}{C_s} \sum_{d[(i,j),(k,l)] \le s} Z_{kl} e^{-\lambda d[(i,j),(k,l)]},$$

where $C_s = \sum_{d[(i,j),(k,l)] \le s} e^{-\lambda d[(i,j),(k,l)]}$ is the total weight of positions adjacent to Z_{ij} at a distance less than s.

Larger s leads to a smoother surface. Discussion on selection of parameter λ and effective recursive calculation of the smoothed surface can be found in Supplementary Material S3.

2.3.2 One-dimensional plots Genome-wide analysis and correlation of biological features require simple and convenient representation and visualization of biological information, which is difficult to achieve with the 2D surface of the structural potential. Thus, based on the surface, we introduce two 1D measures, which are less informative compared with the surface, but easier to interpret and to correlate with various genome features.

The maximum squared Z-score (MZ) of a position is the maximum of squared Z-score among all sequences covering this position for sequence lengths not exceeding a threshold L:

$$MZ(i) = \max_{i-l=r-i,\, r-l+1 \le L} Z_{kl}^2 I\{Z_{kl} \le 0\}$$

where I is the indicator function. The plot of $MZ(i)$ is shown in the top panel of Figure 1.

Another measure reflects the density of locally optimal sequences $\rho_w(i)$ at each genome position i and may be used to discriminate islands of locally optimal sequence fragments:

$$\rho_w(i) = \frac{1}{w} \sum_{k,l} Z_{kl}^2 \cdot I\{S_{kl} \in \text{locally optimal output}\}$$
$$I\{i - w \le \tfrac{k+1}{2} \le i + w\}$$

where w is the maximum distance between position i and the center of a locally optimal sequence fragment.

RNASurface program implements efficient calculation of aforementioned measures. Application of plots to 100 kb region of *B.subtilis* is discussed in Supplementary Materials S7.

2.4 Algorithm

Given a genome S of size N and window L, RNASurface uses program RNAslider (Horesh *et al.*, 2009) that efficiently calculates *MFE* for each subsequence with length up to L. This program applies a sparsification technique (Wexler *et al.*, 2007) that speeds up the algorithm to $\sim NL$.

Given a segment S_{ij}, $\{f_k\}$ is a vector of dinucleotide frequencies of its genomic context $S_{i_0 j_0}$ (Section 2.2.1). Parameter $\mu(S_{ij})$ is estimated by quadratic regression (Section 2.2.3):

$$\mu(S_{ij}) = \sum_{1 \le k,l \le 16} f_k \cdot f_l \cdot c_{kl}, \quad (3)$$

where $\{c_{kl}\}$ are the coefficients of the quadratic regression.

Vectors of dinucleotide frequencies of genomic context differ by only two components between S_{ij} and a shifted segment $S_{i+1,j+1}$: S_{i_0, i_0+1} and S_{j_0, j_0+1}. Thus one can define a recursion based on (3).

The same equations hold for the standard deviation $\sigma(S_{ij})$. The recursion relations allow for fast recalculation of parameters. Given *MFE* and parameters, one can estimate Z-score immediately. Additional smoothing of the Z-score surface is implemented in time $\sim Ns$, where s is the smoothing parameter defined earlier in the text (for details see Supplementary Material S3). Further detection of peaks within the Z-score matrix is carried out by direct comparison of each matrix cell with adjacent cells.

3 RESULTS

We analyzed the genome of *B.subtilis* to benchmark the programs and to study regions of structure accumulation. This gram positive bacterium has numerous cis-acting regulatory RNAs and small RNAs. In addition, widespread rho-independent transcription termination (De Hoon *et al.*, 2005) and non-random prevalence of stem-loops structures (Petrillo *et al.*, 2006) make *B.subtilis* an object of specific research interest.

3.1 Performance comparison

There are >200 non-coding and regulatory cis-acting RNAs in *B.subtilis* subsp. *subtilis* str. 168 according to the Rfam database

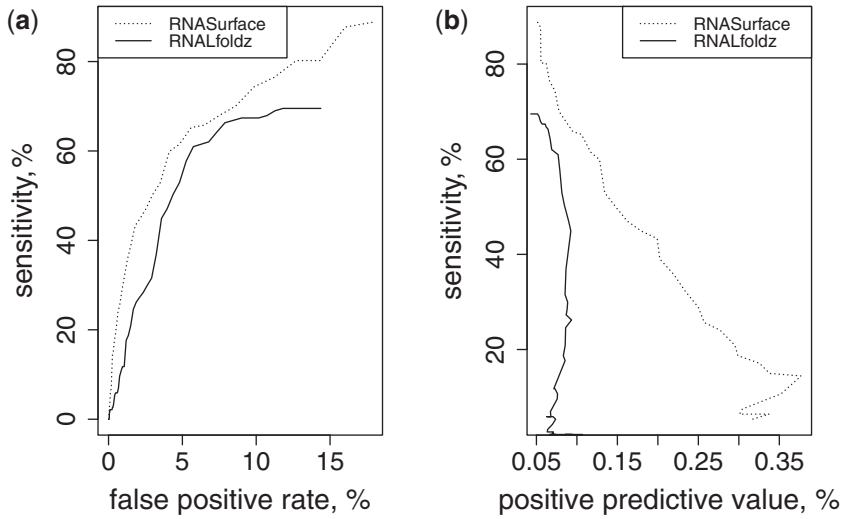


Fig. 3. Comparison of RNASurface and RNALfoldz using different measures. (**a**) Receiver-Operating Characteristic curve. (**b**) Sensitivity versus PPV

(Gardner *et al.*, 2011). Functional RNAs are more stable than random decoys. Thus, one expects that an energy-based program would discriminate them to some extent. To our knowledge, RNALfoldz is the only currently available program that scans genomes for significant RNA secondary structure using statistical information. We compare the quality of discrimination of RNASurface and RNALfoldz, whose outputs are lists of candidates. To accept a candidate as a true prediction, we assess the degree of overlap by the Jaccard measure. The Jaccard measure of two sequences S_1 and S_2 is defined as follows:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

A known functional RNA was considered as correctly predicted if there was an intersecting candidate sequence with the Jaccard measure of at least 75%. Nearly all functional RNAs in *B.subtilis* are <250 nt, thus we set the upper length $L = 250$ and the lower length $m = 50$. The results of programs comparison are presented as a Receiver-Operating Characteristic curve and the sensitivity versus positive predictive value (*PPV*) plot in Figure 3. Sensitivity, false-positive rate (*FPR*) and *PPV* are defined as follows:

$$PPV = TP/\text{predicted positive}$$
$$FPR = FP/\text{all negative}$$
$$\text{sensitivity} = TP/\text{all positive}$$

where *TP* is true positive and *FP* is false positive. To calculate the FPR one has to know the true negative (*TN*), whereas it is not clear how to identify and count non-functional RNA. Thus, we consider *TN* to be huge compared with *TP* and estimate *TN* by the genome length. RNASurface detects structural potential of functional RNA significantly better than RNALfoldz, especially at small *FPR*. Moreover, short RNASurface predictions are robust to increasing window size, whereas RNALfoldz predictions are gradually lost (Supplementary Fig. S2). The length of the majority of annotated RNA elements (87%) is less than 150nt. This decreases the prediction accuracy of RNALfoldz on substantially larger windows. A possible explanation is that the probability of the optimal structure or the base pair decreases with nucleotide distance (Lange *et al.*, 2012) and obeys the zeta-polymer property (Wexler *et al.*, 2007), which influence the prediction quality of closed structure and thus the quality of RNALfoldz predictions. On the other hand, RNASurface relies on the *Z*-score matrix that does not change for smaller length with increasing nucleotide distance.

The complexity of functional structures in genomes varies from simple small hairpins to complicated multidomain structures. We compared the discrimination ability of the programs with respect to the structure complexity. Rho-independent terminators were selected as a widespread example of simple functional hairpins. We used >2000 terminators predicted with accuracy of ~94% (De Hoon *et al.*, 2005). Riboswitches and T-boxes were selected as examples of complex structures. RNASurface identifies complex structures better than RNALfoldz (Supplementary Fig. S3a). Moreover, RNASurface identifies complex structures better than simple ones (Supplementary Fig. 3). In contrast to complex structures, terminators were discriminated weakly by both programs (Supplementary Fig. S3b). To confirm the aforementioned observation about influence of the structure complexity on performance, we classified non-coding (cis-regulatory) RNAs from Rfam by the number of stems comprising their structures (Supplementary Materials S5). The prediction accuracy of RNASurface improves with increasing number of stems in the structure and does not exceed the RNALfoldz performance for only simple hairpin structures (Supplementary Fig. S4). Complex structures often are not closed and are composed of several closed substructures, thus to interpret the results for closed structures, RNALfoldz output could benefit from additional post-processing.

3.2 Structural potential of ncRNAs and genomic regions in *B.subtilis*

We estimate the structural potential for different non-coding RNA classes: riboswitches, T-boxes, ribosomal protein leader

Table 1. Percent (number) of predictions for different types of RNA for three *Z*-score thresholds

Z-score	Riboswitch	T-box	L-leader	small RNA	tRNA	5S rRNA	FPR, %	PPV, %
−1	79 (34)	92 (12)	67 (4)	75 (15)	95 (81)	100 (20)	18	0.05
−2	65 (28)	85 (11)	50 (3)	65 (13)	62 (53)	35 (7)	5	0.1
−3	44 (19)	69 (9)	33 (2)	35 (7)	16 (14)	15 (3)	1	0.25
Total RNAs	43	13	6	20	85	20		

FPR, false-positive rate; PPV, positive-predictive value.

Table 2. Relative abundance of structured regions in various functional parts of the *B.subtilis* genome

Z-score	−2	−3	−4	−5
Coding regions	0.91 (148 441/162 599)	0.68 (21 050/30 963)	0.37 (2010/5399)	0.15 (153/1017)
Upstream regions	0.98 (6442/6601)	1.41 (1809/1280)	2.09 (480/230)	3.05 (131/43)
Downstream regions	2.55 (6166/2420)	5.68 (2793/492)	10.11 (950/94)	12.5 (225/18)
Intercoding regions	1.34 (16 448/12 249)	2.41 (5753/2387)	4.41 (1901/431)	12.52 (551/44)
Intercoding regions in operons	1.71 (274/160)	3.18 (105/33)	5.86 (41/7)	13 (13/1)

Note: Each column features relative abundance of structured segments in functional parts of the genome on condition on structure strength (*Z*-score threshold). Each row features the change of abundance in the corresponding region dependent on the *Z*-score threshold. First and second numbers in parentheses are the observed and expected number of predictions in a selected region with a selected *Z*-score threshold. Abundance is the ratio of these numbers.

regulatory RNAs (L-leaders), small RNAs, transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) (5S and PK-G12). Data were taken from Rfam database (Gardner *et al.*, 2011). Table 1 illustrates results for varying *Z*-score thresholds. Strict *Z*-score threshold of −3 demonstrate moderate rRNA (15%) and tRNA (16%) structural potential compared with cis-acting regulatory RNAs (33–69%) and small RNAs (54%).

Regulatory RNAs are located in specific regions of genomes due to their function. For example, some prokaryotes rely on RNA hairpin formation to terminate transcription and thus have numerous RNA secondary structures with high structural potential downstream of genes (Washio *et al.*, 1998). On the other hand, structures regulating transcription and translation (riboswitches, T-boxes, attenuators) occur primarily upstream of genes. We performed analysis of hits abundance in coding regions, intercoding regions and regions upstream and downstream of genes. Upstream regions were defined as segments from −200 to 50 nt from the start codons, downstream regions were defined as segments from −50 to 150 nt from the stop codons and intercoding regions were defined as segments between two coding regions. To estimate the abundance of structured RNAs in genome regions, we consider a background model, when predictions are distributed uniformly throughout the genome. The observed number of predictions was compared with the expected number of predictions in the background model (Table 2). A considerable overrepresentation of significant structures was observed upstream and downstream of genes. This result may be partially explained by the existence of numerous cis-regulatory RNAs and rho-independent terminators in *B.subtilis*, some of which show high structural potential according to RNASurface predictions (predictions of 200 cis-regulatory RNAs and 434 terminators have *Z*-score less than −2). Strong 12-fold enrichment in the intercoding regions is partially explained by enrichment in the upstream and downstream regions, but excluding upstream and downstream regions still retains >4-fold enrichment. This can potentially be induced by small non-coding RNAs (Irnov *et al.*, 2010; Saito *et al.*, 2009). Few hits were found in the coding regions, although these regions also show signs of selection toward secondary structures (Katz and Burge, 2003). Methodology of *Z*-score estimation does not include codon-usage statistics, which could produce additional selection on RNA secondary structure in coding regions (Park *et al.*, 2013). Thus, our results could underestimate the structural potential of coding regions. However, accurate dissection of selection on protein-coding and structural properties of a coding sequence requires extensive simulations (Zhang *et al.*, 2013) or usage of homologous sequences (Pedersen *et al.*, 2004).

3.3 Time and space requirement

RNASurface requires $O(N(L+s))$ time. Practically, time performance of RNASurface was compared with RNAslider and RNALfold. The *B.subtilis* genome was taken as an input sequence. Execution time as a function of window length up to 600 showed similar results for RNASurface and RNAslider, and several-fold less than for RNALfold (Supplementary Fig. S5). Additional parameters of RNASurface were fixed at the following values: dinculeotide frame $d=600$, smoothing parameter $s=1$ and locally optimal parameter $l=1$. Varying parameters yielded up to 3-fold time increase for d ($d=0$; $L=100$, 300), 2-fold time increase for s ($s=10$; $L=100$, 300) and no increase in time for l in worst cases ($l \geq 15$; $L=100$, 300).

RNASurface requires $O(L^2)$ memory.

4 CONCLUSION

Here, we define and calculate the surface of structural potential. This yields a new definition of locally optimal sequence regions as peaks in this surface. We also introduce several one-dimensional measures that reflect unusually structured regions. An efficient method of surface construction has been implemented in the program RNASurface that uses a novel approach to estimating the significance of RNA secondary structure. RNASurface has the same time requirements as well-known scanning programs that search for secondary structures (RNALfold and RNAslider). We also have implemented a web server that provides visualization tools to explore the surface of structural potential, the detailed description of service capabilities will be reported elsewhere.

Application to *B.subtilis* demonstrated better abilities of RNASurface to discriminate known structured RNAs from random regions. Moreover, we showed that the regions upstream and downstream of genes are significantly enriched in structured motifs.

Thus, surface exploration by RNASurface can bring new understanding of structured patterns in genomes.

ACKNOWLEDGEMENT

The authors are grateful to Mikhail Gelfand for useful discussion and valuable comments on the manuscript.

Funding: Ministry of Education and Science of Russian Federation (projects 8283, 8049) and the Russian Foundation for Basic Research (grant 12-14-91333).

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Bonnet,E. *et al.* (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

Chen,J.H. *et al.* (1990) A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, **6**, 7–18.

Clote,P. *et al.* (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.

De Hoon,M.J. *et al.* (2005) Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comp. Biol.*, **1**, e25.

Djebali,S. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.

Fontana,W. *et al.* (1993) RNA folding and combinatory landscapes. *Phys. Rev. E*, **47**, 2083–2099.

Freyhult,E. *et al.* (2005) A comparison of RNA folding measures. *BMC Bioinformatics*, **6**, 241.

Gardner,P.P. *et al.* (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.*, **39**, D141–D145.

Giegerich,R. *et al.* (2004) Abstract shapes of RNA. *Nucleic Acids Res.*, **32**, 4843–4851.

Gruber,A.R. *et al.* (2010a) RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures. *Lect. Notes Inform.*, **173**, 12–21.

Gruber,A.R. *et al.* (2010b) RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, **15**, 69–79.

Gultyaev,A.P. *et al.* (2010) Influenza virus RNA structure: unique and common features. *Int. Rev. Immunol.*, **29**, 533–K556.

Hofacker *et al.* (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.

Horesh,Y. *et al.* (2009) RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics*, **10**, 76.

Irnov *et al.* (2010) Identification of regulatory RNAs in *Bacillus subtilis*. *Nucleic Acids Res.*, **38**, 6637–6651.

Katz,L. and Burge,C.B. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.*, **13**, 2042–2051.

Keller,T.E. *et al.* (2012) Reduced mRNA secondary-structure stability near the start codon indicates functional genes in prokaryotes. *Genome Biol. Evol.*, **4**, 80–88.

Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.

Lange,S.J. *et al.* (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res.*, **40**, 5215–5226.

Layton,D.M. and Bundschuh,R. (2005) A statistical analysis of RNA folding algorithms through thermodynamic parameter perturbation. *Nucleic Acids Res.*, **33**, 519–524.

Le,S.Y. *et al.* (1988) A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, **4**, 153–159.

Le,S.Y. *et al.* (2003) Discovering well-ordered folding patterns in nucleotide sequences. *Bioinformatics*, **19**, 354–361.

Lee,J.T. (2012) Epigenetic regulation by long noncoding RNAs. *Science*, **338**, 1435–1439.

Lorenz *et al.* (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.

Martin,K.C. and Ephrussi,A. (2009) mRNA Localization: gene expression in the spatial dimension. *Cell*, **136**, 719–730.

Mathews,D.H. *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 7287–7292.

Mathews,D.H. *et al.* (2010) Folding and finding RNA secondary structure. *Cold Spring Harb. Perspect. Biol.*, **2**, a003665.

Park,C. *et al.* (2013) Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl Acad. Sci. USA*, **110**, E678–E686.

Pedersen,J.S. (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.*, **32**, 4925–4936.

Pervouchine,D.D. *et al.* (2012) Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. *RNA*, **18**, 10–15.

Petrillo,M. *et al.* (2006) Stem-loop structures in prokaryotic genomes. *BMC Genomics*, **7**, 170.

Saito,S. *et al.* (2009) Novel small RNA-encoding genes in the intergenic regions of *Bacillus subtilis*. *Gene*, **428**, 2–8.

Schattner,P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Res.*, **30**, 2076–2082.

Smit,S. *et al.* (2009) RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic Acids Res.*, **37**, 1378–1386.

Tinoco,I. and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.

Vitreschak,A. *et al.* (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.*, **20**, 44–50.

Wan,X.F. *et al.* (2006) Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. *J. Bioinform. Comput. Biol.*, **4**, 1015–1031.

Washietl,S. *et al.* (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Washio,T. *et al.* (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res.*, **26**, 5456–5463.

Wexler,Y. *et al.* (2007) A study of accessible motifs and RNA folding complexity. *J. Comput. Biol.*, **14**, 856–872.

Workman,C. and Krogh,A. (1999) No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.*, **27**, 4816–4822.

Zhang,Y. *et al.* (2013) SPARCS: a web server to analyze (un)structured regions in coding RNA sequences. *Nucleic Acids Res.*, **41**, W480–W485.

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.