Published online ??

Nucleic Acids Research, ??, Vol. ??, No. ?? ***1–6***
doi:10.1093/nar/gkn000

SPARCS: a web server to analyze (un)structured regions in coding RNA sequences

Yang Zhang [1,2,*], Yann Ponty [3,4,*], Mathieu Blanchette [1], Éric Lecuyer [2,5,†], Jérôme Waldispühl [1,†]

[1] School of Computer Science & McGill Centre for Bioinformatics, McGill University, Montréal, QC, Canada, [2] Institut de Recherches Cliniques de Montréal (IRCM), Montréal, QC, Canada, [3] Laboratoire d'Informatique (LIX) – CNRS UMR 7161, École Polytechnique, 91128 Palaiseau, France, [4] AMIB team/project, INRIA Saclay, Batiment Alan Turing, 91128 Palaiseau, France and [5] Département de Biochimie, Université de Montréal, Montréal, QC, Canada.

Received ??; Revised ??; Accepted ??

ABSTRACT

More than a simple carrier of the genetic information, mRNA coding regions can also harbour functional elements that evolved to control different post-transcriptional processes, such as mRNA splicing, localization and translation. Functional elements in RNA molecules are often encoded by secondary structure elements. In this paper, we introduce **SPARCS** (Structural Profile Assignment of RNA Coding Sequences), an efficient method to analyze the (secondary) structure profile of protein coding regions in mRNAs. First, we develop a novel algorithm that enables us to sample uniformly the sequence landscape preserving the dinucleotide frequency and the encoded amino acid sequence of the input mRNA. Then, we use this algorithm to generate a set of artificial sequences that is used to estimate the Z-score of classical structural metrics such as the sum of base pair probabilities and the base pair entropy. Finally, we use these metrics to predict structured and unstructured regions in the input mRNA sequence. We applied our methods to study the structural profile of the *ASH1* genes, and recovered key structural elements. A web server implementing this discovery pipeline is available at http://csb.cs.mcgill.ca/nasp together with the source code of the sampling algorithm.

INTRODUCTION

Sequence analysis in the post-genomic era has revealed the multiplicity of selective pressures applied on the genetic code and therefore a frequent overlap of functional elements. Recent studies suggested that coding regions of messenger RNAs can often include secondary structure elements involved in post-transcriptional regulatory processes (1, 2, 3). While many programs have been developed to analyze folding properties of large non-coding RNAs (4) or untranslated regions of mRNAs (5), these tools cannot be directly applied to study the structural properties in coding regions. Indeed, the sequence of codons that specify the amino acid chain might bias the thermodynamic folding properties of the polynucleotide, thus preventing accurate estimate of the statistical significance of local structural motifs. Similar issues are encountered in the context of large scale studies and techniques aiming as defining RNA structure characteristics on a genome-wide scale (6, 7). Actually, assessing the statistical significance of observed phenomena or patterns requires the definition of a reliable and expressive background model (a.k.a. the null hypothesis). In particular, any sequence property that is a natural consequence of a well-understood mechanism should be captured by the background model, so that it will generically appear in random sequences. Including these features in the background model will lead to an increased statistical significance for *novel* phenomena.

A classic exploratory approach starts with a random generation of sequences that share similar properties as a reference set of sequences. Various metrics can then be evaluated, possibly leading to diverging distributions of values within the random and reference sets. The significance of such an observation can be empirically assessed using classic statistical tools (Z-score, P-value...) . To implement such an approach in the context of mRNAs, one must restrict random sequences to synonymous sequences (i.e. the set of sequences that encode the same amino acid sequence). Such sequences can trivially be generated, uniformly at random, by simply choosing, for each amino acid, one of its alternate codons. Another constraint, essential when analyzing structural properties of RNA molecules, is the preservation of the overall dinucleotide frequencies (DF). Such a constraint has been popular in the field of RNA bioinformatics following the study of Workman and Krogh (8), and builds on the rationale that preserving the DF maintains the feasibility of stacking base pairs, arguably the main contributor to RNA stability. Efficient methods have been proposed for such a model, drawing an analogy with the random generation of a Euler path in a De Bruijn-like graph, whose edges represent the dinucleotides (9, 10).

When attempting to infer an evolutionary pressure from the observation of structural features within mRNA sequences,

*Both authors contributed equally to this work.
†To whom correspondence should be addressed. Tel: +1-514-398-5018; Fax: +1-514-398-3883; Email: jeromew@cs.mcgill.ca; eric.lecuyer@ircm.qc.ca

© ?? The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

both constraints should ideally be satisfied. Unfortunately, the algorithms used to capture these two constraints rely on radically different principles, and cannot be easily combined into an algorithm that would, at the same time, preserve the dinucleotide frequency and an amino acid sequence. For this reason, Katz and Burge proposed `DiCodonShuffle` (11), a heuristic algorithm based on a swapping procedure, which repeatedly exchanges codons while preserving the DF. As shown by Shabalin *et al* (12), such a model preserves the periodic pattern of base-pair frequencies observed within coding regions of mRNAs. However, this method is only *asymptotically* uniform, and a bias towards certain sequences may be anticipated in samples produced in finite time (depending on the initial sequence and the number of swaps). Furthermore, as noted by the authors, the codon/DF preserving swaps may disconnect the underlying Markov chain, causing some legit sequences to be completely inaccessible by the sampling procedure. The impact of such limitations turned out to be more than purely theoretical, and we observed (see Figure 1) that the diversity (indicated by the sequence entropy) of generated sequences was much lower for `DiCodonShuffle` than for our truly uniform procedure, indicating a substantial bias in the method.

In this paper, we introduce **SPARCS** (Structural Profile Assignment of RNA Coding Sequences) a web server that predicts structured, unstructured and disorder regions in coding RNA sequences. Building on recent algorithmic advances (13, 14), we developed a novel sampling algorithm that enables us to sample *uniformly* random sequences preserving the encoded protein sequence as well as the dinucleotide frequencies. Combined with multiple classical metrics (e.g. base pair probabilities and base pairing entropy), this sampling algorithm enables the calculation of accurate Z-scores Z-scores and the prediction of strongly and weakly structured regions, along with disordered regions in exons – an insight that could not be fully achieved using previously existing sampling techniques.

SPARCS takes as input the coding region of an mRNA and proceeds in two steps. First, it generates a set of random sequences preserving the encoded amino acid sequence and the dinucleotide frequency of the input sequence. Next, it uses `RNAplfold` (4) to compute thermodynamic properties (e.g. the sum of base pair probabilities, base pairing entropy) of each sequence (input RNA and random samples), and compare these metrics to calculate, for each position in the input sequence, a Z-score estimating the statistical significance of the secondary structure profile.

SPARCS outputs a graph showing the Z-score of the sum of base pair probabilities and the base pairing entropy. It also provides a list of segments with predicted strongly and weakly structured segments. In addition, is also predicts disordered regions (i.e. regions with multiple suboptimal structures). In order to conduct further analysis, the user can also download the set of random sequences generated with **SPARCS**. The web server and the source code are available at http://csb.cs.mcgill.ca/nasp.

METHOD OVERVIEW

The methodology of **SPARCS** is to combine the following procedures, starting from a given RNA sequence:

1. Use a novel statistical sampling algorithm to generate a set of random sequences that preserves both the encoded amino acid sequence and the DF of the input sequence.
2. Use `RNAplfold` (4) to compute thermodynamical properties of the input sequence and all random sequences generated.
3. Predict regions that are significantly structured, unstructured and disordered, based on a comparison of thermodynamic properties between input and random sequences.

Multivariate Boltzmann sampling of protein-coding sequences under dinucleotide frequency constraint

Our approach builds on a multivariate Boltzmann sampling scheme, initially introduced in the context of enumerative combinatorics (13), and previously applied to control the GC-content of sampled RNA sequences within the RNAmutants software (14). This approach initially relaxes the goal of preserving the DF, and draws sequences that strictly preserve the amino acid sequence while only achieving, on the average, the prescribed DF. A further rejection of unsuitable sequences, whose DF differ too much from the targeted DF, filters the generated sequences, reestablishing the uniformity within the selected subset. The produced sequences therefore feature both correct DF and coding capacity while being generated with uniform probability.

Namely, let $\mathbf{S}$ be an amino acid sequence, and $\mathbf{dc}^* = (\mathsf{dc}^*_{\mathsf{AA}}, \mathsf{dc}^*_{\mathsf{AC}}, \mathsf{dc}^*_{\mathsf{AG}} \ldots)$ be the vector of targeted dinucleotide frequencies, the algorithm repeats the following steps until the desired number of samples is reached:

1. Draw a set of structures encoding $\mathbf{S}$, with respect to a weighted distribution;
2. Estimate expected DF from sample;
3. Collect suitable sequences;
4. Update weights to match expected DF with target.

Weighted distribution. We associate a weight π_{XY} to each dinucleotide XY. This weight is inherited multiplicatively by any RNA sequence in $\mathsf{rna}(\mathbf{S})$, the set of sequences compatible with a targeted amino acid sequence $\mathbf{S}$. This implicitly defines a probability distribution over $\mathsf{rna}(\mathbf{S})$ where any RNA sequence $w \in \mathsf{rna}(\mathbf{S})$ has probability

$$\mathbb{P}(w) = \frac{\pi(w)}{\sum_{w' \in \mathsf{rna}(\mathbf{S})} \pi(w')} \text{ with } \pi(w) = \prod_{i=1}^{|w|-1} \pi_{w_i . w_{i+1}}.$$

Quite importantly, it should be noted that any pair of sequences having an equal DF will also have equal relative probability. Therefore, generating a set of sequences, and

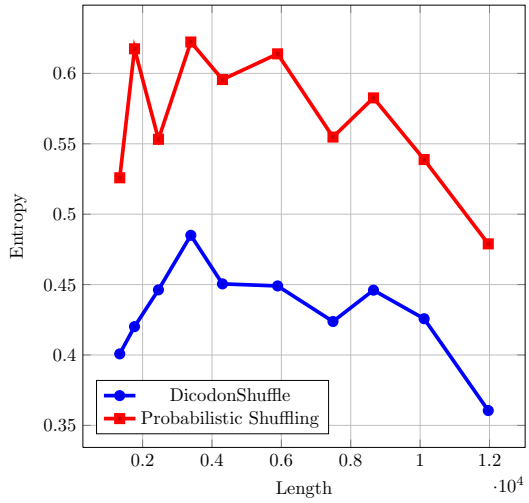


Figure 1. Entropy comparison between sequences generated by `DiCodonShuffle` (11) and our probabilistic shuffling method. For both methods, 1000 sequences are generated and, for **SPARCS**, the relative tolerance was set to $\varepsilon = 10^{-1}$. Sequences produced using `DiCodonShuffle` show much less diversity than those generated using **SPARCS**, either indicating a substantially limited accessibility of compatible sequences, or a substantial bias (non-stationarity) due to the bounded nature of their random walk.

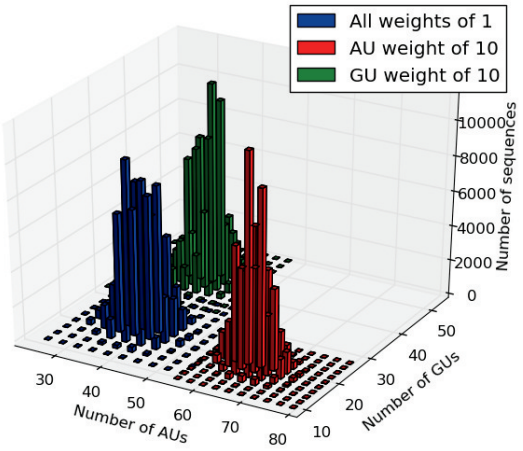


Figure 2. Impact of weighted distribution on the number of occurrences of dinucleotides AU and GU. Either in the uniform distribution ($\pi(\mathsf{XY}) = 1$, blue), or setting larger weights to AU ($\pi(\mathsf{AU}) = 10$, red) or GU ($\pi(\mathsf{GU}) = 10$, green), 100 000 sequences compatible with an mRNA sequence encoding 179 amino acid (the first two exons of *oskar* gene in *D. melanogaster*) were randomly generated. The concentration of the distribution, and the shift in expected DF observed for different weights, are the key ingredients of our method, allowing for an efficient approach based on adaptive sampling.

retaining only the sequences that do feature the targeted DF, gives a unbiased set of sequences. This property also holds for sequences generated using different weights, therefore they can be gathered across different iterations of the adaptive sampling without introducing a bias.

Self-adaptive calibration of weights. The weighting scheme may be used to *shift* the expected number of occurrences of each dinucleotide, as illustrated by Figure 2. Let us denote by V_{XY} the number of copies of XY in a random sequence generated in the weighted model. For instance, setting π_{GU} to 0 will cancel the probability of any sequence featuring any occurrence of GU, and the expected number of GU will therefore drop to 0. Conversely, setting $\pi_{\mathsf{GU}} \to +\infty$ will only grant positive probability to sequences that maximize the number of copies of GU.
To find a weight that matches the expected DF with the targeted one, we use a heuristic strategy to figure out weights that achieve, on the average, the targeted DF. To that purpose, we initially set $\pi(\mathsf{XY}) := \mathrm{dc}^*_{\mathsf{XY}}$ and, after each iteration of the adaptive sampling, we update each weight to $\pi(\mathsf{XY}) \cdot \mathrm{dc}^*_{\mathsf{XY}}/\mu_{\mathsf{XY}}$, where μ_{XY} is the expected value for V_{XY}, estimated from the sample. The process typically converges after a few iterations, leading to a good approximation of the best weight set.

Random generation. To draw a sequence of $\mathsf{rna}(\mathbf{S})$ within the weighted distribution, one needs to choose a compatible codon for each of the n amino acid in $\mathbf{S}$. Such choices cannot be made independently, since the overlap between consecutive codons contributes to an additional dinucleotide, ultimately impacting the weight of a generated sequence.
Following the general principle of the recursive approach for random generation (15, 16), we precompute the total weight $\mathcal{Z}_{b,i+1}$ of every sequence accessible upon choosing some codon ending with a base $b \in [\mathsf{A},\mathsf{U},\mathsf{C},\mathsf{G}]$ at the i-th position. Such weights can be efficiently computed using dynamic programming based on

$$\mathcal{Z}_{b,n+1} = 1 \quad \text{and} \quad \mathcal{Z}_{b,i} = \sum_{c \in \mathsf{cod}(\mathbf{S}_i)} \pi(b.c) \cdot \mathcal{Z}_{c_{-1},i+1}$$

where $\mathsf{cod}(\mathbf{S}_i)$ is the set of codons compatible with the i-th amino acid in $\mathbf{S}$, and c_{-1} is the last nucleotide of c. Since the first amino acid ($i = 1$) is not preceded by any nucleotide, it must be treated slightly differently by setting $b := \varnothing$ and $\pi(\varnothing.c) := 1$.
During the random generation, these precomputations are used to assign probabilities to each of the possible codons, such that each sequence is generated which respect to the weighted distribution. Namely, one picks a codon $c \in \mathsf{cod}(\mathbf{S}_i)$ for the i-th amino acid, in the context of a previous nucleotide b, with probability

$$p_{b,c,i} = \frac{\pi(b.c) \cdot \mathcal{Z}_{c_{-1},i+1}}{\mathcal{Z}_{b,i}}.$$

The sampling algorithm starts on the first codon ($i := 1$ and $b := \varnothing$), and iterates over the amino acid sequence $\mathbf{S}$ in increasing order, picking a codon with the above probabilities, and updating b to the last nucleotide in the elected codon. After picking the last codon, it can be shown that the generated sequence is indeed in $\mathsf{rna}(\mathbf{S})$, and has probability which is proportional to its weight (cf supp. mat.). The complexity of the algorithm is in $\Theta(k \cdot n)$ time and space for sampling k sequences, each consisting of n codons.

Overall time and space requirement. We empirically observed, and could formally prove using Drmota

theorem (17) for non-degenerate cases, that V_{XY} asymptotically follows a Normal law of mean in $\Theta(n)$ and standard deviation σ_{XY} in $\Theta(\sqrt{n})$. Furthermore, the covariations between numbers for different dinucleotides remains provably limited, and the joint distribution of the V_{XY} for every dinucleotide XY asymptotically follows a 16-variate Normal law. Consequently, the probability of generating a sequence having expected DF scales like $\Theta(n^{-16/2})$ and it takes, on the average, $\Theta(n^8)$ attempts to obtain such a sequence. The average-case complexity of a rejection procedure for the uniform sampling is in $\Theta(k\cdot n^9)$ time, after a linear time and space preprocessing.
Such a large time complexity may be impractical for real-life applications. However, if a small relative tolerance $\varepsilon\in\Theta(1/\sqrt{n})$ is allowed on every targeted dinucleotide count, leading any sequence w to be accepted if its dinucleotide counts are such that

$$(1-\varepsilon)\cdot\mathsf{dc}^*_{\mathsf{XY}}\leq\mathsf{dc}_{\mathsf{XY}}(w)\leq(1+\varepsilon)\cdot\mathsf{dc}^*_{\mathsf{XY}},$$

for every dinucleotide XY. Under this setting, the probability of acceptance only decreases like $o(C^{-16})$, where C is a constant which only depends on the covariance matrix. In particular, if

$$\varepsilon=3\cdot\max_{\mathsf{XY}}(\sigma_{\mathsf{XY}}/n)\in\Theta(1/\sqrt{n})\quad\text{(The 3 \textit{std.-dev.} rule)},$$

then the probability of acceptance becomes greater than $0.99^{16}\approx 85\%$, and the average-case complexity of the method becomes asymptotically equivalent to $\Theta(k\cdot n)$, at the cost of loss of uniformity which is typically negligible, and can be efficiently corrected through a post-processing step (13).

Secondary structure prediction

The secondary structures of both the input mRNA sequence and random sequences are predicted using the `RNAplfold` software, distributed within the Vienna RNA package (18). `RNAplfold` considers all possible locally stable secondary structures for an input RNA sequence, and calculates base pair probabilities, assuming a Boltzmann equilibrium. As recommended by Lange *et al.* (19), we use a window size of $W+50$ nucleotides ($W=150$ by default), and retain only those base pairs separated by at most W positions, and set a base pair probability cut off threshold to 0.1.

Characterization of the structural profile

We screen the input sequence with a sliding window of W nucleotides and evaluate the standardized score (Z-score) for each window w on two classical metrics: $\mathcal{B}(w)$ the sum of base pair probabilities and $\mathcal{H}(w)$ the base pair entropy. Let $\mathcal{C}$ be the set of all valid base pairs in the sliding window and $p_{i,j}$ the probability of a base pair (i,j). We define the **sum of base pair probabilities** as the sum of all base pair probabilities assessed by `RNAplfold` within the frame, such that

$$\mathcal{B}(w)=\sum_{(i,j)\in\mathcal{C}}p_{i,j}.$$

The sum of base pair probability estimates the stability of the secondary structures in the conformational landscape and thus quantifies the structural potential of the sequence.

Similarly, we define the **base pair entropy** as the Shannon entropy of the base pair probabilities, such that

$$\mathcal{H}(w)=-\sum_{(i,j)\in\mathcal{C}}p_{i,j}\cdot\log(p_{i,j}).$$

The base pair entropy aims to evaluate whether many alternate sub-optimal structures exist in the conformational landscape. For each nucleotide position, the Z-scores of all windows are averaged out to give the structural profile at a single nucleotide resolution.

We use these metrics to characterize a structural profile consisting in three, mutally-exclusive, types of regions, based on two user-defined thresholds $t_{\mathcal{B}}$ and $t_{\mathcal{H}}$:

- **Structured regions:** A region is said to be *structured* when the Z-score of the base pair probability exceeds $t_{\mathcal{B}}$ and the Z-score of the base pair entropy is lower than $-t_{\mathcal{H}}$. This configuration indicates stable structures with few competitors.
- **Unstructured regions:** A region is *unstructured* when the Z-score of the base pair probability and the Z-score of the base pair entropy are respectively lower than $-t_{\mathcal{B}}$ and $-t_{\mathcal{H}}$. In that case, the energy landscape is *flat* with no dominant structure.
- **Disordered regions:** A region is *disordered* when the Z-score of the base pair probability and the Z-score of the base pair entropy respectively exceed $t_{\mathcal{B}}$ and $t_{\mathcal{H}}$. This configuration suggests the presence of multiple stable and competing structures in the conformational landscape.

By default, **`SPARCS`** uses thresholds on the Z-score of 0.2 to discriminate high or low values. As illustrated in the next section, these settings aim to classify structural domains in the input sequences. Nonetheless, more stringent values can be specified, for instance if the user wishes to detect strongly (un-)structured regions.

Analysis of Ash1 gene in yeast

We illustrate the insights brought by **`SPARCS`** on the well-studied *ASH1* gene in yeast. Using mutagenesis studies and comparative sequence analysis, four functional elements have been identified in the *ASH1* mRNA. Each of them has been shown to be sufficient to localize a reporter mRNA to the bud of dividing yeast cells (20). Out of the four elements, three (E1,E2A and E2B) are located within the coding region of *ASH1* mRNA.

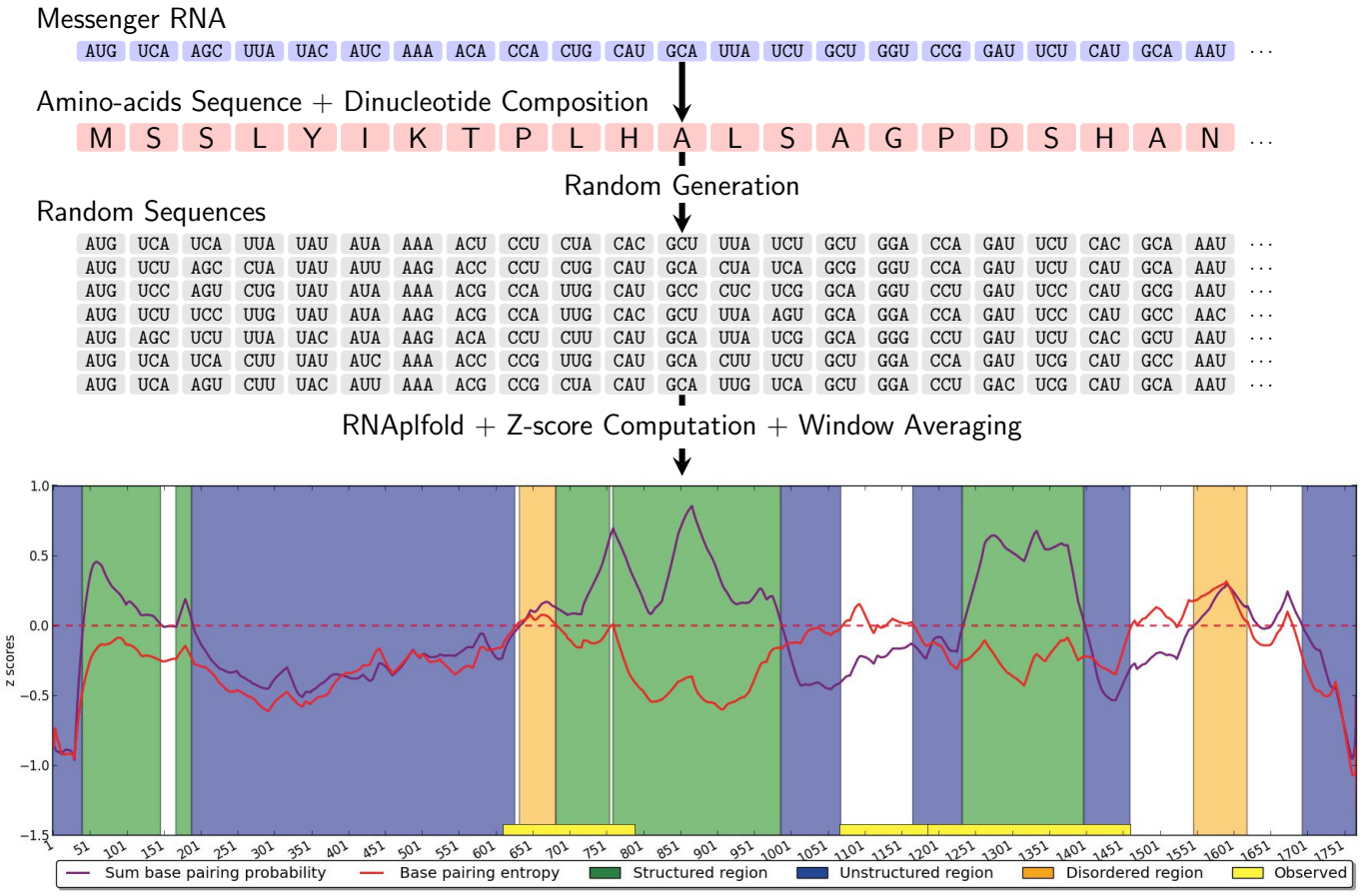


Figure 3. Analysis of the protein-coding region of the *ASH1* gene in yeast. The Z-scores of the base pair probability are represented in magenta and those of the base pair entropy in red. Structured, unstructured and disordered regions are displayed in green, blue and orange, and the functional elements E1, E2A and E2B are indicated at the bottom of the figure with yellow boxes. Dashed lines show the thresholds for determining high or low Z-score values.

Figure 3 shows the output of **SPARCS** for the *ASH1* mRNA coding region . The Z-scores of the sum of base pair probabilities are represented in magenta and those of the base pair entropy in red. Structured, unstructured and disordered regions are displayed in green, blue and orange, and the functional elements E1, E2A and E2B are indicated at the bottom of the figure with yellow boxes. As mentioned above, here we aim to detect structural domains and tendencies in the structural profile rather than focusing on the prediction of single elements. Therefore, we used a threshold of 0.

Our results show that the E1 (positions 625 to 775) match predicted disordered and structured regions. The presence of disordered region at the beginning of the element could be explained by the presence of internal loops and alternate base pairings in the predicted secondary structure (See Gonzales *et al* (21) and Chartrand *et al* (20)). Interestingly, the elements E2A (positions 1081 to 1199) and E2B (positions 1200 to 1447) are both surrounded by unstructured regions, possibly to avoid interactions between these elements. Noticeably, unlike the E2A element, the E2B element is particularly stable and structured. Outside these functional segments, we identify a large unstructured region (from 200 to 600) before the E1 element which could help to stabilize the E1 element or, hypothetically, to facilitate translation. By contrast, we identify a strongly structured regions between the E1 and E2 elements. This prediction could reveal a buffer that aims to prevent these elements interacting. Finally, our analysis also suggests a structured region at the beginning of the sequence (positions 50 to 200). To the best of our knowledge, this region has not been experimentally studied, motivating further comparative studies.

NASP SERVER

The **SPARCS** web server takes an RNA/DNA sequence or a FASTA file as input. Upon validation, a first set of 1 000 random sequences, preserving both the DF and encoded animo acid sequence of the input sequence, is generated. A second set of 1 000 random sequences, called the uniform model, is generated to preserve only the amino acid sequence. The input sequence and the 2 000 random sequences are then fed to `RNAplfold` to predict their base pairing properties. The Z-score is computed for a sliding window of user-specified width (defaulting to 150 nts), and all Z-scores are averaged for every position to evaluate the statistical significance of the secondary structure profile.

SPARCS finally outputs a single Z-score plot based on our metrics: sum of base pair probabilities, base pair entropy, structural potential region, unstructured potential region and

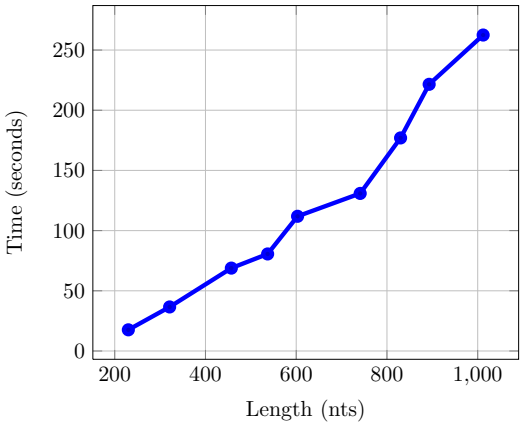


Figure 4. Typical runtime of **SPARCS** for sequence lengths varying from 200 to 1000 nts.

disordered potential region. The dashed line(s) indicates the Z-score thresholds for the sum of base pair probabilities and base pair entropy, respectively. Users may specificy custom thresholds for both of the base pair probability and base pair entropy metrics.

SPARCS runs on a server hosted at University McGill, which has 8 cores and has a total of 63 GB of memory. Each core is an Intel(R) Xeon(R) CPU X5570 at 2.93GHz, with 8192 KB cache. Figure 4 shows the overall runtime on the server as a fonction of the mRNA length, for mRNA sequences ranging from 200 to 1000 nts, and reveals a linear trend.

FUNDING

This work was supported by an IRCM masters scholarship (YZ); the Natural Sciences and Engineering Council of Canada (YZ, JW and EL); and the French *Agence Nationale de la Recherche* through the MAGNUM project [ANR 2010 BLAN 0204 to YP].

REFERENCES

1. Hui Chen and Mathieu Blanchette. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol*, 7 Suppl 1:S9, 2007.
2. Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of rna secondary structure in yeast. *Nature*, 467(7311):103–7, 2010.
3. Michael F Lin, Pouya Kheradpour, Stefan Washietl, Brian J Parker, Jakob S Pedersen, and Manolis Kellis. Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes. *Genome Res*, 21(11):1916–28, 2011.
4. Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–5, Mar 2006.
5. Michal Rabani, Michael Kertesz, and Eran Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc Natl Acad Sci U S A*, 105(39):14885–90, 2008.
6. Joseph M Watts, Kristen K Dang, Robert J Gorelick, Christopher W Leonard, Julian W Bess, Ronald Swanstrom, Christina L Burch, and Kevin M Weeks. Architecture and secondary structure of an entire hiv-1 rna genome. *Nature*, 460(7256):711–716, Aug 2009.
7. Robert C Spitale, Pete Crisalli, Ryan A Flynn, Eduardo A Torre, Eric T Kool, and Howard Y Chang. Rna shape analysis in living cells. *Nat Chem Biol*, 9(1):18–20, Jan 2013.
8. Christopher Workman and Anders Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822, 1999.
9. Walter M. Fitch. Random sequences. *Journal of Molecular Biology*, 163(2):171 – 176, 1983.
10. Stephen F Altschul and Bruce W Erickson. Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Molecular Biology and Evolution*, 2(6):526–538, 1985.
11. Luba Katz and Christopher B Burge. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res*, 13(9):2042–51, 2003.
12. Svetlana A. Shabalina, Aleksey Y. Ogurtsov, and Nikolay A. Spiridonov. A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research*, 34(8):2428–2437, 2006.
13. Olivier Bodini and Yann Ponty. Multi-dimensional boltzmann sampling of languages. In *DMTCS Proceedings - AOFA'10*, volume 0, pages 49–64, 2010.
14. Jérôme Waldispühl and Yann Ponty. An unbiased adaptive sampling algorithm for the exploration of RNA mutational landscapes under evolutionary pressure. *Journal of Computational Biology*, 18(11):1465–79, 2011.
15. Herbert S. Wilf. A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects. *Advances in Mathematics*, 24:281–291, 1977.
16. Alain Denise, Yann Ponty, and Michel Termier. Controlled non uniform random generation of decomposable structures. *Theoretical Computer Science*, 411(40-42):3527–3552, 2010.
17. Michael Drmota. Systems of functional equations. *Random Structures and Algorithms*, 10(1-2):103–124, 1997.
18. Ronny Lorenz, Stephan Bernhart, Christian Honer zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter Stadler, and Ivo Hofacker. Viennarna package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.
19. Sita J Lange, Daniel Maticzka, Mathias Möhl, Joshua N Gagnon, Chris M Brown, and Rolf Backofen. Global or local? predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res*, 40(12):5215–26, 2012.
20. Pascal Chartrand, Xiu Hua Meng, Stefan Huttelmaier, Damiane Donato, and Robert H Singer. Asymmetric sorting of ash1p in yeast results from inhibition of translation by localization elements in the mRNA. *Mol Cell*, 10(6):1319–30, 2002.
21. Isabel Gonzalez, Sara B Buonomo, Kim Nasmyth, and Uwe von Ahsen. ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Curr Biol*, 9(6):337–340, Mar 1999.

Supplementary Data are available at NAR online: Supplementary methods