Evidence of higher order patterns in information transmission between nucleotide sequences and folded molecular shapes of RNA

Christopher Barrett
Virginia Bioinformatics Institute
1015 Life Sciences Circle
Blacksburg, VA, USA
cbarrett@vbi.vt.edu

Fenix W. Huang
Virginia Bioinformatics Institute
1015 Life Sciences Circle
Blacksburg, VA, USA
fenixprotoss@gmail.com

Christian M. Reidys
Virginia Bioinformatics Institute
1015 Life Sciences Circle
Blacksburg, VA, USA
duck@santafe.edu

ABSTRACT

This contribution is a short version of a full paper submitted to Bioinformatics. DNA data transcribe into single stranded RNA, which folds into specific configurations. On the level of contact structures these are described by RNA secondary structures. Here we stipulate that RNA structures provide semantics for sequential DNA data. Accordingly we study the correlation between RNA sequences and RNA structures. We compute the partition function of sequences with respect to a fixed structure. We present a Boltzmann sampler and obtain the a priori probability of specific sequence patterns in such samples. We present a detailed analysis for the two **PDB**-structures, 2JXV (hairpin) and 2N3R (3-branch multi-loop). We localize where specific sequence patterns occur, contrast the energy spectrum of Boltzmann sampled sequences versus those sequences that refold into the same structure and derive a criterion to identify native structures.

Keywords

RNA sequence-structure relation, partition function, Boltzmann sampling, entropy

1. INTRODUCTION

In this paper we study the information transfer from RNA sequences to RNA structures. This question is brought into the context of processing of DNA data, specifically the role of DNA nucleotide sequences being transcribed into RNA.

DNA, or deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Its information is stored as a code made up of four chemical bases: adenine (**A**), guanine (**G**), cytosine (**C**), and thymine (**T**). DNA bases pair, forming units called base pairs. Each base is attached to a sugar and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. The latter are arranged in two strands that form a double helix.

RNA or ribonucleic acid (RNA) is unlike DNA, single-stranded. An RNA strand has a backbone made of alternating sugar (ribose) and phosphate groups. Attached to each sugar is one of four bases–adenine (**A**), uracil (**U**), cytosine (**C**), or guanine (**G**). There exists various types of RNA in the cell: messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). More and more RNAs have been found to play a role in a variety of other processes like for instance regulating gene expression. Recent transcriptomic and bioinformatic studies suggest the existence of thousands of so called non coding RNA, ncRNAs, i.e. RNA that does not translate into a protein. [6, 3]

Structurally less constrained, RNA folds into structures by forming in particular the Watson-Crick base pairs **G**-**C**, **C**-**G**, **A**-**U** and **U**-**A** as well as the Wobble base pairs **G**-**U** and **U**-**G**. In the following we consider RNA secondary structures, i.e. contact structures, that when drawing the sequence in a straight line and all Watson-Crick and **G**-**U** base pairs as arcs in the upper half-plane, have no crossing arcs, see Fig. 1.

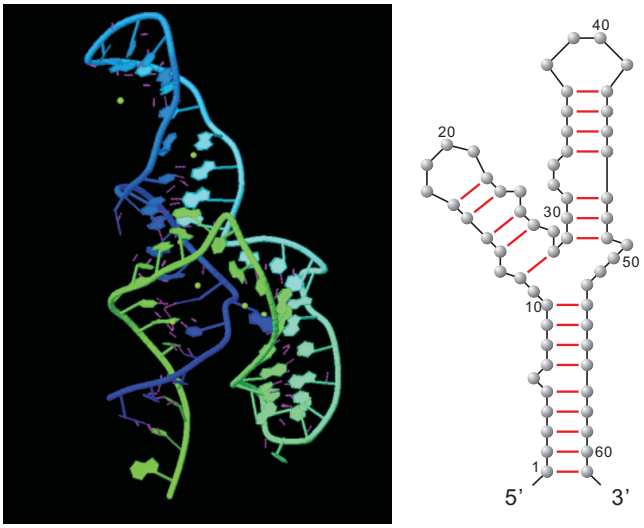

Figure 1: The 3D-structure of 2N3R from PDB [1] and its secondary structure.

.

BICT 2015, December 03-05, New York City, United States

DOI 10.4108/eai.3-12-2015.2262552

DNA information processing refers to replication, transcription and translation. Moreover we have RNA replication [10], reverse transcription (from RNA to DNA in e.g. retroviruses) [19] and a direct translation from DNA to protein [13, 20].

In the following we offer an alternative view of DNA information processing. Here we focus on the information transcendence from DNA/RNA sequences to the folded RNA (modulo transcription). We speculate that the sequential DNA information transcribes into single stranded RNA in order to interpret DNA data.

Currently DNA data are viewed as a sequence of nucleotides. We consequently run alignment tools as a means of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences[15] Here we suggest to entertain the idea that that the transcription into RNA with the implied self-folding is a way of lifting DNA information to a whole new level: RNA structures provide the semantics of DNA data.

In order to study this idea we consider a theoretical "proof of concept" framework: the folding of RNA sequences into minimum free energy (mfe) secondary structures [21]. Pioneered by Waterman more than three decades ago [18] and subsequently studied by Schuster *et al.* in the context of the RNA toy world [17] there is an abundance of information about this folding. In particular we have fairly accurate energy values for computing the so called loop-based mfe that are employed by the folding algorithms [22, 9]. However, much more detailed work has been done in [11, 12, 5] which will be the basis of a refined analysis of the framework proposed here.

In [14] McCaskill observed that the dynamic programming routines folding mfe structures [14] allow to compute the partition function of all possible structures for a given sequence. The partition function is tantamount to computing the probability space of structures that a fixed sequence is compatible with. Predictions such as base pairing probabilities are obtained in [9, 8] and are parallelized in [7]. [4] derives a statistically valid sampling of secondary structures in the Boltzmann ensemble and calculates the sampling statistics of structural features.

We shall augment this type of analysis by considering the "dual" of McCaskills partition function, i.e. the partition function of all sequences that are compatible with a fixed structure. Let S be a secondary structure over n nucleotides. Then the partition function of S is given by

$$Q(S) = \sum_{\sigma \in \Omega_4^n} e^{-\frac{\eta(\sigma,S)}{kT}}, \tag{1}$$

where $\eta(\sigma, S)$ is the energy function based on the loop-based energy model given in [11], k is the universal gas constant and T is the temperature.

Having computed the partition function $Q(S)$ as well as the $Q(y_i, y_j)$ terms puts us in position to Boltzmann sample sequences for fixed secondary structure S. Here the probability of a sequence σ to be sampled is given by

$$\mathbb{P}(\sigma) = \frac{e^{-\frac{\eta(\sigma,S)}{kT}}}{Q(S)}.$$

Slightly more generally we consider also the pairing

$$\varepsilon \colon \Omega_4^n \times \mathcal{S}_n \longrightarrow \mathbb{R}. \tag{2}$$

2. RESULTS

In this section we discuss to what extend a structure determines particular sequence patterns and what differentiates native from random structures.

Since the energy model underlying the current analysis does not take non-canonical base pairs into account, we defer a detailed analysis of the mutual information to a later study where we use the MC-model [16].

Let $x_{i,j}$ be a concrete subsequence on the interval $[i, j]$, having probability $\mathbb{P}(x_{i,j})$. Its Shannon entropy $E_{i,j}$ is given by

$$E_{i,j} = \sum_{\forall x_{i,j}} \mathbb{P}(x_{i,j}) \log_4 \mathbb{P}(x_{i,j}).$$

By construction, $0 \leq E_{i,j} \leq (j - i + 1)$, where $E_{i,j} = (j - i+1)$ when all $x_{i,j}$ have the same probability, i.e., uniformly distributed, and $E_{i,j} = 0$ when $x_{i,j} = y_{i,j}$ is completely determined, i.e., $\mathbb{P}(y_{i,j}) = 1$. Let $R_{i,j} = 1 - (E_{i,j}/(j - i + 1)$ be the concentration of $[i, j]$, i.e. $R_{i,j} = 0$ for random sequences and $R_{i,j} = 1$ if there exists only one pattern $p_{i,j}$. We display the collection of $R_{i,j}$ as a matrix (heat-map), in which we color the higher $R_{i,j}$ darker and $R_{i,j} = 0$ is displayed as white. Due to computational limitations, we compute $R_{i,j}$ for $j - i + 1 \leq 8$.

The heat maps presented here are obtained by Boltzmann sampling an ensemble of 10^4 sequences from $Q(S)$. We present the energy distribution of this ensemble in Fig. 3 and Fig. 5 and in addition the energy spectrum of those sequences that actually fold into S via the classic folding algorithm. (We write the program using the same energy function to avoid bias.) The Inverse folding rate (IFR),

$$\text{IFR} = \frac{\#\text{ of sequences folding into } S}{\#\text{ of sampled sequences}}$$

measures the rate of successful refolding from that ensemble.

Let σ be a sequence from a Boltzmann sample of size 10^4 w.r.t. the structure S. Let $\bar{S}$ denote the structure that σ folds to. We consider

$$\Delta G(\sigma, S) = |\eta(\sigma, S) - \eta(\sigma, \bar{S})|$$

and compare the $\Delta G(\sigma, S)$ of several native structures contained in PDB with those of a several random structures (obtained by uniformly sampling RNA secondary structures).

In the following, we examine some RNA examples from Protein Data Bank (PDB) [1]

PDB: 2JXV The RNA structure 2JXV represents a segment of an mRNA, having length 33. The structure exhibits a tetra-loop, an interior loop and two stacks of length 8 and 5, respectively, see Fig. 2. We Boltzmann sample 10^4 sequences for this structure observing an AU ratio of 18.18%, while CG ratio is 81.82%. The IFR reads 95.16%, i.e. almost all sampled sequences refold into 2JXV. The heat map of 2JXV is given in Fig. 2. We observe that the tetra-loop determines specific patterns. This finding is not entirely straightforward as the hairpin-loops are the last to be encountered when Boltzmann sampling. I.e. they are the most correlated loop-types in the sense that structural context influences them the most.

The energy distribution of the Boltzmann sample is presented in Fig. 3 and we observe that the inverse folding solution is not simply the one that minimizes the free energy w.r.t. 2JXV. with the best energy. $\Delta\eta(\sigma)$-data are not displayed here in view of the high IFR.

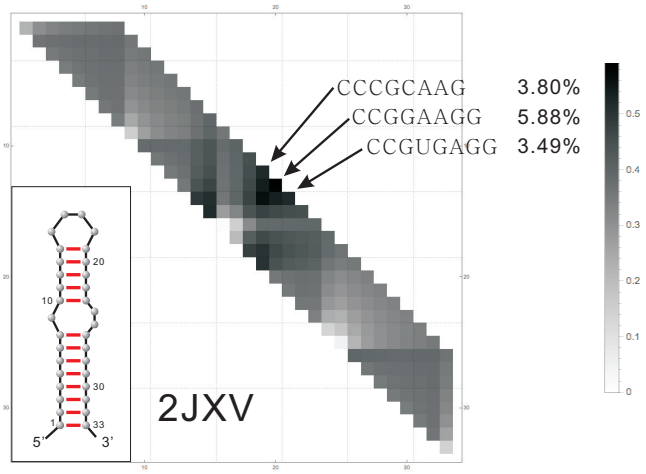

Figure 2: The secondary structure of 2JXV and its heat-map. We display the most frequent sampled patten for the largest interval having $R_{i,j} > 0.52$. The sample size is 10^4.

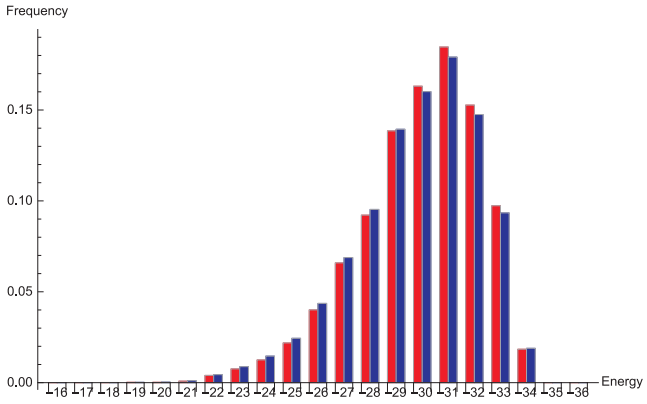

Figure 3: The energy distribution of the Boltzmann sample for 2JXV. We display the frequency of sequences having a particular energy (blue) and the frequency of sequences that fold into 2JXV.

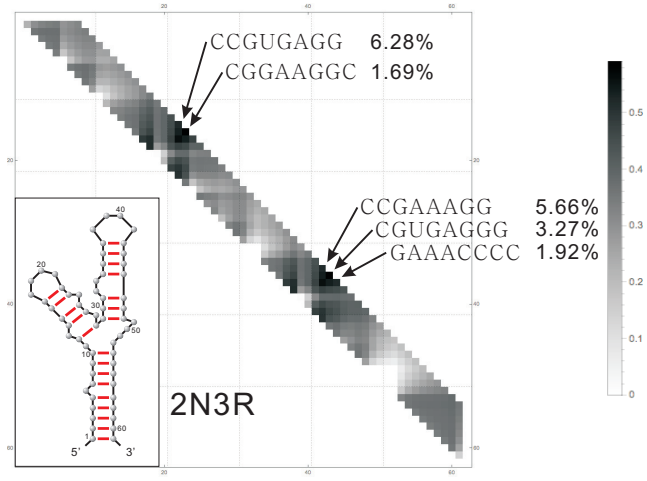

Figure 4: The secondary structure of 2N3R and its heat map. We show the most frequent patterns for the largest interval having $R_{i,j} > 0.52$. The sample size is 10^4.

PDB: 2N3R The structure 2N3R consist of 61 nucleotides and has a 3-branch muti-loop, two tetra loops, interior loops and helixes, see Fig. 4. The ratios of AU and CG pairs are 19.67% and 80.33%, respectively, again in a Boltzmann sample of 10^4 sequences. The IFR is still quite high, despite the fact that 2N3R is much more complex than 2JXV, its IFR is 0.69%. We illustrate the heat-map of 2N3R in Fig. 4.

Comparing the sequence segments $[17, 24]$ and $[37, 44]$, both of which being tetra loops with additional two nucleotides. The $R_{i,j}$ values of these segments are similar, approximately 0.59, however, their most frequently sampled patterns appear at different rates. For $[17, 24]$ this pattern is `CGGAAGGC` and it occurs with a Boltzmann sampled frequency of 1.69% and pattern probability 1.44%, while for $[37, 44]$ it is `CGUGAGGG` with sampled frequency 3.27% and pattern probability 3.24%. This shows that pattern frequency distributions are strongly correlated with structural context.

The energy distribution of the Boltzmann sample is given in Fig. 5 (A) and we display the $\Delta\eta(\sigma)$-data in Fig. 5 (B) where we contrast the data with $\Delta\eta(\sigma)$-values obtained from Boltzmann sampling 10^4 sequences of 5 random structures of the same length. We observe that the $\Delta G(\sigma, S)$-values for 2N3R are distinctively lower than those for random structures.

The above examples indicate that sequence-structure correlations can be used to locate regions where specific embedded patterns arise. Furthermore we observe that studying $Q(S)$ has direct implications for inverse folding. This is in agreement with the findings in [2].

This implies that sequences carry embedded patterns that cannot be understood by considering the sequence of nucleotides. At this point we have no concept of what these patterns are and provided a rather conventional notion of "embedded pattern". However, even when considering specific nucleotide patterns in hairpin loops, we observe significant context dependence on the structure. Other loops affect the energy of the hairpin loop and thus determine this particular subsequence. We observe that the embedded patterns can, for certain structures, be quite restricted, possibly elaborate and are not entirely obvious. In any case, the analysis cannot be reduced to conventional sequence alignment. The heat-maps introduced here identify the regions for which only a few select patterns appear and computed the *a priori* probabilities of their occurrence.

This type of analysis will be carried out for the far more advanced MC-model [16], where non-canonical base pairs can be incorporated. This will in particular enable us to have a closer look at the hairpins of the tRNA structure. In addition we believe that this line of work may enable us to arrive at non-heuristic inverse foldings.

As mentioned above, the present analysis is just a first step and discusses embedded patterns in the sense of subsequent nucleotides. However our framework can deal with any embedded pattern. We think a deeper, conceptual analysis has to be undertaken aiming at identifying how a collection of structures provides sequence semantics. Quite possibly this can be done in the context of formal languages. We speculate

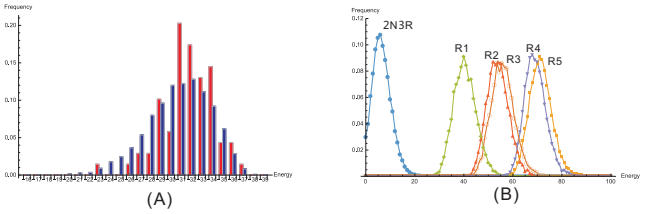

Figure 5: (A) The energy distribution of sampled sequences. The frequency of sequences having a particular energy level (blue), the frequency of sequences folds into 2N3R (red). (B) $\Delta\eta(\sigma)$-data for 2N3R versus $\Delta\eta(\sigma)$-data of five random structures.

that advancing this may lead to a novel class of embedded pattern recognition algorithms beyond sequence alignment.

3. ACKNOWLEDGMENTS

We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM, number FP7-ICT-318121 Special thanks to Reza Rezazadagen, Madhav Marathe, Rebecca Wattam, Henning Mortvei for discussions. We are grateful to Kevin Shinpaugh and the computational team at Virginia Bioinformatics Institute for their help and support.

4. REFERENCES

[1] Protein data bank, 2015. `http://www.rcsb.org/pdb/home/home.do`.

[2] A. Busch and R. Backofen. Info-rna–a fast approach to inverse rna folding. *Bioinformatics*, 22(15):1823–31, 2006.

[3] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308(5725):1149–54, 2005.

[4] Y. Ding and C. E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31:7280–7301, 2003.

[5] C.B. Do, D.A. Woods, and S. Batzoglou. Contrafold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–8, 2006.

[6] S. R. Eddy. Non-coding RNA genes and the modern rna world. *Nat. Rev. Genet.*, 2(12):919–29, 2001.

[7] M. Fekete, I.L. Hofacker, and P.F. Stadler. Prediction of RNA base pairing probabilities on massively parallel computers. *J. Comput. Biol.*, 7:171–182, 2000.

[8] I. L. Hofacker. The vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.

[9] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[10] E. V. Koonin, A. E. Gorbalenya, and K. M. Chumakov. Tentative identification of RNA-dependent RNA polymerases of dsRNA viruses and their relationship to positive strand RNA viral polymerases. *FEBS Lett.*, 252(1-2):42–6, 1989.

[11] D Mathews, M Disney, J Childs, S Schroeder, M Zuker, and D. Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci*, 101:7287–7292, 2004.

[12] D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, 2004.

[13] B. J. McCarthy and J. J. Holland. Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci. USA*, 54(3):880–886, 1965.

[14] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[15] D. M. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2nd ed. edition, 2004.

[16] M. Parisien and F. Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452:51–55, 2008.

[17] P. Schuster. Genotypes with phenotypes: adventures in an RNA toy world. *Biophys. Chem.*, 66(2-3):75–110, 1997.

[18] T.F. Smith and M.S. Waterman. RNA secondary structure. *Math. Biol.*, 42:31–49, 1978.

[19] H.M Temin and S. Mizutani. RNA-dependent DNA polymerase in virions of rous sarcoma virus. *Nature*, 226(5252):1211–3, 1970.

[20] T. Uzawa, A. Yamagishi, and T. Oshima. Polypeptide synthesis directed by DNA as a messenger in cell-free polypeptide synthesis by extreme thermophiles, thermus thermophilus HB27 and sulfolobus tokodaii strain 7. *J. Biochem*, 131(6):849–53, 2002.

[21] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. (Suppl. Studies)*, 1:167–212, 1978.

[22] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.