PROJECT REPORT: PAUSING SITES AND COTRANSCRIPTIONAL RNA FOLDING

NILS GUBELA

Abstract. With RNA secondary structures a prediction of the RNAs function can be made. Thermodynamic models help us understand the equilibrium distribution of RNA structures but the transcription process might alter the conformation found at the end of transcription. In the present report we study the pausing behaviour of the RNA polymerase to understand if it influences the distribution of structures at transcription end. Therefore a program was implemented that finds optimal pausing sites to increase a structures occupancy. This leads to the design of artificial sequences with distribution of structures at the transcription end that are controlled by pausing. An analysis of experimentally confirmed pausing sites shows that the pauses increased the occupancy of the observed structure like almost no other pausing site along the sequence.

Contents

1. Introduction

Noncoding RNA make up most of the transcribed RNA in the human body. They do not serve as a blueprint for proteins but have a function on their own. This function is determined by the secondary structure, i.e. isosteric base pairs. Ground states, base pairing probabilities, as well as thermodynamic properties can be computed but there are more factors that influence the native structure. The nascent RNA chain can fold during transcription and therefore end up in states that were otherwise unlikely to observe [1] or it can speed up the folding into the minimal free energy (MFE) structure by sequential folding [2]. There are currently three strategies for cotranscriptional structure prediction: Stochastic simulation (e.g. Kinfold [3]), master equation methods (e.g. BarMap [4]) and deterministic prediction of a single folding trajectory (e.g. Kinwalker [5]). A combination of the master equation method and single trajectory prediction is the heuristic "DNA to RNA Transformer" DrTransformer [6]. Master equation methods are useable for very short sequences (up to 70 nucleotides) and single trajectory prediction only selects the one most populated secondary structure at each transcription step. DrTransformer calculates for each transcription step an energy landscape with a conformation graph that yield the transition rates of structures. A new nucleotide is added to a structure and then its neighbourhood is searched for new, energetically better conformations. New occupancies are calculated via kinetic simulation and structures with low occupancies are removed. DrTransformer outputs the distribution of structures at each transcription step and at the transcription end together with the equilibrium distribution.

DrTransformer uses for the kinetic simulation at each step a simulation time of 20 ms as default. RNA polymerase (RNAP) adds nucleotides to the growing transcript rapidly at most DNA positions (in 10–50 ms) but takes much longer (seconds to minutes) at other positions [7].

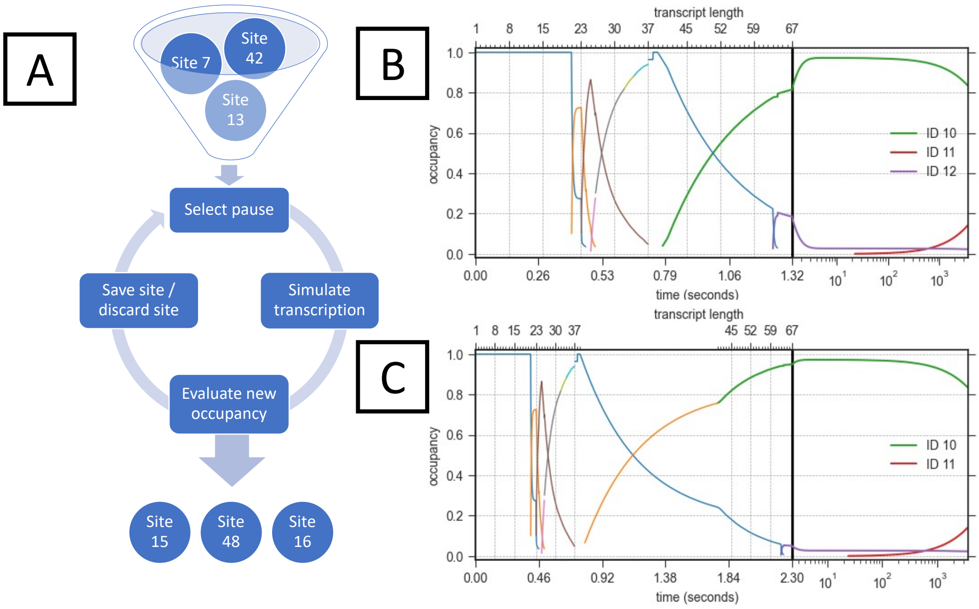


Figure 1. Schematic description and application example of the pausing site finding algorithm. (A) Pausing site candidate is selected, the transcription with the pause is calculated using DrTransformer, the new occupancy of the structure is evaluated and a decision is made to save or discard the pausing site. (B) Occupancies of structures during and after transcription of test sequence CCUCUUAGUCAUGGUAAACUAUGCCAGUUUGAAGUGGAGGU-UUUAGGUUGGCAAUUAAUAUCAGUUC. Without pausing the structure with ID 10 has an occupancy of 0.83 at the transcription end. (C) Including a pause site at nucleotide 39 increases the occupancy of structure 10 to 0.92 at transcription end.

These slow steps are called pauses. They are sequence dependent and reproducible. The location and/or duration of the pause sites can alter folding [8]. But much like cotranscriptional folding the exact influence of pausing sites on the RNA structure is unknown. DrTransformer offers an option to manually set the transcription time at certain positions. This results in a longer kinetic simulation and a possible change in the distribution of structures. We use this feature of DrTransformer to build a pausing site finding algorithm that given a sequence structure pair returns a pausing strategy to increase the structures occupancy at transcription end. The main application of this tool is to gather insight into the purpose of pausing. We want to understand if the RNAP uses pausing as a mechanism to favour unfavourable confirmations during transcription. We therefore design sequences with different folding behaviour with and without pausing. We also compare experimentally found pausing sites to the proposed sites of the pausing sites finding algorithm.

2. Methods

The center of this project is an algorithm implemented in Python 3 that finds pausing sites that increase the occupancy of one structure at the end of transcription (see Figure 1 A). It selects a site of the transcribed DNA at random and checks weather the inclusion of a pause site increases the occupancy of a predefined sequence at the end of transcription. DrTransformer v0.10 is called at each step to simulate cotranscriptional folding with the specified pausing site. If the occupancy is higher than in the previous step, we save this site as a pausing site and continue to select random sites until we either checked $2n$ possibilities, where n is the sequence length, or obtained three pausing sites. This means that on average each site is tested two times. It is experimentally confirmed that the RNAP dwells one to three times during the transcription of certain RNA [9, 10, 11, 12]. Gajos et al. found active genes with no or only one high-confident pausing site and genes with pauses every 140 nucleotides with an algorithm for peak detection from single-nucleotide resolution profiling data for investigating transcriptional pausing in human cells [13]. Therefore, three pausing sites is the maximum of pausing sites that the algorithm considers. Since this is an heuristic tool, there is no need to optimize the occupancy of the given structure but only to check weather or not it can be increased.

To get a first understanding of the possible capabilities of pausing of the RNAP we want to answer the question if pausing can increase the occupancy of all possible structures in the ensemble. We therefore generate a set of short random RNA sequences that contain at each position one of the four bases A,C,G,T with equal probability. We select two representative structures from the ensemble, the MFE calculated with RNAfold v2.1.9 [14] and another structure that lies in an energy interval of 6 kcal/mol above the MFE and has a barrier of at least 5 to the MFE. To get this second structure we use RNAsubobt v2.1.9 of the ViennaRNA package [14] to create the ensemble above the MFE and barriers v1.8.1 [15] to filter the ensemble for local minimal structures with the proposed barrier. The sequence structure pairs are then processed by the pausing site finding algorithm. A total of 500 sequence structure pairs of varying length between 50 and 120 nucleotides are examined.

To gather evidence that the RNAP dwells at certain sites to regulate the occupancies of structures we design sequences to enforce this behaviour. We use RNAblueprint [16] to uniformly sample RNA structures with given constrains. This is incorporated into a simulated annealing algorithm with objective function that controls the energy difference of two structures in which the desired sequence should fold into. The goal is to obtain structures that predominantly fold with regular transcription rate into the MFE and that can fold with pausing into an energetically non-optimal structure. The following two structural constrains are chosen:

```
1 (((((..((((.....(((((.....)))))...)))))))))((((.....(((((.....)))))...))))........
2 ...............(((((.....((((((((((((((((......))))))))))))))))).....))))).........
```

Structure 1 is divided into two parts that share no base pairs. The first part can fold during transcription and is not influenced by the second part. This should also favour the folding of the second part. The 5′-end is very structured which means that the structure can easily fold during transcription. Structure 2 has an unstructured 5′-end which competes with the many base pairs in the 5′-end of structure 1 that can be formed at the beginning of transcription. It is worth mentioning that structure 1 and 2 do not have a single base pair in common and have a base pair distance of 44.

Wong et al [12] found pausing sites in noncoding RNAs of Escherichia coli in vitro. According to their results there are two pausing sites present in E. coli SRP RNA at position U82 and U84. We use the E. coli SRP RNA to test the pausing site finding algorithm and further investigate the experimentally found pausing sites with DrTransformer. As control we insert a pausing site at any position of the sequence of length five seconds to see if the experimentally and theoretically obtained pausing sites have significant influence on the occupancy of the structure compared to other sites.

3. Results

From the 500 analysed random sequence structure pairs the occupancy of 24% could be increased with pausing. This means that the RNAP is certainly not able to fold into any structure of the ensemble just by including pausing sites. For 7% of the sequence structure pairs the observed increase generated by pausing was more than 0.1. Most of the pausing sites were found for MFE

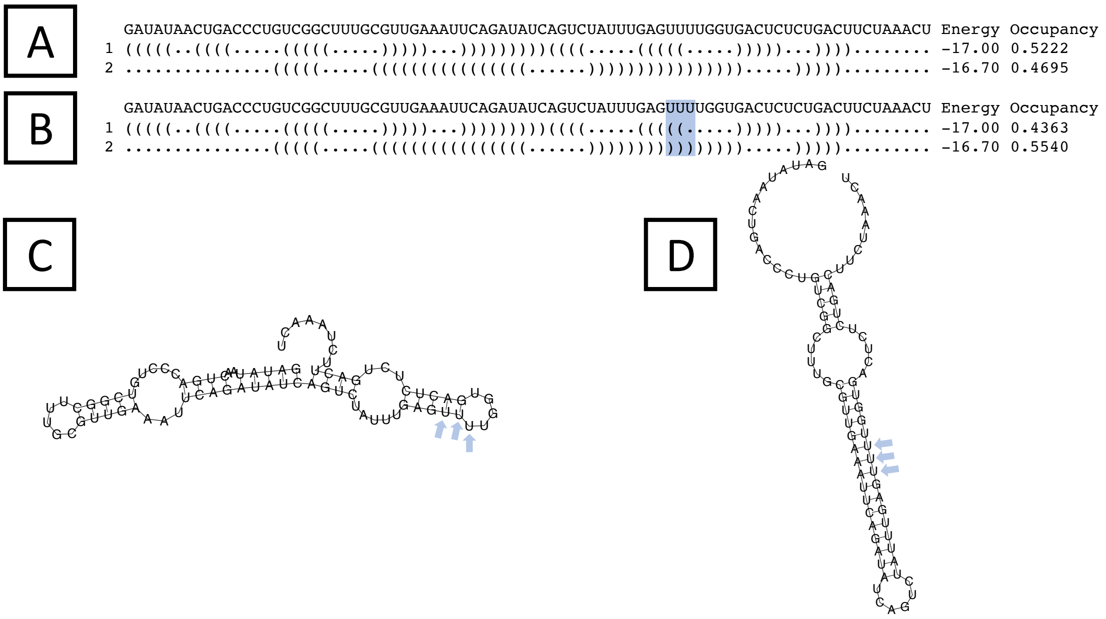
Table 1. Results of analysis of random sequences sorted per sequence length: Column one contains the length of the sequences. Column two contains the number of sequence structure pairs per length. Column three contains the number of sequence structure pairs for which a pausing site was found. The number in parenthesis is the percentage relative to the number of sequence structure pairs. Column four contains the number of MFE structures for which pausing was found. Column four contains the number of sequence structure pairs that had an increase of more than 0.1 after using the pausing site. Column six contains the number of sequence structure pairs that were not observable after standard transcription but could be observed after the inclusion of a pausing site.

Length	Seq-Struc Pairs	Pausing found	MFE increased	Increase > 0.1	Increase from 0
50	40	5 (12.5%)	4 (10%)	1 (2.5%)	0
60	40	5 (12.5%)	4 (10%)	2 (5%)	0
70	40	8 (20%)	7 (17.5%)	1 (2.5%)	0
80	80	22 (27.5%)	17 (21.25%)	6 (7.5%)	2 (2.5%)
90	80	13 (16.5%)	13 (16.25%)	5 (6.25%)	1 (1.25%)
100	86	23 (26.744%)	19 (22.09%)	6 (6.98%)	1 (1.16%)
110	80	26 (32.8%)	20 (25%)	8 (10%)	0
120	55	18 (34%)	13 (24.53%)	6 (11.32%)	4(7.55%)
Total	500	120 (24%)	97 (19.4%)	35 (7%)	8(1.6%)

structures (97 from 120). There are eight sequence structure pairs that were not observable after a transcription with equal rate at each site but became observable after the inclusion of a pausing site. For short sequences it is more unlikely to find a pausing site than for longer sequences. For a detailed overview of the results consult Table 1.

We use the simulated annealing algorithm with RNAblueprint to obtain sequences that may fold in two dissimilar structures. One noteworthy example can be found in Figure 2. Structure 1 which is divided into two parts and is easy to fold during transcription is the MFE of the found sequence. If the RNAP adds nucleotides with a constant rate at each site the occupancy of structure 1 at the end of transcription is 0.5222 while structure 2 is observable with a rate of 0.4695. The pausing site finding algorithm proposes three consecutive pausing sites at position U56, U57 and U58. If the RNAP dwells at these three sites for 5 seconds, non-MFE structure 2 is observed predominantly at transcription end. Note that the equilibrium distribution of structure 1 is 0.6152 and of structure 2 it is 0.3781. This means that with pausing we are able to create a short uprise in the occupancy of structure 2 before it regresses back to the equilibrium. Using the combination of RNAsubopt and barriers from the random sequence study on this sequence yields the MFE structure 1 and structure 2 which means that there is a high energy barrier between these two structures. Further investigation with barriers show that structure 1 and structure 2 are minima in unconnected barrier trees (see Figure 3).

Wong et al. [12] discovered pausing sites on various E. coli RNAs. E. coli SRP RNA has 114 nucleotides and the experimentally observed structure contains four interior loops and a hairpin loop (see Figure 4B). There were two pausing sites observed at U82 and U84. The experimentally observed structure can not be found in the DrTransformer output with or without pausing. The closest observable structure in terms of base pair distance can be seen in Figure 4. The difference is one additional base pair between the experimentally unpaired U38 and A68. The energy of the experimentally observed structure is -55.4 kcal/mol while the structure observable by DrTransformer has the slightly lower energy of -56.1 kcal/mol. The latter has an occupancy of 0.1339 after transcription with constant rate at each site. The pausing site finding algorithm suggests three pausing sites of length 5 seconds at G54, G64 and C65. Simulating transcription with these



```
  GAUAUAACUGACCCUGUCGGCUUUGCGUUGAAAUUCAGAUAUCAGUCUAUUUGAGUUUUGGUGACUCUCUGACUUCUAAACU Energy Occupancy
1 (((((..((((.....(((((.....)))))...)))))))))((((.....(((((.....)))))...))))........ -17.00 0.5222
2 ...............(((((.....((((((((((((((((......)))))))))))))))).....)))))......... -16.70 0.4695
```

```
  GAUAUAACUGACCCUGUCGGCUUUGCGUUGAAAUUCAGAUAUCAGUCUAUUUGAGUUUUGGUGACUCUCUGACUUCUAAACU Energy Occupancy
1 (((((..((((.....(((((.....)))))...)))))))))((((.....(((((.....)))))...))))........ -17.00 0.4363
2 ...............(((((.....((((((((((((((((......)))))))))))))))).....)))))......... -16.70 0.5540
```

Figure 2. Occupancy and structure of designed RNA sequence after transcription. (A) Result after transcription with constant rate: Two structures are populated, the MFE (structure 1) has slightly higher occupancy of 0.5222. (B) Result after pausing at site U56, U57 and U58: The non-MFE structure (structure 2) is populated more often than the MFE. (C) Plot of the MFE structure 1. The arrows indicate the positions of the three pausing sites. (D) Plot of structure 2. The arrows indicate the position of the three pausing sites.

three pausing sites yield an occupancy of 0.9968 at transcription end of the structure of interest. We use DrTransformer to analyse the two experimentally proposed pausing sites. Pausing at position U82, U84 or U82 and U84 has the same result, which is an increased occupancy of 0.4776 at transcription end. To verify that these pausing sites have different behaviour than all other sites in the sequence, simulations of cotranscriptional folding with a single pausing site at each position are performed with DrTransformer. The resulting influence on the occupancy of the structure can be found in Figure 4D. The length of five seconds was chosen from prior observation that the equilibrium distribution at each site is reached after five seconds. We note that the main contribution to the increase of the occupancy of the structure at transcription end comes from the pausing at site G54. Without this pause the other two proposed sites G65 and C66 have a slight negative effect on the end occupancy. The two experimentally confirmed pausing sites U82 and U84 lie in a small region between G81 and G85 that all increase the end occupancy to 0.4776. Pausing at all sites after position 90 has a positive effect on the end occupancy. Pausing in the 3′-end of this sequence has a positive effect on the end occupancy since the occupancy of the structure increases after the end of transcription. If we disregard pausing at the 3′-end then the experimentally observed pausing sites of the RNAP are reflected in the simulations. There is one optimal pausing site at position G54 that was not observed experimentally but the only other increasing sites were chosen by the RNAP. The pausing site finding algorithm proposed different pausing sites than experientially confirmed. A closer look at the influence of a pause at each site on the end occupancy reveals that it does make sense for the RNAP to dwell at U82 and U84. There are five sites between G81 and G85 that have similar influence on the end occupancy. Pausing somewhere in this region will yield the desired effect. While there is a single global optimal pausing site at G54 it may be difficult for the RNAP to stop at precisely one position. Pausing at G53 or A55 has no effect on the end occupancy.

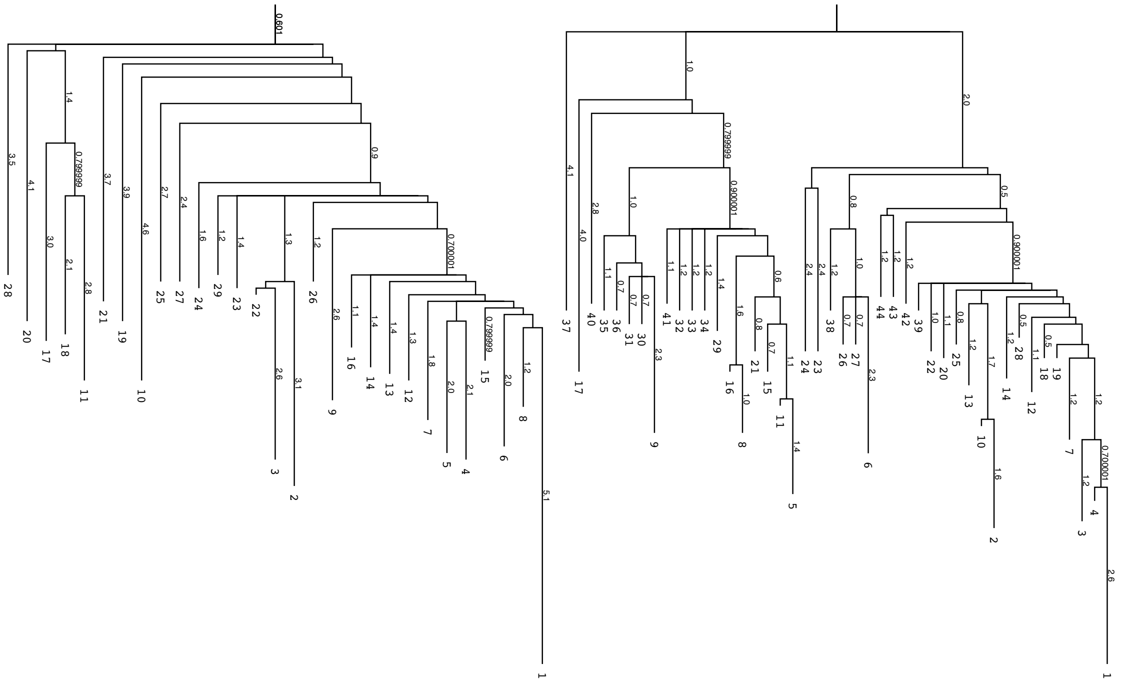


Figure 3. Two connected barriers trees of the designed sequences from Figure 2. (Left) The connected barrier tree where the label 1 corresponds to structure 1. (Right) The connected barrier tree where the label 1 corresponds to structure 2. There is no connected tree that contains both structure 1 and structure 2.

4. Discussion

The pausing site finding algorithm successfully proposes pausing sites that increase a structures occupancy at the end of transcription. It enables the design of artificial RNA sequences that have a folding behaviour that is controlled by the RNAP. Compared with experimentally found pausing sites the pausing site finding algorithm optimizes the occupancy of a certain structure. In reality there are many possibilities for the RNAP to dwell and control the distribution of structures. It is therefore important to look at the influence of any single site on the end distribution (see Figure 4D). These two examples outline the importance of cotranscriptional pausing and are valuable hints that we have a correct understanding of the purpose of pausing sites. Especially the pausing profile of the E.coli SRP RNA makes clear that the location of the two experimentally confirmed pausing sites lead to a meaningful change in the end distribution of structures. This means that observing this structure can be explained by the inclusion of the two pausing sites since they increase the structures occupancy.

The meaningfulness of the random sequence data for real sequences is questionable. Even though an effort was made to only include structures that are observable in the DrTransformer output with high probability, most of the structures were actually not observable. It is possible that they are just very unlikely to be observed at transcription end but they could also be coarse grained to a similar structure. It would lead to more insight if the study would be repeated with real RNA sequences. To curate a dataset with native structures that are also observable in the DrTransformer output lies beyond the scope of this project.

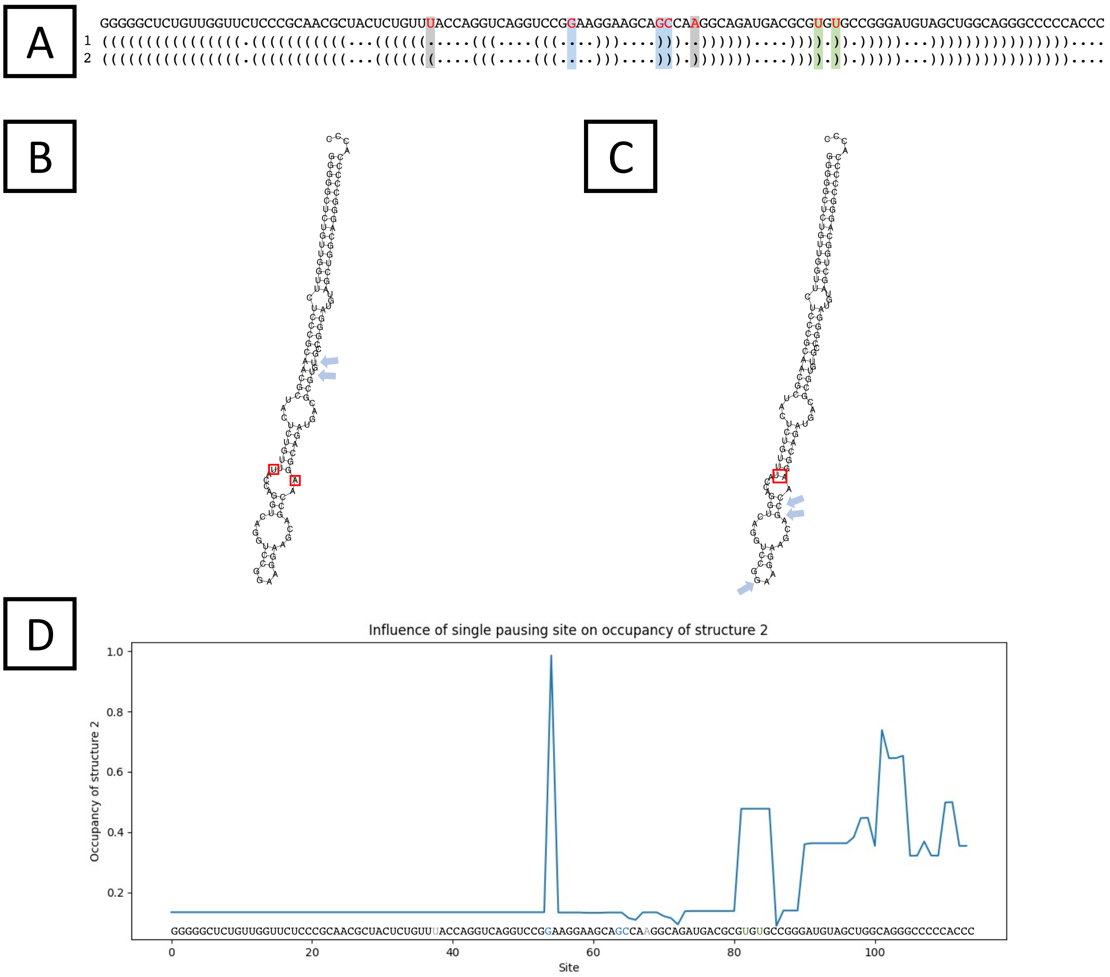


Figure 4. Occupancy and structure of E. coli SRP RNA after transcription.(A) Sequence of E. coli SRP RNA together with the experimentally observed structure by Wong et al. [12] labelled by 1 and the structure that is processed by DrTransformer labelled by 2. The one different base pair U38-A68 is highlighted in grey. The two pausing sites discovered by Wong et al. are highlighted in green (U82 and U84). The three pausing sites suggested by the pausing site finding algorithm are highlighted in blue (G54, G65, C66). (B) Plot of experimentally observed structure for the E. coli SRP RNA. The two experimentally confirmed pausing sites are marked by arrows. The two non paired nucleotides that are paired in structure 2 are marked by red boxes. (C) Plot of the structure found calculated by DrTransformer. The three pausing sites proposed by the pausing site finding algorithm are marked by blue arrows. The additional base pair to structure 1 is enclosed by a red box. (D) Influence of single five second long pause at any site on the occupancy of structure 2 at the end of transcription. The sequence is written over the x-axis for orientation. The additional base pair is marked in grey, the three pausing sites found by the pausing site finding algorithm are marked in blue and the two experimentally found pausing sites are marked in green.

The presented algorithm and concepts should further be applied to more biological data. Gajos et al. examined the pausing behaviour of RNA polymerase II in human cells with an algorithm for peak detection from single-nucleotide resolution profiling data. This large scale study can be replicated with the algorithm at hand to see if the proposed pausing sites match the ones from Gajos et al. and to understand if they are related to structural properties of the transcribed RNA. Another potential application to biological data is homology search. Watts et al. showed that RNAP pauses at very similar genic locations among individuals, and in some cases, at the same nucleotide positions [11]. Structural similarity is already used for homology search, incorporating pausing information could potentially be an additional indicator for relatedness.

References

[1] Kramer, F. R., & Mills, D. R. (1981). *Secondary structure formation during RNA synthesis.* Nucleic acids research, 9(19), 5109-5124. https://doi.org/10.1093/nar/9.19.5109

[2] Xayaphoummine, A., Viasnoff, V., Harlepp, S., & Isambert, H. (2007). *Encoding folding paths of RNA switches.* Nucleic acids research, 35(2), 614-622. https://doi.org/10.1093/nar/gkl1036

[3] Flamm, C., Fontana, W., Hofacker, I. L., & Schuster, P. (2000). *RNA folding at elementary step resolution.* Rna, 6(3), 325-338. https://doi.org/10.1017/S1355838200992161

[4] Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., Scheuermann, G., & Stadler, P. F. (2010). *BarMap: RNA folding on dynamic energy landscapes.* Rna, 16(7), 1308-1316. https://doi.org/10.1261/rna.2093310

[5] Geis, M., Flamm, C., Wolfinger, M. T., Tanzer, A., Hofacker, I. L., Middendorf, M., . . . & Thurner, C. (2008). *Folding kinetics of large RNAs.* Journal of molecular biology, 379(1), 160-173. https://doi.org/10.1016/j.jmb.2008.02.064

[6] Badelt, S., Lorenz, R., & Hofacker, I. Manuscript in preparation.

[7] Landick, R. (2021). *Transcriptional Pausing as a Mediator of Bacterial Gene Regulation.* Annual review of microbiology, 75, 291-314. https://doi.org/10.1146/annurev-micro-051721-043826

[8] Artsimovitch, I., & Landick, R. (2000). *Pausing by bacterial RNA polymerase is mediated by mechanistically distinct classes of signals.* Proceedings of the National Academy of Sciences, 97(13), 7090-7095. https://doi.org/10.1073/pnas.97.13.7090

[9] Wickiser, J. K., Winkler, W. C., Breaker, R. R., & Crothers, D. M. (2005). *The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch.* Molecular cell, 18(1), 49-60. https://doi.org/10.1016/j.molcel.2005.02.032

[10] Gromak, N., West, S., & Proudfoot, N. J. (2006). *Pause sites promote transcriptional termination of mammalian RNA polymerase II.* Molecular and cellular biology, 26(10), 3986-3996. https://doi.org/10.1128/MCB.26.10.3986-3996.2006

[11] Watts, J. A., Burdick, J., Daigneault, J., Zhu, Z., Grunseich, C., Bruzel, A., & Cheung, V. G. (2019). *cis elements that mediate RNA polymerase II pausing regulate human gene expression.* The American Journal of Human Genetics, 105(4), 677-688. https://doi.org/10.1016/j.ajhg.2019.08.003

[12] Wong, T. N., Sosnick, T. R., & Pan, T. (2007). *Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures.* Proceedings of the National Academy of Sciences, 104(46), 17995-18000. https://doi.org/10.1073%2Fpnas.0705038104

[13] Gajos, M., Jasnovidova, O., van Bömmel, A., Freier, S., Vingron, M., & Mayer, A. (2021). *Conserved DNA sequence features underlie pervasive RNA polymerase pausing.* Nucleic acids research, 49(8), 4402-4420. https://doi.org/10.1093/nar/gkab208

[14] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). *ViennaRNA Package 2.0.* Algorithms for molecular biology, 6(1), 1-14. https://doi.org/10.1186/1748-7188-6-26

[15] Flamm, C., Hofacker, I. L., Stadler, P. F., & Wolfinger, M. T. (2002). *Barrier trees of degenerate landscapes.* https://doi.org/10.1524/zpch.2002.216.2.155

[16] Hammer, S., Tschiatschek, B., Flamm, C., Hofacker, I. L., & Findeiß, S. (2017). *RNAblueprint: flexible multiple target nucleic acid sequence design.* Bioinformatics, 33(18), 2850-2858. https://doi.org/10.1093/bioinformatics/btx263

E-mail address: nils.gubela@univie.ac.at