# RRIkinDP: Direct Paths RNA-RNA Interaction Kinetics Model

Maria Waldl[1,2,3], Irene K. Beckmann[1,4], Sebastian Will[5], and Ivo L. Hofacker[1,6]

[1]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Vienna, Austria
[2]Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, Währingerstraße 42, 1090 Vienna, Austria
[3]Center for Anatomy and Cell Biology, Medical University of Vienna, 1090 Vienna, Austria
[4]Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, A-1030, Vienna, Austria
[5]Laboratoire d'informatique de l'École polytechnique (LIX), Institut Polytechnique de Paris, 91120 Palaiseau, France
[6]Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria
*maria@tbi.univie.ac.at, irene@tbi.univie.ac.at, sebastian.will@polytechnique.edu, ivo@tbi.univie.ac.at*

July 28, 2023

**Abstract**

Interactions between RNAs are an essential mechanism in gene regulation. State-of-the-art tools predict RNA-RNA interactions based on thermodynamic stability but largely neglect kinetic effects. To overcome this limitation, we propose novel models of RNA-RNA interaction dynamics. On this basis we can improve our understanding of general principles that govern RNA-RNA interaction formation and improve target prediction tools.

While the dynamics of secondary structure formation of single RNAs have been successfully modeled using transition systems between conformations [2, 10], analogous approaches for RNA-RNA interaction quickly lead to infeasibly large systems. Therefore, we propose reducing the interaction system to the direct trajectories (shortest paths) from possible first contacts to full hybridization. This key idea enables studying general principles and relevant features of the interaction formation as well as model details; e.g. the relative speed of intra- and intermolecular folding.

Due to the design of our RNA kinetics model, features like energy barriers can be computed efficiently. In the future, this could enable refining genome-wide target predictions through kinetic criteria. Beyond these immediate practical improvements, it could shed light on general principles like the long-debated influence of the accessibility of the initial contact site.

RRIkinDP is is available on GitHub at `https://github.com/mwaldl/RRIkinDP` .

# 1 Introduction

We present a novel model for RNA-RNA interaction formation along direct paths from an initial contact or seed structure to a given candidate structure. By concentrating on these direct paths, we aim to capture important folding events while effectively restricting the state space and thereby enabling efficient computational evaluations of kinetic interaction features. In this text, we introduce the direct paths interaction formation model and present a dynamic programming algorithm designed to identify optimal folding barriers. Additionally, we present a pipeline for computing various interaction features, including thermodynamic features, features of putative first contact sides and kinetic features. These features are derived either by applying the dynamic programming algorithm or directly from the previously defined state space. We provide a plotting script for visualizing the state space as 2D energy landscape for a first visual inspection of an interaction candidate. Furthermore, we introduce a machine learning approach to assess how well the computed features are suited to differentiate native RNA-RNA interactions from a randomized background set of interactions and to quantify how much novel information these features provide.

# 2 Methods

Our primary goal is to evaluate the kinetic features and feasibility of a given *candidate interaction structure*. Targeting a candidate structure notably differs from exploring the full space of interaction structures that can be formed by two RNAs for example according to the nearest neighbor (2D) energy model [7, 9]. Such a candidate structure can be obtained, for example, through thermodynamic interaction prediction [8]. We specifically focus on single-site interactions, where the interaction site cannot be interrupted by intramolecular structures (intramolecular base pairs).

In this work, we are generally interested in two interacting RNAs with positions $1, \ldots, n$ and $n+1, \ldots, n+m$. We define RNA structures and interactions at the level of secondary structure as sets of base pairs $(k, l)$, $1 \leq k < l \leq n+m$ between the nucleotides of the RNA(s). On these structures $R$ we impose common restrictions in the context of RNA secondary structure and interaction prediction [4]: both intra-molecular structures $P_R$ and $Q_R$ are non-crossing; moreover the set of inter-molecular base pairs $I_R$ is non-crossing.

The interaction base pairs $I_R$ of any interaction structure $R$ are naturally ordered "$5' \to 3'$" by their position in the first RNA: $(a, b) \prec (a', b')$ if and only if $a < a'$. Note that the non-crossing condition implies $b' < b$. We specifically

focus on single-site interactions, where the minimal and maximal interaction base pairs, respectively $(a_5, b_3)$ and $(a_3, b_5)$ limit the single interaction site. We require that interaction sites are not interrupted by intramolecular structures (intramolecular base pairs), i.e. the positions $a_5, \ldots, a_3$ and $b_5, \ldots, b_3$ are unpaired, i.e. do not occur in $P_R$ and $Q_R$.

Given the single-site candidate structure $C$, we define direct folding paths and structures along those paths based on $I_C = \{p_1, p_2, ...p_n\}$, $p_1 \prec p_2 \cdots \prec p_n$. We assume that an interaction starts from a first contact, which can be a base pair $(p_i)$ or a short helix $(\{p_i, ..., p_j\})$, and then extends by adding the next interaction base pair to the right or left end of the current intermediary structure with respect to the full candidate interaction. Consequently, any intermediary interaction can only consist of consecutive base pairs with respect to the full candidate interaction. By using this definitions, we can fully define any intermediary interaction structure $S_{i,j}$ based on the index $i$ of the first base pair and the index $j$ of the last base pair. In contrast to the inter-molecular base pairs, the corresponding intramolecular structures are treated implicitly by the energy model.

## Energy model

To score different folding paths within our direct paths model, we first need to compute the free energy $E(i, j)$ of the individual states (interaction structures) $S_{i,j}$ along the path. We employ an energy model similar to the RNAup model [8], which computes the free energy of an RNA-RNA complex as the contribution of the interaction structure plus the cost of making the interaction site accessible in both intramolecular structures.

The free energy contribution of the interaction structure $I$, denotes as $E^{\text{hybrid}}$, is defined as the sum of free energy contribution of individual loops in the interaction structure according to nearest neighbour model.

The contribution of the intramolecular structures is included as the difference in free energy between the unconstrained single RNA structures and the free energy of the single molecule structures given that the interaction site is constraint to be unpaired. This difference $E^{\text{unpaired}}$ quantifies the energy needed to make the interaction site unpaired and therefore accessible for interaction formation (interaction base pairs).

In more detail, we implemented two different models for computing the unpairing energies $E^{\text{unpaired}}$. In the first limiting case, we assume that the intramolecular structure equilibrates instantly. In this case $E^{\text{unpaired}}$ can simply be computed as the difference in free energy between the unconstrained ensemble of structures and a constrained ensemble of structures where the interaction site is forced to stay unpaired. In contrast, the second case assumes that only intramolecular base pairs that directly conflict with the interaction structure are removed and that the intramolecular structures can not adapt to the extent of the interaction site in any other respect. For this minimal intramolecular refolding model, a start structure for the individual RNAs has to be specified, e.g. the minimum free energy structure of the individual non interacting molecules. The accessibility constraint is full-filled by removing any base pair in the intramolecular structures that is in direct conflict with the interaction site. $E^{\text{unpaired}}$ is computed as the difference between the free energy of the such adjusted intramolecular structure and the intramolecular start structure. In

reality, we would expect intra- and inter-molecular folding rates to be similar. However, since we cover both limiting cases, we expect the correct behaviour to lie somewhere in between the two models.

## The energy landscape

As any interaction structure $S_{i,j}$ along direct paths can be uniquely described by two indices we can easily visualize the whole state space as a 2D energy landscape, where the x-coordinate is the index of the first interaction base pair and the y-coordinate the index of the last base pair. The free energy of an interaction structure is represented by the color shade of the corresponding cell. An example for such an energy landscape is described in the Example section (see also Figure 1).

## Computing the minimum folding barrier with dynamic programming

The folding rate along a particular path should be dominated by the energy barrier between start state and the saddle point, defined as highest energy state along the path. The energy barrier may therefore be a suitable score to compare different paths. In particular, we are interested in finding the saddle point with lowest energy $B_s$ over all direct folding path from a given start interaction to the full interaction. The start interaction is defined by its first base pair with the index $s_l$ and last base pair with the index $s_r$. Given a start interaction, the folding path with the lowest saddle point is found by optimizing the order of extensions to the left and to the right. Starting from a predefined seed, and direct path to the full interaction can be described by a sequence of extensions moves/steps. In our model, there are only two types of moves - an extensions to the left $l$ and an extension to the right $r$ and given the seed interaction $S_{s_l,s_r}$ any direct path consists of a fixed number of left and right moved. Thus, any optimization of a direct path is an optimization of the order of left and right extensions. The sequence of extensions moves produces a sequence of interaction states - the folding path. Our goal is to optimize the sequences of extension moves such that the energy of the highest energy structure along the path is minimal. This optimization problem can decomposed in a recursive fashion:

$$B_s(i,j) = \min \begin{cases} \max \begin{cases} E(i,j) \\ B_s(i+1,j) \end{cases} & \text{if } i < s_l \\ \max \begin{cases} E(i,j) \\ B_s(i,j-1) \end{cases} & \text{if } j > s_r \\ E(s_l, s_r) & \text{if } s_l = i \text{ and } s_r = j \end{cases} \tag{1}$$

where every extension step corresponds to one iteration of the recursion.

$B_s(i,j)$ is the minimal saddle point along the folding path from the seed state $S_{s_l,s_r}$ to the current interaction structure $S_{i,j}$. $S_{i,j}$ could either be reached by extending the interaction $S_{i+1,j}$ by one base pair to the left or by extending the interaction $S_{i,j-1}$ by one base pair to the right. In the first case the path up until $S_{i+1,j}$ either includes a saddle point $B_s(i+1,j)$ that is higher than the energy

$E(i,j)$ of the current structure and $B_s(i,j) = B_s(i+1,j)$ or the energy of the current interaction $E(i,j)$ is larger than the previous saddle point $B_s(i+1,j)$ and therefor the new saddle point $B_s(i,j) = E(i,j)$, given that $S_{i,j}$ was reached through a left extension in the last step. Analogous, if $S_{i,j}$ was reach by an extension to the right from $S_{i,j-1}$, we can derive the minimum barrier so far by comparing $B_s(i,j-1)$ and $E(i,j)$. This results in the minimum saddle points given that the last step was a left extension and given that the last step was a right extension. Based on those two options the extension step that results in a lower folding barrier is selected. The optimal extension sequence is back-traced recursively from the full interaction $S_{f_l,f_r}$ until the start structure is reached $S_{s_l,s_r}$ and the saddle point $B_s(f_l,f_r)$ of the best full folding path(s) is obtained. Any $E(i,j)$ can be computed from the two energy contributions described in the energy model section as follows.

$$E(i,j) = E^{\text{hybrid}}(i,j) + E^{\text{unpaired}}(i,j). \tag{2}$$

# 3 Example

To introduce the energy landscapes and interaction features that can computed by RRIkinDP, we present an application of our tool on the minimum free energy interaction of the small RNA DsrA and a 300nts area around the translation start side of rpoS, as predicted by Intarna [1, 6] . The corresponding nucleotide seqeunces can be obtiained from NCBI (DsrA: NC_000913.3/2025223-2025313, rpoS: NC_000913.3/2865573-2864581).

The binding of DsrA leads to a structure change in the rpoS 3' UTR and thereby makes the translation start side accessible [5, 3]. According to Intarna predictions, the thermodynamically most favourable interaction structure contains 27 base pairs and is formed between the positions 10-40 in DsrA and 175-204 in the rpoS fragment. A graphical representation of the candidate interaction structure can be found in the lower left of Figure 1 .

RRIkinDP generates all possible interaction structures on direct paths from a first base pair to the full candidate interaction structure. This structures and their neigborhood relation ate vizualized as energylandscape in Fig. 1. Details on how to read the energylandcspa plot are provided in the figure caption. Within this energy landscape one can for example identify short stable structure that could server as initial interaction seeds. In Table 1 the free energy of all potential seed structures of length for is listed. Based on this list there is potential for a stable seed interaction around $S_{13,16}$ and at the left most end of the interaction candidate ($S_{24,27}$).

In addition to the state space, RRIkinDP computes the minimum barrier state for direct paths from each possible start side to the full candidate structure. The free energy of this barrier state (saddle point) is also listed in Table 1 (column $E_{\text{barrier}}$). If the free energy of the saddle point is than free energy of the seed interaction, we assume that the interaction formation is slowed down. Possibly, the seed interaction could even dissociate instead of folding into the full candidate structure. An example for this is the left most seed interaction $S_{24,27}$.·

| seed structure | $E_{\text{seed}}$ | $E_{\text{barrier}}$ |
|---|---|---|
| $S_{1,4}$ | 5.47 kcal/mol | 5.47 kcal/mol |
| $S_{2,5}$ | 2.85 kcal/mol | 5.24 kcal/mol |
| $S_{3,6}$ | 0.38 kcal/mol | 5.24 kcal/mol |
| $S_{4,7}$ | 1.47 kcal/mol | 5.24 kcal/mol |
| $S_{5,8}$ | 3.05 kcal/mol | 5.24 kcal/mol |
| $S_{6,9}$ | 9.31 kcal/mol | 9.31 kcal/mol |
| $S_{7,10}$ | 9.18 kcal/mol | 9.18 kcal/mol |
| $S_{8,11}$ | 10.45 kcal/mol | 10.45 kcal/mol |
| $S_{9,12}$ | 4.53 kcal/mol | 4.63 kcal/mol |
| $S_{10,13}$ | 3.69 kcal/mol | 3.69 kcal/mol |
| $S_{11,14}$ | 0.86 kcal/mol | 0.86 kcal/mol |
| $S_{12,15}$ | -0.64 kcal/mol | -0.64 kcal/mol |
| $S_{13,16}$ | -0.91 kcal/mol | -0.91 kcal/mol |
| $S_{14,17}$ | -0.17 kcal/mol | -0.17 kcal/mol |
| $S_{15,18}$ | 0.55 kcal/mol | 0.55 kcal/mol |
| $S_{16,19}$ | 3.16 kcal/mol | 3.16 kcal/mol |
| $S_{17,20}$ | 2.94 kcal/mol | 2.94 kcal/mol |
| $S_{18,21}$ | 3.03 kcal/mol | 3.03 kcal/mol |
| $S_{19,22}$ | 6.26 kcal/mol | 6.26 kcal/mol |
| $S_{20,23}$ | 6.53 kcal/mol | 6.53 kcal/mol |
| $S_{21,24}$ | 7.88 kcal/mol | 7.88 kcal/mol |
| $S_{22,25}$ | 3.55 kcal/mol | 3.55 kcal/mol |
| $S_{23,26}$ | 1.6 kcal/mol | 1.6 kcal/mol |
| $S_{24,27}$ | -0.88 kcal/mol | 0.66 kcal/mol |

Table 1: This is the caption for the second table.

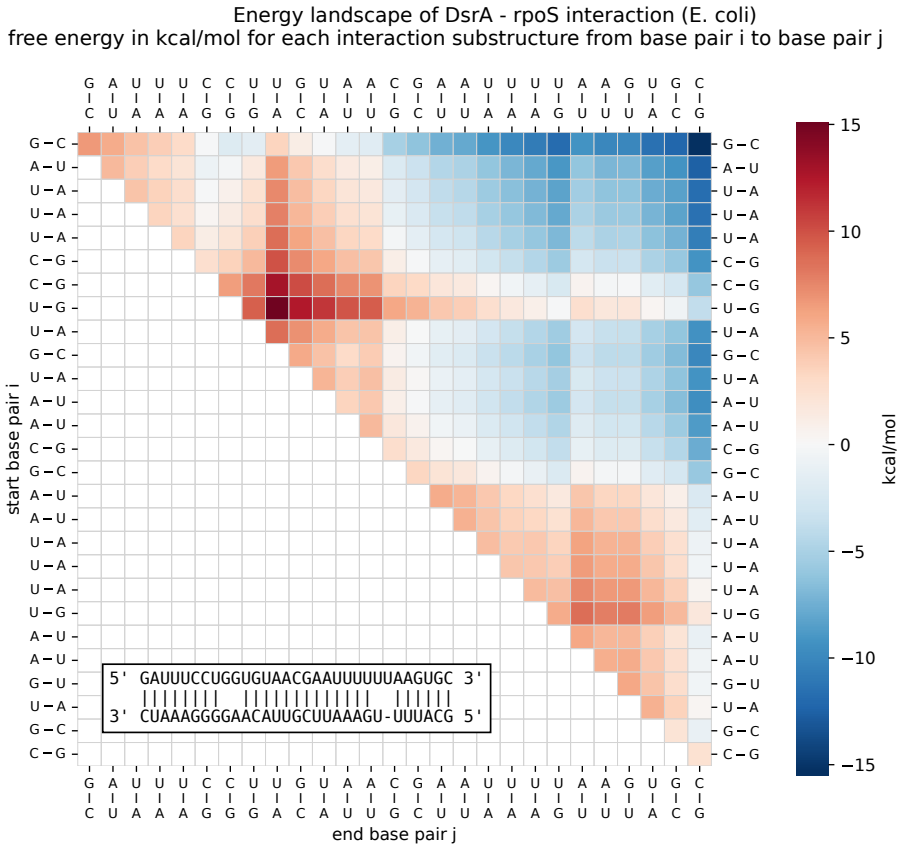


Figure 1: Energylandscape of the predicted minimum free energy interaction between DsrA and the 5' UTR of rpoS in E. coli: The energy landscape represents all possible sub-interactions that can occur along a direct path from a single base pair to the full Dsra-rpoS. Sub-interactions $S_{i,j}$ are defined by their first base pair i (row) and their last base pair j (column), with the their free energy $\Delta G$ in kcal/mol indicated by the red to blue color scale. Cells on the diaginal correspond to single base pair interactions. Moving one cell to the right (left) in this matrix corresponds to extending (shortening) the interaction by one base pair on the right. Analogously, moving one cell up (down) in the matrix corresponds to extending (shortening) the interaction by one base pair on the left.

## References


[1] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 10 2008.

[2] Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, March 2000.

[3] Richard A. Lease, Michael E. Cusick, and Marlene Belfort. Riboregulation in ¡i¿escherichia coli¡/i¿: Dsra rna acts by rna:rna interactions at multiple loci. *Proceedings of the National Academy of Sciences*, 95(21):12456–12461, 1998.

[4] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6(1):1–14, December 2011.

[5] Nadim Majdalani, Christofer Cunning, Darren Sledjeski, Tom Elliott, and Susan Gottesman. Dsra rna regulates translation of rpos message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proceedings of the National Academy of Sciences*, 95(21):12462–12467, 1998.

[6] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Research*, 45(W1):W435–W439, 05 2017.

[7] David H Matthews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceeding of the National Academy of Science of the USA*, 101:7287–7292, May 2004.

[8] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006.

[9] Douglas H Turner and David H Matthews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 38:D280—D282, Oct 2009.

[10] Michael T. Wolfinger, W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L. Hofacker, and Peter F. Stadler. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, 37(17):4731, April 2004.