THE INTERPLAY BETWEEN SINGLE-STRANDED BINDING PROTEINS ON RNA SECONDARY STRUCTURE

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Yi-Hsuan Lin, M.S., B.S.
Graduate Program in Physics

The Ohio State University
2015

Dissertation Committee:
Ralf Bundschuh, Advisor
Eric Braaten
Mohit Randeria
Michael Poirier
Abstract

Interactions between RNAs and RNA-binding proteins (RBPs) are significant in post-transcriptional regulation. In this process, an mRNA molecule is bound by many proteins and/or microRNAs to modulate its function. It is therefore an interesting question how these multiple RBPs collaborate to enable combinatorial gene regulation. Here, we propose a possible mechanism which can support this RBP-RBP collaboration, termed “cooperativity”. Such a cooperativity can exist merely based on fundamental principles of statistical physics and thermodynamics of RNA structure folding, without considering any further details of RNA and RBP properties. The theory is based on the idea that a successfully binding RBP will prohibit the formation of some originally allowed RNA structures, thus changing the statistical properties of the RNA structure ensemble, as well as the binding probabilities of other RBPs on the same RNA. In addition, this mechanism does not require direct physical interactions between RBPs, and thus supports the long-range characteristic of the cooperativity. Focusing on an RNA with two binding sites, we first calculate the correlation function between the RBPs on the RNA-RBP complex, verifying that this cooperativity exists. We then derive a characteristic difference of free energy differences, i.e. $\Delta\Delta G$, as a quantitative measure of this structure-mediated cooperativity. We apply this measure to a large number of human mRNAs, and discover that this cooperativity is a generic feature. Interestingly, this cooperativity not only affects binding sites in close proximity along the sequence but also configurations in which one binding site is located in the 5'UTR and the other is located in the 3'UTR of the mRNA. Some intriguing interplays between RBPs, microRNA binding sites, and UTR sequences are also disclosed. In the last chapter, we extend our model to handle multiple...
sequence-specified protein binding sites. We apply this extended model to the binding reaction between the protein HuR and several RNA sequences, theoretically calculating their dissociation constants and comparing with experimental results. We discover that RNA secondary structures are crucial in the interplay between HuR and RNA sequences, verifying the importance of the structure-mediated cooperativity in realistic RNA-protein binding reactions.
To my mom.
Acknowledgments

First I would like to thank my advisor Ralf Bundschuh for introducing me to the fantastic world of biology. In addition, he showed me what a professor should be as a paragon of a leader, researcher and instructor.

I would also like to thank Omar Tabbaa for many discussions about biology and physics, as well as the progress and the meaning of science. I really enjoyed those inspirational debates and appreciated that they greatly improved my English.

I feel indebted to Shizhong Zhang. In many regards, he was my first advisor for scientific research. I very much thank him for his help with how to rationally ask a question and both study and solve a problem.

The whole of my life in Columbus is just a surprising story. Probably what I appreciate most is my encounter with Yu-Wen. I am thankfully pursuing this gift of destiny and I preciously believe that time will prove everything between us.
VITA

March 19, 1987 . . . . . . . . . . Tainan, Taiwan Parents: Young-Chang Lin and Jui-Chu Chuang
June 2005 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Tainan First Senior High School, Tainan, Taiwan
2005–2007 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . National Taiwan University, Taipei, Taiwan
May 2009 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . B.S., University of Illinois, Urbana-Champaign, IL, USA
August 2012 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . M.S., The Ohio State University, Columbus, OH, USA
May 2015 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Ph.D Physics, Columbus, OH, USA

Publications


Fields of Study

Major Field: Physics
Table of Contents

Abstract ................................................................. ii
Dedication ............................................................... iv
Acknowledgments ......................................................... v
Vita ............................................................................ vi
List of Figures .................................................................. x
List of Tables ............................................................... xv

Chapters

1 Introduction ............................................................. 1
  1.1 The complexity in the central dogma of molecular biology ....... 1
  1.2 Combinatorial post-transcriptional regulation ..................... 2
  1.3 The long-range cooperativity mediated by RNA secondary structure . 3
    1.3.1 RNA secondary structure .................................. 3
    1.3.2 Mechanism for the long-range cooperativity ................. 4
  1.4 Investigation of the structure-mediated cooperativity ............ 7

2 Theory ........................................................................ 9
  2.1 Theory of RNA secondary structure ................................ 9
  2.2 Modeling of RNA-protein binding .................................... 12
  2.3 RNA-protein binding on the simplest RNA folding model ........ 14
  2.4 RNA-protein binding on the RNA folding model with constant loop cost .... 15
    2.4.1 Molten phase .............................................. 17
    2.4.2 Glass phase ............................................... 21
  2.5 RNA-protein binding on the RNA folding model with loop-size dependent cost .... 23
  2.6 RNA-protein binding on a realistic RNA folding model .......... 28
  2.7 Summary .................................................................. 30

3 Cooperativity in natural mRNAs between the 5' and 3'UTRs .......... 33
  3.1 Cooperativity in terms of free energy ............................... 33
  3.2 Preparation of RNA sequences and protein binding sites .......... 35
    3.2.1 Sequences .................................................. 35
    3.2.2 Locations of protein binding sites ......................... 37
  3.3 Biologically significant cooperativity in mRNA sequences ........ 37
    3.3.1 Distance dependence of structure-mediated cooperativity .... 38
3.3.2 Structure-mediated cooperativity is significant for large numbers of mRNA molecules ......................................................... 40
3.3.3 Cooperativities in human mRNAs and random RNA sequences have different statistical characteristics .......................... 43
3.3.4 Cooperativity is driven by the UTRs and not by the coding sequences ................................................................. 44
3.3.5 Cooperativities in C. elegans mRNAs are weaker than those in human mRNAs .......................................................... 44
3.3.6 RNA structure mediated cooperativity between the sequence ends persists in long mRNA molecules ......................... 48
3.4 Human microRNA binding sites .................................................. 48
3.4.1 Preparation and measure of the RBP-microRNA cooperativity .......................... 50
3.4.2 Human microRNA binding sites show significant negative cooperativities .......................................................... 51
3.5 Artificial evolution of highly cooperative sequences ...................... 53
3.6 Summary ................................................................. 55

4 Sequence-dependent protein binding affinity ................................. 58
4.1 Sequence specificity of RNA-binding proteins .............................. 58
4.2 Model for multiple RBP binding sites on an RNA molecule ........... 59
4.3 The sequence-specified bare dissociation constants ......................... 61
4.3.1 RNAcompete experiment .................................................. 61
4.3.2 Derivation of the sequence-dependent binding affinities ................. 61
4.4 Interplay between sequence-dependent protein bindings and RNA secondary structures ............................................... 63
4.5 Summary ................................................................. 67

5 Conclusion ........................................................................ 69

Bibliography ........................................................................ 72

Appendices

A Molten-phase partition function of the model including a constant loop cost .................................................................................. 81

B Calculation of the molten-phase correlation function for the model including a constant loop cost to first order in loop cost ............. 84
  B.1 \( S_d \), the partition function for the changed structures in \( Z_d \) ................................................................. 85
  B.2 \( S_{dd}^* \), contributions to \( Z_{dd} \) when both footprints are in the same base stack 88
  B.3 Correlation function .......................................................... 89

C Calculation of the molten-phase correlation function for the model including a constant loop cost to second order in loop cost ........ 92
  C.1 \( S_d \), the partition function for the changed structures in \( Z_d \) ................................................................. 92
  C.2 \( S_{dd}^* \), contributions to \( Z_{dd} \) when both footprints are in the same base stack . 98
  C.3 \( S_{dd}^{(1)} \), contributions to \( Z_{dd} \) when both footprints are in stacks ................................................................. 99
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A secondary structure of an RNA. The solid line shows the backbone of the RNA molecule. Dots are nucleotides, i.e. the bases of the RNA. Dashed lines are base-pair bonds. The regimes of multiple consecutive bonds are called stems, and the regions surrounded by unpaired bases are loops. Depending on the number of stems emanating from the loops, they have different names as indicated in the figure.</td>
</tr>
<tr>
<td>1.2</td>
<td>(a) The base-pair bonds forming pseudoknots. Pseudoknots are excluded in the discussion of secondary structures because (b) short ones do not contribute much to the total binding energy, and (c) the longer ones are kinetically unlikely to form because of their complex double-helical structures.</td>
</tr>
<tr>
<td>1.3</td>
<td>An example of structure-mediated cooperativity, where open boxes indicate unoccupied protein binding sites and filled boxes indicate occupied protein binding sites. $G$ is the free energy of the RNA-protein complex in each step, $G_0$ is the free energy of the RNA secondary structure alone, $\Delta G_s$ is the free energy associated with breaking the base pairs in the stem containing the two protein binding sites, $c_1$, $c_2$, $K_{D,1}$, and $K_{D,2}$ are the concentrations and dissociation constants of the first and second protein, respectively, $R$ is the gas constant, and $T$ is the temperature in Kelvin. (a) To have the first protein bind to bases in a stem of the structure in the absence of the second protein, the stem has to be broken, resulting in a free energy difference of $\Delta G_{0\to1} = \Delta G_s - RT \ln(c_1/K_{D,1})$. (b) Once the second protein is bound to the opposing bases in the stem, the first protein can directly bind to its binding site without the loss of free energy by breaking additional base pairs, yielding $\Delta G_{2\to12} = -RT \ln(c_1/K_{D,1}) &lt; \Delta G_{0\to1}$. Thus, in this case the two proteins exhibit positive cooperativity.</td>
</tr>
<tr>
<td>2.1</td>
<td>The four possible RNA-protein complexes to consider. Lines are RNA backbones and dots are bases. Transparent and black blocks represent the unoccupied and occupied protein binding sites, respectively. Only the bases which are not in the occupied binding regions participate in base pairing.</td>
</tr>
</tbody>
</table>
2.2 In the molten phase, the base-pair binding energies of different nucleotides become identical. The limited partition functions with consideration of protein binding sites no longer depend on sequences and retard to functions of lengths of segments between binding sites.

2.3 Structures in which the footprint is in a stem, i.e., inserted between two consecutive base-pair bonds. Such structures contribute differently in $Z_0(n + m)$ (left) and $Z_d(n, m)$ (right).

2.4 Numerical calculation of the protein protein correlation $g(D)$ on RNA polymers of $N = 200, 600, 1000, 2000$ based on the molten-phase homopolymer model, with $q = \exp(0.2)$, $s_0 = 5$ and $c_i/K_d(i) = 0$. The footprint is set to be $l = 6$. The correlation satisfy a power law $g(D) \sim D^{-3/2}$ (solid line) in the regime $N \gg D \gg l$.

2.5 Numerical calculation of the correlation function $G(D)$ on RNA polymers of 200 nucleotides for normal protein concentrations at low temperature, $k_BT = 0.3u$. The loop cost is $s_0 = 5$. The quenched average is taken over 400,000 random sequences. Footprints are $l = 6$ in the main graph and $l = 1$ in the inset. Protein binding parameters are $K_{d,1}^{(0)} = K_{d,2}^{(0)} = 0.1nM$. Concentrations are $c_1 = c_2 = 100nM$, yielding $c_k/K_d,k \approx 3.84$ by $e^{\beta \Delta F_k} \approx 0.00384$ for both $k = 1, 2$ in main graph, and $c_1 = c_2 = 2nM$ to have $c_k/K_d,k \approx 3.72$ by $e^{\beta \Delta F_k} \approx 0.186$ in the inset. The correlation functions of the two graphs follow power laws as $G(D) \sim 1/D^{0.9}$.

2.6 Numerical calculation of the correlation function $G(D)$ on RNA polymers of 200 nucleotides for normal protein concentrations at low temperature, $k_BT = 0.3u$. The loop cost is $s_0 = 5$ and $\nu = 0.5$. The quenched average is taken over 400,000 random sequences. Temperature is set to $k_BT = 0.3u$ and the secondary structure parameters are the same as those of Fig. 2.5. No power law dependence of the correlation function appears.

2.7 A structure in which the footprint is naked. Dashed lines represent the two independent substrands in the calculation for partition functions. Such structures contribute identically in $Z_q(n, m)$ and $Z_0(n + m)$.

2.8 Numerical calculation of correlation function $G(D)$ on RNA polymers of 200 nucleotides for normal protein concentrations at low temperature, $k_BT = 0.3u$. The loop cost is $s_0 = 5$ and $\nu = 0.5$. The quenched average is taken over 100,000 random sequences. Protein binding parameters are $K_{d,1}^{(0)} = K_{d,2}^{(0)} = 0.1nM$ and $c_1 = c_2 = 100nM$, with $e^{\beta \Delta F_k} \approx 0.000401$ for both $k = 1, 2$. The footprint is equal to $l = 6$ base pairs. The correlation function follows a power law as $G(D) \sim 1/D^{0.8}$.
2.9 Numerical calculation of protein-protein correlation $G$ on RNA polymers of 200 nucleotides. The calculation is based on the Vienna package and the quenched average is taken over 100,000 random sequences. Temperature is set to $T = 37^\circ C$. The parameters for protein binding are $c_1/K_d^{(0)} = c_2/K_d^{(0)} = 100$ with $e^{\beta \Delta F_k} \approx 0.00410$, which can be easily achieved by RNA binding proteins and are chosen to allow the proteins to compete with the RNA base pairing without completely outcompeting it. The length of the binding sites (footprint) is equal to 6 bases. Distance is expressed by the number of base pairs (bp) between the two binding sites. This log-log plot shows a regression to a power-law decay function $G(D) \sim 1/D^{0.9}$.

3.1 Definition of the outside distance, $D_{out}$, as the sum of the distances from each of the two protein binding sites to their corresponding ends in units of nt. Black dots represent bases and the two blocks represent the protein binding sites with footprint length $l$ (in this figure $l = 4nt$). In our numerical calculation we restrict ourselves to the symmetric case $n_1 = n_2$, and use a footprint size of $l = 6nt$.

3.2 The mean (circles) and standard deviation (diamonds) of the cooperativity free energy $\Delta \Delta G$ of 2282 human sequences, as a function of the outside distance $D_{out}$ between the two protein binding sites. The mean of the cooperativity free energy $\Delta \Delta G$ is close to zero for all outside distances, while its standard deviation is on the order of or greater than the thermal energy scale of $RT = 0.6 \text{kcal/mol}$. The dashed vertical lines label two different outside distances $D_{out}$ for which the whole distribution of the cooperativity free energies are compared in Fig. 3.3. The maximum of the standard deviation of the cooperativity free energy occurs at $D_{out} = 8nt$.

3.3 The number of mRNA molecules with respect to given cooperativity free energies $\Delta \Delta G$ (in bin sizes of 0.2kcal/mol) at outside distances between the protein binding sites of $D_{out} = 8$ (red solid line) and $D_{out} = 150$ (green dashed line). The inset shows the same data on a logarithmic scale. The two distributions are both centered at 0. The distribution for $D_{out} = 8$ (when the protein binding sites are close to their respective ends of the molecule) extends much wider than that for $D_{out} = 150$ when the protein binding sites are further away from the ends of the molecule. Vertical black dashed and dash-dotted lines label $|\Delta \Delta G| = RT$ and $|\Delta \Delta G| = 2RT$, respectively.

3.4 The number of mRNA sequences with biochemically relevant protein binding cooperativity among the 2282 investigated human mRNA sequences. The two data sets correspond to $|\Delta \Delta G|/k_B T > 1$ (circles) and $|\Delta \Delta G|/k_B T > 2$ (diamonds). The data shows that at any given distance $D_{out}$ a significant fraction of molecules has a significant cooperativity free energy.
3.5 The standard deviations of the cooperativity free energy $\Delta\Delta G$ for different mRNA ensembles including 2282 sequences each, as functions of the distance $D_{out}$ between the protein binding sites. Statistics of four sequence ensembles are shown: natural human mRNAs (circles), dinucleotide-shuffled human sequences (diamonds), sequences with randomized UTRs (crosses), and natural C. elegans mRNAs (squares). All four sequence ensembles have standard deviations on the order of or greater than the thermal energy scale of $RT = 0.6\text{kcal/mol}$, implying that a biologically relevant $\Delta\Delta G$ is a general property for mRNA sequence ensembles. However, different sequence ensembles show different statistical characteristics.

3.6 The fractions of sequences that have $|\Delta\Delta G(D_{out})| > 2RT$ in different sequence ensembles: natural human mRNAs (circles), dinucleotide-shuffled human sequences (diamonds), sequences with randomized UTRs (crosses), and natural C. elegans mRNAs (squares). All four ensembles show maxima of the fraction at small $D_{out}$. However, their distance dependences are different.

3.7 The standard deviations of the cooperativity free energy $\Delta\Delta G$ of 2282 human mRNAs and 1377 CDS-replaced sequences. The two standard deviations are very similar to each other, implying that the cooperativity depends on UTRs instead of CDSs.

3.8 The cooperativity free energy of four randomly selected mRNA molecules longer than 2000nt as a function of outside distance $\Delta\Delta G(D_{out})$. Horizontal lines label the scales of $\pm RT$ and $\pm 2RT$. For all four molecule $\Delta\Delta G$ exceeds these energy scales and thus develops significant cooperativity at specific values of the outside distance $D_{out}$.

3.9 The number of microRNA-mRNA binding pairs, for which the strongest cooperativity with respect to a binding site in the 5'UTR corresponds to a given value of $\Delta\Delta G_{max}$, in bins of 0.2kcal/mol. While the number distribution of human binding pairs does not show significant differences compared to randomly shuffled binding pairs for most values of $\Delta\Delta G_{max}$, a peak at $\Delta\Delta G_{max} \approx -3.8\text{kcal/mol}$ suggest that there are a group of human sequences with strongly negative cooperativities between the 3'UTR-bound microRNA and putative 5'UTR-bound proteins.

3.10 Cooperativities along artificial evolutionary trajectories starting from five human sequences seeking to maximize the cooperativity free energy. Compared to the cooperativities in natural human sequences, which are on the order of $RT$, the artificial sequences generated by mutations support much stronger cooperativities of $|\Delta\Delta G| > 10RT$. The original sequences for initiating the evolutions are (from top to bottom at the final evolutionary step of 1000): NM_001272086, NM_001244390, NM_001104548, NM_022645, and NM_006308.
4.1 Comparison between the experimentally observed $K_d$ and the $K_{d, eff}$ calculated by two theoretical models, for the HuR binding with respect to several AU-rich sequences listed in Table 4.1. The concentration parameter $\gamma$ is determined by the regression between experimental results and the model with RNA structure, yielding $\gamma = 0.674$ with $r = 0.89$. The model considering RNA structure is more consistent with the experiments.

B.1 A one-to-one mapping from the changed structures to the structures including the certain bond. The footprint is between the $n^{th}$ and the $(n+1)^{th}$. Red dots are the nucleotides forming the base pair stack which includes the footprint. The two types of structures are taken into account in Eq. (B.1), where the first summation is for the structures in (a) and the second one is for (b). All labels for the segment lengths in the plot do not count the red dots.

B.2 The structures included in $S_{st}$. The outer segment lengths $n_1 - 1$ and $n_2 - 1$ do not include the two nucleotides forming the outer bond. The partition function for the inner segment is given by $Z_0(D)$ and thus includes an algebraic component $D^{-3/2}$ for $D \gg 1$.

C.1 Two types of structures which contribute differently in $Z_0(k_1 + k_2)$ (left) and $\tilde{Z}(k_1; k_2)$ (right). Solids lines are base pairs in $Z_0(k_1, k_2)$ and $\tilde{Z}(k_1; k_2)$, and dashed lines are the bonds in the right hand side of Fig. B.1, whose effects are necessary to be considered in the calculation of $\tilde{Z}(k_1; k_2)$. (a) A hairpin loop in $Z_0(k_1 + k_2)$ becomes a base pair stack in a stem in $\tilde{Z}(k_1; k_2)$. (b) A stack in a stem in $Z_0(k_1 + k_2)$ becomes part of a multiloop in $\tilde{Z}(k_1; k_2)$.

C.2 Two examples for the naive expectation for $S_{dd}^{(11)}$. In these two examples, the contractions from base pair stacks to single base pairs result in the same topology, whereas the explicit positions of the remaining base pairs and the distances between two inserted footprints are different. These small deviations from the $n_1^{th}$ and $(n_1 + D)^{th}$ base pairs are ignored in the naive estimation for $S_{dd}^{(11)}$.

C.3 Two configurations included in $S_{dd}^{(11)}$. The two footprints are in two consecutive stacks of the same stem, sharing a mutual bond. The partition function of these two configurations is very similar to $S_{dd}^{*}$, in which the two footprints are in the same stack.

C.4 The three types of structures included in $S_{dd}^{(11)} - S_{dd}^{*} - S_{dd, mutual}$, i.e., when both footprints are in different stems or different base pair stacks of the same stem. Two footprints are inserted between the $n_1^{th}$ and $(n_1 + 1)^{st}$ and the $(n_1 + D)^{th}$ and $(n_1 + D + 1)^{st}$ nucleotide. All labels for segment lengths in the figures do not take into account the red dots, which form the bonds nearby the footprints.

C.5 The structures which (a) contribute differently in $Z_0$ and $Z_d$ and thus included in $S_d$ but (b) contribute identically in the corresponding $\tilde{Z}$ and $\tilde{Z}_d$ and thus should not be taken into account in $\tilde{S}_d$. These structure have to be excluded when approximating $\tilde{S}_d$ as $S_d$. 

xv
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The sequences with measured $K_d$ with respect to HuR, from Meisner, et al. [1].</td>
<td>65</td>
</tr>
</tbody>
</table>
Chapter 1
INTRODUCTION

1.1 The complexity in the central dogma of molecular biology

In biology, an organism is defined by the combination of two phenomena: replication, and metabolism [2]. Replication stores and transmits the genetic information of life, and metabolism manipulates the information to perform biological functions. With the development of molecular biology in the later 20th century, a multitude of significant discoveries have unveiled the close relationship between the two phenomena. In the classic view of molecular biology, the metabolism is a series of mechanisms obeying the commands of replication [2–4]. Following this viewpoint is the statement of the central dogma proposed by F. Crick [5,6], which describes the encryption and flow of genetic information: First, the genomic information is stored in the form of DNA molecules. To release the information, DNA molecules are first transcribed to the form of RNA molecules, and then are translated to proteins. Finally, the proteins perform metabolism in the organism [3,4].

The downstream processes described in the central dogma have been kept as the basic idea of genetics for decades. However, people have been skeptical about the one-directional flow of genetic information for a long time, even since it was proposed for the first time. Actually, Crick himself has already stated that the DNA→RNA arrow was intended to be double-headed, and all the three biomolecules, DNA, RNA, and protein, might be able to regulate themselves, yielding “loops” in the flow of genetic information; however, these were not the popular perceptions at the time [4–6].

Recently, with the help of a large number of experimental observations, the flow of
genetic information has been shown to be more complicated than the direct "DNA to RNA to protein" picture [7, 8]. The transcription from DNA to RNA and the translation from RNA to protein do not merely decrypt genetic information for the use of metabolism. Instead, before being expressed in the form of proteins for metabolism, genetic information has been regulated at multiple levels [3, 4, 7, 8]. In eukaryotic organisms, the transcription of DNAs results in precursor messenger RNAs (pre-mRNA), which then undergo a series of modifications, such as splicing, capping, and polyadenylation, to become “mature” [3, 4]. Once an mRNA is mature, it is subjected to processes such as RNA localization [9, 10], translation initiation [11, 12], and RNA decay steps [13, 14], which together determine the fate of the molecule. Control over the fate of an RNA is known as post-transcriptional gene regulation, guiding an mRNA to translate to proteins at the appropriate space and time.

1.2 Combinatorial post-transcriptional regulation

The post-transcriptional regulation involves the interactions between an mRNA molecule and its binding partners, for example the RNA-binding proteins (RBPs) [15–18] or small RNA molecules such as microRNAs [19–22]. The importance of these RNA binding partners can be gauged from the fact that there are close to 800 RBPs annotated in the proteome of the human embryonic kidney [23] and 860 in HeLa cells [24], as well as 1100 and 770 microRNAs in human and mouse recorded in the microRNA.org database [25].

The post-transcriptional regulation is an important component of information processing in a cell. Many of such gene regulation processes are combinatorial, and it is therefore an interesting question how these multiple RBPs, microRNAs, and other RNA binding factors collaborate to accomplish the regulation. The binding events of these molecules have to be interdependent, or in biochemical terms “cooperative”. As the sizes of the binding partners are much smaller than the distances between them, it has been speculated for long that the effective range of this interdependency of binding partners has to be much greater than the sizes of these binding partners, to integrate numerous binding partners on a long RNA molecule.
The effective range of the interdependency can even be as large as the length of the whole mRNA molecule. Considering the fact that most RBP and microRNA binding sites are located in the untranslated regions (UTRs) at the ends of a messenger RNA [26], it is intriguing how cooperativity crossing the whole mRNA molecule is achieved, especially in light of recent evidence that the protein bridge between the 3' and 5'UTRs in eukaryotes mediated by proteins that simultaneously bind the poly(A) tail and the cap may not be essential [27]. For some transcripts the two UTRs might be brought together physically by binding of dimeric RNA binding proteins to one binding site in each UTR as recently hypothesized for the GLD-1 protein in Caenorhabditis elegans [28]. The existence of such combinatorial regulations implies a “long-range” cooperativity between the RNA binding partners. However, the detailed mechanisms of this cooperativity are still unclear.

1.3 The long-range cooperativity mediated by RNA secondary structure

In this thesis we present a natural mechanism for the long-range cooperativity that is not restricted to specific binding partners but relies on RNA structure alone [29, 30]. In our mechanism, the cooperativity is merely supported by the fundamental principles of statistical physics and thermodynamics of RNA structure folding, without requiring any further detailed properties of RBPs or microRNAs.

1.3.1 RNA secondary structure

RNA is one of the fundamental biopolymers. It plays an important role in many biological functions [3, 4, 19, 31]. Each RNA molecule is a heteropolymer consisting of the four different nucleotides A, U, G and C in a specific order called the primary structure of the molecule. These nucleotides have a strong propensity to form Watson-Crick (i.e., G–C and A–U) base pairs. As the backbone of RNA is soft and flexible, an RNA molecule is able to bend back onto itself, allowing its own nucleotides to pair with each other, and thus forming intricate patterns of stems containing runs of base pairs stacked on top of each other connected by
flexible linkers of unpaired nucleotides called a secondary structure [32].

By definition, a secondary structure of an RNA molecule, for which an example is given in Fig. 1.1, describes its configuration as a base pairing pattern without specifying the three-dimensional arrangement of the molecule. In general, RNA molecules fold first into secondary structures, and these structures then fold up into specific three-dimensional (tertiary) structures which enable some RNAs to perform catalytic functions while other RNAs mainly function as templates for transmitting genomic information.

Even in the context of a higher-order structure, base pairing, i.e. the secondary structure, is the major contribution to the total folding energy [33]. Thus, both from the dynamic and energetic point of view, secondary structure is the most important component of RNA folding. Therefore, it is appropriate to study secondary structures alone when trying to understand RNA folding phenomena; a long tradition in the field that we are also following in this thesis.

As we mentioned above, each secondary structure is described by all of its formed base pairs, denoted as \((i, j)\) for a bond between the \(i^{\text{th}}\) and \(j^{\text{th}}\) nucleotide with \(1 \leq i < j \leq N\) where \(N\) is the sequence length. Different base pairs \((i, j)\) and \((i', j')\) are either independent \((i < j < i' < j')\) or nested \((i < i' < j' < j)\). Pseudoknots, i.e. configurations with \(i < i' < j < j'\), are usually excluded to make the structures more tractable both analytically and numerically. In practice, this exclusion is reasonable: long pseudoknots are kinetically forbidden (see Fig. 1.2), and short ones do not contribute much to the total binding energy. Moreover, because the RNA backbone is highly charged and pseudoknots increase the density of the molecule, their formation is relatively disfavored in low-salt conditions [34] and can even be “turned off” in experiments to focus the study on the secondary structures [33]. Thus, pseudoknots are commonly considered part of the tertiary structure of RNA and will not be discussed further in this thesis.

1.3.2 Mechanism for the long-range cooperativity

In order to keep the language simpler, we will throughout the rest of this thesis refer to the binding sites as protein (or RBP) binding sites, although they may represent binding
Figure 1.1: A secondary structure of an RNA. The solid line shows the backbone of the RNA molecule. Dots are nucleotides, i.e. the bases of the RNA. Dashed lines are base-pair bonds. The regimes of multiple consecutive bonds are called stems, and the regions surrounded by unpaired bases are loops. Depending on the number of stems emanating from the loops, they have different names as indicated in the figure.

Figure 1.2: (a) The base-pair bonds forming pseudoknots. Pseudoknots are excluded in the discussion of secondary structures because (b) short ones do not contribute much to the total binding energy, and (c) the longer ones are kinetically unlikely to form because of their complex double-helical structures.
Figure 1.3: An example of structure-mediated cooperativity, where open boxes indicate unoccupied protein binding sites and filled boxes indicate occupied protein binding sites. $G$ is the free energy of the RNA-protein complex in each step, $G_0$ is the free energy of the RNA secondary structure alone, $\Delta G_s$ is the free energy associated with breaking the base pairs in the stem containing the two protein binding sites, $c_1$, $c_2$, $K_{D,1}$, and $K_{D,2}$ are the concentrations and dissociation constants of the first and second protein, respectively, $R$ is the gas constant, and $T$ is the temperature in Kelvin. (a) To have the first protein bind to bases in a stem of the structure in the absence of the second protein, the stem has to be broken, resulting in a free energy difference of $\Delta G_{0 \rightarrow 1} = \Delta G_s - RT \ln \left( \frac{c_1}{K_{D,1}} \right)$. (b) Once the second protein is bound to the opposing bases in the stem, the first protein can directly bind to its binding site without the loss of free energy by breaking additional base pairs, yielding $\Delta G_{2 \rightarrow 12} = -RT \ln \left( \frac{c_1}{K_{D,1}} \right) < \Delta G_{0 \rightarrow 1}$. Thus, in this case the two proteins exhibit positive cooperativity.
sites of microRNA or other binding partners as well. Qualitatively, the mechanism for this cooperativity between RBP binding sites is as follows: Many RBPs bind to single-stranded regions of RNA [35–37]. Therefore, a natural competition arises between the formation of intramolecular base pairs in the RNA and the binding of proteins to unpaired bases. A successfully binding RBP will then prohibit the formation of some originally allowed RNA structures, thus changing the statistical properties of the RNA structure ensemble, as well as the binding probabilities of other RBPs on the same RNA. It can be viewed as that the first RBP “helps” (for a positive change) or “inhibits” (for a negative change) the binding of the second RBP, and thus a “structure-mediated” cooperativity occurs between the two binding sites.

In Fig. 1.3 we exhibit an example of positive cooperativity between two RBPs. In the case of structural RNAs, for example tRNAs, the structure is essential for its specific biological function. However, in this thesis we are interested in mRNAs, which are not primarily designed to fold into a specific structure. In the ensemble of all secondary structures of an mRNA sequence, both structures with positive and negative cooperativity can exist. It is thus necessary to manipulate statistical physics to systematically take into account the combinatorial cooperativity resulting from all these different structures, as we are going to study in this thesis.

1.4 Investigation of the structure-mediated cooperativity

It is worth to comprehensively investigate the mechanism for the structure-mediated cooperativity we proposed above, as it is only based on the least required postulations, while the consequence is robust. It would be very interesting if this mechanism supports the long-range interdependency and thus offers a theory allowing multiple RBPs to communicate with each other efficiently even though they are bound on different regions on the same mRNA.

In this thesis we are going to investigate this mechanism as follows. In Chapter 2 we construct a simplified model for RBP-RNA interaction. We derive the linear RBP-RBP
correlation function in this model, as the quantity describing the structure-mediated cooperativity. Through the derivation, we show that the cooperativity is long range, and it is a generic property in all RNA sequences [29].

As we confirm that the mechanism supports the long-range characteristic of the cooperativity, in Chapter 3 we apply the mechanism to a human RNA database, focusing on the cooperativity between two binding sites at the 5’UTR and the 3’UTR respectively. For the use in molecular biology and biochemistry, we derive a characteristic difference of free energy differences, i.e. $\Delta\Delta G$, as a quantitative measure of this cooperativity. We verify that our mechanism supports biologically relevant cooperativity in natural RNA sequences. This cooperativity occurs not only between protein binding sites that are within 100 nt or so along the RNA but also between the 5’UTR and the 3’UTR, crossing the whole RNA molecule. Some intriguing interdependencies between RBPs, microRNA binding sites, and UTR sequences are also disclosed. [30].

In the last chapter, we extend our model to handle multiple sequence-specified protein binding sites. We apply this extended model to investigate the binding reactions of an RNA-binding protein, HuR, with respect to several RNA sequences. We fetch the required parameters from the data of RNAcompete experiments [36, 37]. Based on our theory, we calculate the sequence-dependent dissociation constants of HuR on these RNA sequences, and compare the results with previous experiments [1]. We discover that RNA secondary structures are crucial in the interplay between HuR and RNA sequences, verifying the importance of the structure-mediated cooperativity in realistic RNA-protein binding reactions. Several of the more technical calculations are relegated to the Appendices.
In this chapter, we consider only the case of two binding sites per RNA molecule, which is the simplest system to investigate this phenomenon in. We define a linear correlation function between the binding partners bound to an RNA molecule as the observable to quantify this interdependency, and investigate its properties with respect to the RNA structures. We will find that this linear correlation function decays algebraically as a function of the distance between two protein binding sites, $D$. We then discuss the linear correlation function for the homopolymer state in the molten phase of RNA secondary structures as well as for the heteropolymer state in the glass phase of RNA [34, 38, 39]. Such algebraically decaying correlation function provides long-range interactions between binding proteins or microRNAs on an RNA. Therefore, we conclude that long-range interdependency of binding sites that is necessary for the implementation of logical operations in post-transcriptional regulation is a generic property of RNA secondary structures and does not require any direct protein-protein interactions. Finally we also numerically establish a power-law behavior of the correlation function between RNA-binding proteins using the Vienna RNA package [40], which represents the state of the art of quantitative modeling of realistic RNA secondary structures, verifying that our simplified model does not lose the generality.

### 2.1 Theory of RNA secondary structure

In order to model RNA secondary structures, each structure $S$ needs to be assigned a folding energy $E[S]$. To calculate the energy $E[S]$ of a secondary structure $S$, it is necessary
to clarify the contributions from different constituents of the structure. It is a good approximation to consider these contributions as local, i.e. to calculate the total energy $E[S]$ by summing over the contributions from each of the independent constituents, such as the stems and loops (see Fig. 1.1). The dominant contribution from stems is the stacking energy, which is associated with two consecutive base-pair bonds and depends on the bases making up the stack. The contribution from a loop is more sophisticated. First, compared to a free chain, a loop of unpaired bases is entropically less favorable, and a free energy penalty, which depends on the loop length, has to be taken into account. Second, an enthalpy cost is involved, e.g., in bending the backbone for small loops and most importantly in the opening of a junction. The combination of these two costs has been measured as a substantial length-independent term for loop initialization plus a relatively smaller length-dependent term \[41\]. In a complete model for real RNA, all these contributions depend not only on the size of a loop but also on its type (hairpin, interior, bulge, multiloop, see Fig.1.1) and sequence.

The partition function of an RNA sequence with $N$ nucleotides is calculated by considering all its possible secondary structures, summing over all of them as

$$Z(N) = \sum_{S \in \Omega(N)} e^{-\beta E[S]},$$  \hspace{1cm} (2.1)

where $\Omega(N)$ denotes the set of all possible secondary structures for the given sequence \[34,38,42,43\], and $\beta = 1/k_BT$ follows the traditional variable definition in statistical Physics. The set $\Omega(N)$ is not only constrained by the aforementioned secondary structure definition, but also constrained by the mechanical properties of the backbone of the RNA molecule. For a real RNA, the width of the double helix prohibits the formation of loops with less than three free base pairs. All secondary structures including these small loops have to be excluded from $\Omega(N)$. Considering all these sequence- and structure-dependent factors, the calculation of the partition function $Z(N)$ is complicated, requiring consideration of thousands of parameters, and therefore can only be handled numerically. In practice, the Vienna package \[40\] implements such a complete model for the numerical calculation of
thermodynamic properties of RNA molecules.

For theoretical discussions, however, often simplified models are used to study generic properties of RNA folding. The most popular such model considers only the base-pair binding energy, instead of calculating the complicated stacking energy and loop cost. Thus, for each bond between two bases \( i \) and \( j \), the binding energy \( \varepsilon_{ij} \) is calculated as

\[
\varepsilon_{ij} = \begin{cases} 
-u & \text{if } (i, j) \text{ is an A-U or C-G pair} \\
u & \text{otherwise}, 
\end{cases}
\]  

and the total energy for a structure \( S \) is just the sum of all binding energies, 

\[
E[S] = \sum_{(i,j) \in S} \varepsilon_{ij}.
\]

The partition function can then be calculated by the recursive equation [34, 42–44]

\[
Z_{(i,j)} = Z_{(i,j-1)} + \sum_{k=i}^{j-1} Z_{(i,k-1)} e^{-\beta \varepsilon_{kj}} Z_{(k+1,j-1)},
\]

where \( Z_{(i,j)} \) denotes the partition function of the subsequence starting from the \( i \)th and ending at the \( j \)th nucleotide. The partition function of the whole sequence, \( Z(N) \equiv Z_{(1,N)} \), can thus be calculated by iterating this recursion in \( O(N^3) \) time.

A second-order phase transition has been verified in this simplified model. At high enough temperature, the RNA molecule is in the so-called molten phase. In this phase, sequence dependence of the base-pair binding energy becomes irrelevant. Instead, the RNA molecule behaves after some coarse graining like a homopolymer, with an identical binding energy \( \varepsilon_0 \) for arbitrary pairs of nucleotides rendering its partition function analytically solvable [44]. In this case, \( Z(N) \) is just a function of total sequence length, given as

\[
Z(N) \approx A_0(q) \frac{z_0^N(q)}{N^{3/2}},
\]

where \( q \equiv \exp(-\beta \varepsilon_0), \; z_0(q) = 1 + 2\sqrt{q}, \) and \( A_0(q) = \left[ z_0^3(q)/(4\pi q^{3/2}) \right]^{1/2} \) [43]. As temperature decreases to the critical temperature, the phase transition occurs. The RNA transitions into the glass phase and becomes of noticeably heteropolymer nature, with sequence-dependent binding energy for each base-pair binding. The generic properties of glass phase RNA molecules are investigated by taking the quenched average of all random
sequences, i.e. averaging over their free energies. These properties, including the occurrence of the phase transition itself, have been widely discussed numerically \cite{34,45}, and further verified by theoretical calculations on the basis of the renormalization group \cite{46,47}.

A theoretical discussion is not necessary to be constrained to this simplest model. More complete but also complicated models can be constructed by adding the loop costs or even the stacking energy back into the model of the base-pair binding energy \cite{48,49}. Generally speaking, more detailed properties can be discovered by including more free energy terms and structure constraints in the model. However, the calculations also dramatically become more difficult. To obtain a general idea of the behavior of the linear correlation function, here, instead of using the most detailed models from the start, we first restrict ourselves to the base-pair binding models, starting from the simplest one with only binding energy, and adding more free energy terms later if necessary. In the end we verify our findings using state of the art energy models including all the details necessary for quantitative RNA structure prediction.

### 2.2 Modeling of RNA-protein binding

In order to keep the language simpler, we will throughout the rest of this manuscript refer to the binding sites as protein binding sites, although they may represent microRNA binding sites as well. For the purpose of modeling RNA-protein binding events on RNA secondary structures, we consider merely the simplest and inevitable aspect: a bound protein, with a size \( l \), would exclude all structures including any one of the \( l \) base pairs in the footprint (i.e. the bases bound by the protein). For all other bases we assume that they can form the same base pairs as without the protein binding \cite{50}. Even though, in practice, more sophisticated interactions, such as the excluded volume interaction between RNA and protein, occur around the footprint, we do not include them in our minimal model, for the purpose of a coarse-grained and conceptual investigation.

In this chapter we consider only the case of two binding sites per RNA molecule, which is the simplest system in which to investigate this phenomenon. Such a system has four
possible RNA-protein binding states, as shown in Fig. 2.1. All thermodynamic quantities concerning the RNA protein interactions can be derived from the partition function of this RNA-protein system,

$$Z = Z_0 + Z_1 e^\beta \mu_1 + Z_2 e^\beta \mu_2 + Z_{12} e^{\beta(\mu_1 + \mu_2)}.$$  \hspace{1cm} (2.5)

In this expression $Z_0$ is the partition function over all secondary structures of the RNA without any protein binding; $Z_1$ and $Z_2$ are the limited partition functions, in which all bases at the first or second protein binding site are unpaired, respectively; $Z_{12}$ is the partition function for which all bases at both of the two protein binding sites are unpaired. $\mu_1$ and $\mu_2$ are the chemical potentials for the two proteins. Practically, the protein chemical potentials are controlled by their concentrations in solution as $\mu_k = \mu_{0,k} + k_B T \ln(c_k/c_0)$, where $k = 1, 2$, and $\mu_{0,1}, \mu_{0,2}$ and $c_0$ are characteristic parameters of the specific proteins, determined by experiments. The partition function can thus be rewritten as

$$Z = Z_0 + Z_1 \frac{c_1}{K_{d,1}(0)} + Z_2 \frac{c_2}{K_{d,2}(0)} + Z_{12} \frac{c_1 c_2}{K_{d,1}(0) K_{d,2}(0)},$$  \hspace{1cm} (2.6)

with $K_{d,k}(0) = c_0 e^{-\beta \mu_{0,k}}$ the bare dissociation constant for a protein binding to an otherwise unstructured RNA.

To quantify the interdependency between the protein binding sites, we introduce observables $P_k$ which are one if protein $k$ is bound and zero otherwise. If the binding sites
are independent, the thermodynamic average \( \langle P_1 P_2 \rangle \) decouples into the product \( \langle P_1 \rangle \langle P_2 \rangle \). Thus, we use \( \langle P_1 P_2 \rangle - \langle P_1 \rangle \langle P_2 \rangle \) as a measure of interdependency of the binding sites. To discover generic characteristics of all nucleic acid sequences, we investigate the *quenched average* of this protein-protein correlation function over all random RNA sequences,

\[
G \equiv \langle P_1 P_2 \rangle - \langle P_1 \rangle \langle P_2 \rangle .
\] (2.7)

In practice, we will choose random RNA sequences in which each base is chosen with equal probability from the four possibilities A, U, G, and C, independently of the other bases. While structural RNAs have very specific sequences that ensure their folding into a target structure, the random sequence model is appropriate for messenger RNAs, the sequences of which are not optimized for a specific structure and which are anyways the more interesting targets for post-transcriptional regulation. The linear correlation function is then calculated as

\[
G = c_1 c_2 \frac{\partial^2 \ln Z}{\partial c_1 \partial c_2} = \left( \frac{(Z_0 Z_{12} - Z_1 Z_2)c_1 c_2}{Z^2 K_{d,1}^{(0)} K_{d,2}^{(0)}} \right).
\] (2.8)

In the following, we will investigate this linear correlation function as a function of the distance, \( D \), between the two protein binding sites.

### 2.3 RNA-protein binding on the simplest RNA folding model

As described in Sec. 2.1, our investigation of the linear correlation function for protein binding sites starts from the simplest model, which includes only the base-pair binding energy defined in Eq. (2.2). Note that although we mentioned in Sec. 2.1 that there is a minimum size for hairpin loops in real RNA, we do not impose such a constraint in this model for the purpose of a conceptual discussion. We first consider the high-temperature regime, where the RNA is in the molten phase, with the partition function given in Eq. (2.4). Due to translational invariance, the limited partition functions also retard to functions of sequence length parameters \( D, n_1, n_2, \) and \( l \), defined in Fig 2.2, and can be written as
\[ Z_1 = Z_d(n_1, D + l + n_2), \]
\[ Z_2 = Z_d(n_1 + l + D, n_2), \]
\[ Z_{12} = Z_{dd}(n_1, D, n_2), \]  

where \( Z_d \) and \( Z_{dd} \) are limited molten phase partition functions with one and two stretches of \( l \) unpaired bases, respectively (see Fig. 2.2), and the lengths of the segments are constrained by the length of the whole molecule as \( n_1 + D + n_2 + 2l = N \). The exact form of \( Z_d \) and \( Z_{dd} \) can be derived by considering the following insertion aspect: a limited partition function \( Z_d(n, m) \) is constructed by inserting a blank segment, with the length equal to the footprint \( l \), into the partition function \( Z_0(n + m) \), between the \( n^{th} \) and \( (n + 1)^{th} \) nucleotides. For the model including only pair binding energies, inserting such a blank segment does not affect the energies of any of the structures and thus Eq. (2.9) is further simplified to

\[ Z_1 = Z_2 = Z_0(n_1 + D + l + n_2) = Z_0(N - l), \]
\[ Z_{12} = Z_0(n_1 + D + n_2) = Z_0(N - 2l). \]  

The molten phase correlation function is then given by

\[ g(D) \equiv G(D) \frac{K_{d,1}^{(0)} K_{d,2}^{(0)}}{c_1 c_2} = \frac{Z_0(N)Z_0(N - 2l) - Z_0(N - l)^2}{Z^2(N, l, \{c_k/K_{d,k}^{(0)}\})}, \]  

completely independent of the distance between the protein binding sites and converges to 0 as fast as \( (l/N)^2 \), as can be seen by substituting Eq. (2.4) for all \( Z_0 \), yielding

\[ g(D) = \frac{3z_0^{-2l}}{2[(1 + c_{1,1}^{(0)}/z_0 K_{d,1}^{(0)})[(1 + c_{2,2}^{(0)}/z_0 K_{d,2}^{(0)})]^2 \left( \frac{l}{N} \right)^2 + O \left( \frac{1}{N^3} \right)}. \]  

We conclude that the model including only base pair bonds is not able to explain protein-binding correlations.

2.4 RNA-protein binding on the RNA folding model with constant loop cost

Since the simple energy model does not result in correlations between the binding sites, we are now going one step further than the simplest base-pair binding model and consider
Figure 2.2: In the molten phase, the base-pair binding energies of different nucleotides become identical. The limited partition functions with consideration of protein binding sites no longer depend on sequences and retard to functions of lengths of segments between binding sites.

a loop cost. As discussed in Sec. 2.1, the complete form of the loop cost includes a constant term for loop initialization and a length-dependent term for extension. Since the constant term is generally much greater than the length-dependent term, it is appropriate to take loops into account only through this constant term. Although this loop cost has in principle entropic and enthalpic components, we follow the literature [41] and model it when varying temperature as purely entropic, i.e. take loops into account through a temperature-independent Boltzmann factor, \( \exp(-s_0) < 1 \). We note, however, that this choice only affects the detailed temperature dependence and not the presence of protein-protein correlations per se.

In order to compute the partition function for an RNA folding model with the pairing energies given in Eq. (2.2) and constant loop cost \( s_0 \), two different auxiliary partition functions are required. They are \( Z_{b(i,j)} \), the partition function for structures on a substrand starting at the \( i^{th} \) nucleotide and ending at the \( j^{th} \) one, having a bond between the first and last nucleotides, and \( Z_{(i,j)} \), the partition function starting from the \( i^{th} \) nucleotide and ending at the \( j^{th} \) one without any further constraints. These quantities obey the recursion equations [44,51–53]

\[
Z_{b(i,j)} = g_{ij} \left[ Z_{b(i,j-1)} + \xi (Z_{(i+1,j-1)} - Z_{b(i+1,j-1)}) \right] \quad \text{and} \\
Z_{(i,j)} = Z_{(i,j-1)} + \sum_{m=i}^{j-1} Z_{(i,m-1)} Z_{b(m,j)}, \tag{2.13}
\]
2.3 Structures in which the footprint is in a stem, i.e., inserted between two consecutive base-pair bonds. Such structures contribute differently in $Z_0(n + m)$ (left) and $Z_d(n, m)$ (right).

where $\xi \equiv \exp(-s_0)$ and $q_{ij} \equiv \exp(\beta \varepsilon_{ij})$, which can be iterated numerically to compute the full partition function $Z_{(1,N)}$ for arbitrary sequences and loop penalties $\xi$ in $O(N^3)$ time. Again, we neglect constraints on the size of hairpin loops for simplicity.

2.4.1 Molten phase

Similar to the simplest model of only base-pair binding energy, the recursion in Eq. (2.13) can be analytically solved for a molten-phase RNA with loop cost by substituting a homogeneous $q$ for all $q_{ij}$, yielding (see [48, 49, 54] and Appendix A)

$$Z_0(N) = A(q, \xi) \frac{z^N(q, \xi)}{N^{3/2}} [1 + O(N^{-1})], \quad (2.14)$$

where $A$ and $z$ are non-universal parameters depending on the values of $q$ and $\xi$. Comparing Eqs. (2.14) and (2.4), the partition function with and without the entropy cost have exactly the same form — the loop cost is irrelevant for the asymptotic behavior as also known in the context of DNA melting [55] for a long time.

However, for the limited partition functions $Z_d$ and $Z_{dd}$, the relationships corresponding to Eqs. (2.10) no longer hold when the loop cost is non-zero, i.e. $Z_d(n, m) \neq Z_0(n + m)$. This can be seen as follows: Once an unpaired segment of length $l$ is inserted into $Z_0(n + m)$, it can either create a new loop, or extend an existing loop. The secondary structures taken into account by $Z_0(n + m)$ are thus separated into two groups, reacting differently to the insertion. If a structure of $Z_0(n + m)$ has a stem cross the $n^{th}$ and $(n + 1)^{st}$ base pairs, the insertion of the footprint between the two nucleotides changes the contribution of this structure to $Z_d(n, m)$ by a loop factor $\xi$ since the insertion creates a new loop as shown.
in Fig. 2.3. For all other structures, the contribution to $Z_0(n + m)$ and $Z_d(n, m)$ are the same. It is therefore necessary to distinguish which structure belongs to which one of the two groups. Defining the partition function for all structures with a stem containing the $n^{th}$ and $(n + 1)^{st}$ base pairs (i.e., the partition function for all structures for which the insertion of a footprint generates a new loop) as $S_d(n, m)$, the limited partition function can be expressed as

$$Z_d(n, m) = Z_0(n + m) - (1 - \xi)S_d(n, m).$$

$$\equiv Z_0(n + m) + C_d(n, m) \quad (2.15)$$

Notice that $C_d(n, m) = -(1 - \xi)S_d(n, m)$ is calculated as the contribution to $Z_0(n + m)$ in the absence of the inserted footprint. The first term $Z_0(n + m)$ is just the one used in the simplest model — albeit itself dependent on $\xi$ as given by Eq. (2.14) — and the subsequent term, proportional to $(1 - \xi)$, is the effect of the footprint insertion resulting from a nonzero loop cost.

A similar strategy as Eq. (2.15) can also be applied to the calculation of $Z_{dd}$ by defining its changed term via

$$Z_{dd}(n_1, D, n_2) = Z_0(n_1 + D + n_2) + C_{dd}(n_1, D, n_2).$$

However, since now there are two footprints of the protein, $C_{dd}$ is more complicated than $C_d$ and cannot be simply written down as a term proportional to $(1 - \xi)$. Considering that each insertion of an unpaired stretch of bases into a base stack of a stem (no matter how many stretches are inserted into a stack) changes the contribution of the structure by a factor of $\xi$, four different configurations, affected differently by the insertions, are included in $C_{dd}$:

(i) The first footprint is in a stem but the second one is not, contributing a change of $(\xi - 1)$.

(ii) The second footprint is in a stem but the first one is not, contributing a change of $(\xi - 1)$.

(iii) Both footprints are in different stems or different base pair stacks of the same stem, contributing a change of $(\xi^2 - 1)$. 

18
(iv) Both footprints are on opposite sides of the same base stack of a stem, contributing a change of \((\xi - 1)\).

We describe the partition functions for the first three configurations with the help of the partition functions \(S_{dd}^{(ab)}\), where the labels \(a\) and \(b\) describe constraints on the first and the second binding site respectively. A label of 1 for \(a\) or \(b\) denotes that the corresponding binding site is in a stack, a label of 0 denotes that the binding site is not in a stack, and a label of \(\times\) indicates that there is no constraint for the corresponding binding site. In addition, we introduce the partition function for all configurations in which the locations of both footprints are on opposite sides of the same stack as \(S_{dd}^{*}\). The changed term \(C_{dd}\) can then be written as

\[
C_{dd}(n_1, D, n_2) = (\xi - 1)S_{dd}^{(10)} + (\xi - 1)S_{dd}^{(01)} + (\xi^2 - 1)(S_{dd}^{(11)} - S_{dd}^{*}) + (\xi - 1)S_{dd}^{*} \\
= (\xi - 1)(S_{dd}^{(1x)} - S_{dd}^{(11)}) + (\xi - 1)(S_{dd}^{(x1)} - S_{dd}^{(11)}) + (\xi^2 - 1)(S_{dd}^{(11)} - S_{dd}^{*}) + (\xi - 1)S_{dd}^{*} \\
= (\xi - 1)(S_{dd}^{(1x)} + S_{dd}^{(x1)}) + (\xi - 1)^2 S_{dd}^{(11)} + \xi (1 - \xi) S_{dd}^{*}.
\]

(2.17)

where we omit the arguments \((n_1, D, n_2)\) for each \(S_{dd}^{(ab)}\) and \(S_{dd}^{*}\) for the sake of clarity. We note that the limited partition functions \(S_{dd}^{(1x)}\) and \(S_{dd}^{(x1)}\), since they only contain constraints at the location of one of the footprints, can be exactly expressed as

\[
S_{dd}^{(1x)}(n_1, D, n_2) = S_d(n_1, D + n_2), \text{ and} \\
S_{dd}^{(x1)}(n_1, D, n_2) = S_d(n_1 + D, n_2)
\]

(2.18)

The limited partition function \(Z_{dd}\) is thus given as

\[
Z_{dd}(n_1, D, n_2) = Z_0(n_1 + D + n_2) \\
- (1 - \xi)[S_d(n_1, D + n_2) + S_d(n_1 + D, n_2)] \\
+ (1 - \xi)^2 S_{dd}^{(11)}(n_1, D, n_2) + \xi(1 - \xi)S_{dd}^{*}(n_1, D, n_2).
\]

(2.19)

At this point, we have investigated the effect of protein binding site insertions on all the limited partition functions required to calculate the molten phase correlation function \(g(D)\) for this model with constant loop cost. Unfortunately, the quantities \(S_d\), \(S_{dd}^{(11)}\) and \(S_{dd}^{*}\), can
not be calculated exactly and we have to make two approximations. First, we consider the
limit \(N/2 \approx n_1 \approx n_2 \gg D \gg l\), where \(N = n_1 + D + n_2 + 2l\) is the length of the whole
RNA molecule. Second, we investigate the correlation function \(g(D)\) only perturbatively in
the loop cost or more precisely as an expansion in \((1 - \xi)\). This investigation, the details
of which are given in the appendices, shows that already an infinitesimally small loop cost
will yield a non-zero correlation function that shows a power law dependence on \(D\). We will
later demonstrate numerically (see Fig. 2.4) that this result remains valid in the biologically
relevant regime of small \(\xi\) (or \((1 - \xi) \approx 1)\).

More specifically, expanding the limited partition functions to the appropriate orders of
\((1 - \xi)\) and then inserting these expansions into the definition of the correlation function
\(g(D)\) as shown in Appendices B and C yields
\[
g(D) = \frac{Z_0(N)Z_{dd}(n_1, D, n_2) - Z_d(n_1, D + l + n_2)Z_d(n_1 + l + D, n_2)}{Z(N, D, l, \{c_k/K_{d,k}^{(0)}\})^2}
= (1 - \xi) \frac{A(q, \xi, \{c_k/K_{d,k}^{(0)}\})}{D^{3/2}} + O\left(\frac{1}{N}\right),
\]
in the limit of \(N \gg D \gg l\), where \(N\) is sequence length, \(D\) is distance between protein
binding sites, and \(l\) is the footprint of protein binding. The prefactor \(A\) converges to
\(A(q, \xi)z(q, \xi)^{-2(l+2)}q^2\xi^2\) in the limit \(c_k \ll K_{d,k}^{(0)}\) and \(1 - \xi \ll 1\), and has to be determined
numerically in the general case. The leading order term, which obeys a power law with an
exponent of \(3/2\), is the zeroth order term of \(1/N\) and thus does not vanish as \(N \to \infty\).

As we have pointed out above, the analytical result that the correlation function \(g(D)\)
obey\(s\) the power law is only a perturbative result for small \((1 - \xi)\), i.e., small loop costs.
It is somewhat reassuring that the first and second order terms in this expansion show the
same power law. Nevertheless, loop costs in real RNA tend to be large which is why in Fig.
2.4 we numerically verify the occurrence of the same power-law \(g(D)\) in the molten phase
model with a substantially large loop cost, \(s_0 = 5\). The combination of the perturbative
calculation and the numerical evidence confirms that any non-zero loop cost contributing
to the partition function as \(\xi \neq 1\) qualitatively changes the properties of the protein-protein
correlation function leading to a \(D^{-3/2}\) long-range correlation between the multiple binding
Figure 2.4: Numerical calculation of the protein-protein correlation $g(D)$ on RNA polymers of $N = 200, 600, 1000$, and $2000$ based on the molten-phase homopolymer model, with $q = \exp(0.2)$, $s_0 = 5$ and $c_i/K_d^{(0)} = 0$. The footprint is set to be $l = 6$. The correlation satisfy a power law $g(D) \sim D^{-3/2}$ (solid line) in the regime $N \gg D \gg l$.

2.4.2 Glass phase

In reality, it is believed that RNA molecules at room temperature are not in the molten but rather in the so-called glass phase [34, 38, 39]. Thus, it is necessary to investigate the correlation function $G(D)$ in that phase. When temperature decreases, the differences between the $\varepsilon_{ij}$ become relevant, and the disorder breaks the homopolymer assumption for RNA molecules. Therefore, the limited partition functions can no longer be appropriately expressed in terms of the translationally invariant $Z_d$ and $Z_{dd}$, and Eq. (2.20) is not
applicable, either. Moreover, the effect of the denominator in Eq. (2.8) and thus of the protein-binding parameters \( c_k/K_{d,k}^{(0)} \) is now uncertain. These parameters may now play an important role in the form of the correlation function, instead of only modifying the prefactor as in Eqs. (2.11) and (2.20).

To clarify the effect of protein binding parameters, we consider the ratio \( Z_0/Z_k \), where \( k = 1, 2 \), and compare its value with the corresponding protein-binding term \( c_k/K_{d,k}^{(0)} \), to quantify the protein concentrations. We thus introduce the effective dissociation constant for each individual RNA sequence by taking into account the free energy difference between free and bound RNAs,

\[
\Delta F_k \equiv -k_B T \ln(Z_0/Z_k).
\]  

(2.21)

The generic effective dissociation constant for all random RNA sequences is then derived by considering the quenched average of the free energy difference, \( \overline{\Delta F_k} \), yielding

\[
K_{d,k} = K_{d,k}^{(0)} e^{-\beta \overline{\Delta F_k}}
\]  

(2.22)

Using this concept, we consider the concentration dependence of \( G(D) \) in the following three different regimes: (i) dilute concentration, where \( c_k \ll K_{d,k} \), (ii) saturated concentration, where \( c_k \gg K_{d,k} \), and (iii) normal concentration, where \( c_k \approx K_{d,k} \). In the dilute regime, proteins essentially never bind to the RNA while in the saturated regime they remain nearly always bound. Thus, most biochemical reactions occur in regime (iii).

In Fig. 2.5 we show the correlation function of the simplified model using base pairing energies from Eq. (2.2), a loop cost of \( s_0 = 5 \), and random sequences with equal probabilities for the four possible nucleotides for normal protein concentrations (\( c_k/K_{d,k} \approx 4 \)). A similar calculation resulting from a shorter footprint is shown in the inset (\( l = 1 \) compared to the \( l = 6 \) in the main graph) to allow an evaluation of the effect of footprint size. Again, we observe a power law dependence of the correlation function on the distance between the binding sites. We verified this power law dependence in the normal protein concentration regime numerically for a whole range of temperatures \( k_B T/u = 0.1 \sim 0.9 \) (which spans both sides of the glass transition as determined by a peak in the specific heat in the vicinity of
$k_BT \approx 0.3u$, data not shown).

However, in contrary to the molten phase results in the last subsection, it turns out that protein concentration is an important factor in the glass phase. Fig. 2.6 shows the correlation function for the same model and parameters as above but for protein concentrations in the dilute and the saturated regime, respectively. In this case, the correlation function does not show a power law behavior. These discoveries suggest that while the power-law correlation is sensitive to temperature and protein binding parameters, it generically occurs precisely in the biologically relevant regime where $c_k \approx K_{d,k}$.

### 2.5 RNA-protein binding on the RNA folding model with loop-size dependent cost

In Sec. 2.1 we explained that the loop cost comprises initialization and extension components. In practice, this extension penalty is due to the entropic loss upon forming a closed loop, and is thus proportional to $\ln(L)$ with $L$ the loop length [48]. Taking this logarithmic dependence into account numerically results in an algorithm for RNA partition functions of complexity $O(N^4)$, much less efficient than the aforementioned $O(N^3)$ one. Thus, current state of the art RNA structure prediction tools such as MFOLD [56] and the Vienna package [40] linearize the loop cost for multiloops and interior loops by approximating $\ln(L) \approx L - 1$, supposing that loops with very large $L$, where the difference between the logarithmic and linear dependence becomes noticeable, are extremely energetically unfavorable and thus forbidden.

Thus, we consider here the effect of such a linear loop cost. The total Boltzmann factor for a loop with $L$ free bases now becomes $\exp(-s_0 - \nu L)$ with $\nu$ the extension penalty per free base in a loop. To take into account the effect of the extension penalty, one more auxiliary partition function, $Z_{m,(i,j)}$ for a substrand from the $i^{th}$ to the $j^{th}$ base in the context of a closed base pair $(i',j')$ with $i' < i < j < j'$ is required in addition to the partition functions $Z_{(i,j)}$ and $Z_{b(i,j)}$ used in Eq (2.13). The new recursive equations for calculating the three partition functions are then given by
Figure 2.5: Numerical calculation of the correlation function $G(D)$ on RNA polymers of 200 nucleotides for normal protein concentrations at low temperature, $k_B T = 0.3 u$. The loop cost is $s_0 = 5$. The quenched average is taken over 400,000 random sequences. Footprints are $l = 6$ in the main graph and $l = 1$ in the inset. Protein binding parameters are $K_{d,1}^{(0)} = K_{d,2}^{(0)} = 0.1 \text{nM}$. Concentrations are $c_1 = c_2 = 100 \text{nM}$, yielding $c_k/K_{d,k} \approx 3.84$ by $e^{eta \Delta F_k} \approx 0.00384$ for both $k = 1, 2$ in main graph, and $c_1 = c_2 = 2 \text{nM}$ to have $c_k/K_{d,k} \approx 3.72$ by $e^{eta \Delta F_k} \approx 0.186$ in the inset. The correlation functions of the two graphs follow power laws as $G(D) \sim 1/D^{0.9}$. 
Figure 2.6: Numerical calculation of the protein-protein correlation \( g(D) \) on model RNA polymers of 200 nucleotides for (a) extremely dilute and (b) extremely saturated protein concentrations. The quenched average is taken over 400,000 random sequences. Temperature is set to \( k_B T = 0.3u \) and the secondary structure parameters are the same as those of Fig. 2.5. No power law dependence of the correlation function appears.

\[
Z_b(i,j) = q_{ij} \left[ Z_b(i+1,j-1) + \xi(Z_m(i+1,j-1) - Z_b(i+1,j-1)) \right],
\]

\[
Z_m(i,j) = \tilde{\xi} Z_m(i,j-1) + \sum_{k=i}^{j-1} Z_m(i,k-1) Z_b(k,j),
\]

\[
Z(i,j) = Z(i,j-1) + \sum_{k=i}^{j-1} Z(i,k-1) Z_b(k,j),
\]

(2.23)

with \( \tilde{\xi} \equiv \exp(-\nu) \). Again, we neglect constraints on the size of hairpin loops for simplicity.

The extension loop cost results in a new property different from those in the aforementioned two RNA folding models, yielding very different asymptotic RNA secondary structure partition functions. This new property becomes apparent by adding a free energy contribution of \( k_B T \nu \) to every base, which amounts to a physically irrelevant overall shift in all free energies by \( N k_B T \nu \). For paired bases this amounts to shifting the pairing energy to \( u' = u + 2k_B T \nu \). For all bases inside of a closed bond (i.e. the ones included in the partition function \( Z_m \)), this transformation just retards the model back to the one including only a constant loop cost; for the unpaired bases outside of any bond,
however, it results in an additional *gain* in free energy of $k_B T \nu$. Such a gain in free energy on the free bases can be viewed as the work resulting from a constant force stretching the RNA molecule, $k_B T \nu = f \cdot x$, thus mapping an RNA molecule with extension loop cost to a molecule with constant loop cost under tension.

A second order phase transition has been discovered in such RNA molecules under tension [48,57,58]. In the homopolymer case, as the stretching force is weak, the asymptotic RNA partition function still holds the form $Z_0(N) \approx A z^N / N^{3/2}$. Once the force crosses a threshold, the phase transition occurs, and the asymptotic partition function becomes $Z_0(N) \approx A' z^N$, a purely exponentially increasing function of sequence length [57,58]. In our model with extension loop cost, the phase transition is effectively driven by tuning the temperature and thus tuning the parameter $u' / k_B T - 2 \nu$. At high temperatures, the $\nu$-term is dominant and thus the RNA molecule is in the stretched phase characterized by the purely exponential partition function. Below a critical temperature, the $\nu$-term is irrelevant and the molecule is in the regular molten or glass phase.

According to the calculation in Appendix D, the phase transition occurs at

$$q = q_c = \frac{(1 - \bar{\xi})(2 - \bar{\xi})^2}{(1 - \tilde{\xi})(1 - \tilde{\xi})^2 + \xi(2 - \tilde{\xi})^2},$$

(2.24)

in which $\xi$ and $\tilde{\xi}$ are defined in Eq. (2.23), and $q$ is the identical $q_{ij}$ in the molten phase. For realistic RNAs, $1 \gtrsim \bar{\xi} \gg \xi$ is generally satisfied, and thus the ratio between the effective molten-phase binding energy and temperature is approximately

$$\frac{\epsilon_{\text{molten}}}{T} = \log(q_c) \simeq 2(1 - \tilde{\xi}),$$

(2.25)

which is much less than one and thus it is convincingly that the stretching transition is typically far above the glass transition.

If the parameters are chosen such that the molecule is below the critical temperature of the stretching transition but above the glass transition temperature, i.e. where $Z_0(N) \approx A z^N / N^{3/2}$, we can again calculate the protein-protein correlation function. In fact, this calculation is very similar to the one for the RNA folding model including a constant loop cost. For most structures, their contributions to the partition function $Z_d(n, m)$ (where a
protein binding footprint is to be inserted after the \(n^{th}\) base pair) are simply the ones for \(Z_0(n + m)\) multiplied by a penalty factor \(\tilde{\xi}^l\) or \(\xi \tilde{\xi}^l\), since either the inserted \(l\)-base long footprint extends an existing loop or creates a new loop, respectively. Considering these two types of configurations, the limited partition function \(Z_d\) for a model with extension loop cost is simply the one for the model with constant loop cost multiplied by \(\xi \tilde{\xi}^l\) (with appropriate parameters \(A(q, \xi, \tilde{\xi})\) and \(z(q, \xi, \tilde{\xi})\)). There are some structures, in which the footprint neither creates a new loop nor extends an existing loop but is “naked” (see Fig. 2.7). These structures contribute identically in \(Z_d(n, m)\) and \(Z_0(n + m)\), i.e., without an additional factor \(\tilde{\xi}^l\). However, the partition function for these naked structures is simply \(Z_0(n)Z_0(m) = O(z^N N^{-3})\), which can be ignored in the limit of large \(N/2 \approx n \approx m\) compared to the leading order term in \(Z_d\) of order \(O(z^N N^{-3/2})\) as we have seen numerous time in the calculations for the model with constant loop cost. Similarly, in \(Z_{dd}\) (the partition function into which two footprints of length \(l\) are to be inserted), the limited partition functions for structures with at least one naked footprint are all at the order of \(O(z^N N^{-3})\) and thus negligible, leading to a \(Z_{dd}\) identical to the one for the model including constant loop cost multiplied by \(\xi \tilde{\xi}^{2l}\).

In consequence, for a model including a linearized extension loop cost, once the asymptotic partition function holds the form \(Z_0(N) \approx A z^N / N^{3/2}\), the asymptotic limited partition functions \(Z_d\) and \(Z_{dd}\) are identical to those in the model including only a constant loop cost up to factors \(\tilde{\xi}^l\) and \(\xi \tilde{\xi}^{2l}\), respectively, and thus the protein-protein correlation function has the same form as that in Eq. (2.20), with only additional factors of \(\tilde{\xi}^l\) multiplying the protein concentrations.

Figure 2.7: A structure in which the footprint is naked. Dashed lines represent the two independent substrands in the calculation for partition functions. Such structures contribute identically in \(Z_d(n, m)\) and \(Z_0(n + m)\).
Since in practice the extension cost per free base in a loop, $\nu$, is much less than the loop initialization cost $s_0$, the critical temperature of the force-driven transition is typically high enough such that the force-driven phase transition occurs earlier than the glass-molten phase transition as temperature decreases. The properties of the glass phase (limited) partition functions are thus believed not to be affected by the stretched phase and thus to be similar to those of the model including only a constant loop cost. Therefore, we also expect a similar behavior of the correlation function to that in the model with constant loop cost. This expectation is verified numerically in Fig. 2.8, where the correlation function between two protein binding sites, again, decays as a power law of the distance between the sites in the normal concentration regime. It is worth pointing out that the quenched average over random sequences for the correlation of the model with loop extension penalty converges faster than that of the model with only constant loop penalty and yields an even more convincing power law. This might imply that, in the glass phase, for the model with extension penalty, the energy distribution of secondary structures has a narrower peak around the global maximum; for the model with only constant penalty, however, the energy distribution is wider, and thus the average has to be taken over much more random sequences for convergence. The relationship between the energy landscape in the glass phase and the loop penalty as a function of loop length would be a valuable topic for more discussion in the future.

2.6 RNA-protein binding on a realistic RNA folding model

In the previous sections, we have established that the protein-protein correlation function is a power law in serveral simplified models of RNA folding as long as they include a loop cost. While simplified models are useful to disentangle the mechanism of and the minimum requirements for the power law behavior, an obvious question is if this finding is an artifact of the simplified model or if it applies to real RNA molecules as well. To this end, we numerically study the same protein-protein correlation function as before using the Vienna package [40], which represents the state of the art in quantitative RNA secondary structure
Figure 2.8: Numerical calculation of correlation function $G(D)$ on RNA polymers of 200 nucleotides for normal protein concentrations at low temperature, $k_B T = 0.3u$. The loop cost is $s_0 = 5$ and $\nu = 0.5$. The quenched average is taken over 100,000 random sequences. Protein binding parameters are $K_{d,1}^{(0)} = K_{d,2}^{(0)} = 0.1\text{nM}$ and $c_1 = c_2 = 100\text{nM}$, with $e^{\beta \Delta F_k} \approx 0.000401$ for both $k = 1, 2$. The footprint is equal to $l = 6$ base pairs. The correlation function follows a power law as $G(D) \sim 1/D^{0.8}$. 
prediction using thousands of measured free energy parameters including the very important stacking free energies. We apply the constraint folding capabilities of the Vienna package to exclude the footprints (i.e., the bases bound by the protein) from participating in base pairing, which allows us to calculate the limited partition functions $Z_1$, $Z_2$, and $Z_{12}$ for arbitrary RNA sequences and distances between the protein binding sites.

Based on our experience with the simplified models, we choose the protein concentrations and binding constants in the biologically reasonable regime (iii) where the binding sites are neither essentially empty nor completely saturated. We average over 100,000 random sequences of length $N = 200$ bases with equal probability for each of the four nucleotides. The resulting protein-protein correlation function, shown in Fig. 2.9, again follows a power law.

We would like to point out that, even though our simplified model and the Vienna package render very similar exponents for the power-law correlation, we do not give much credence to the precise value of this value, since it is dependent on strong finite-size effects. E.g., it appears that the effective exponent depends somewhat on protein concentrations and its absolute value decreases (i.e. the power law decays slower) as the ratios $c_k/K_{d,k}^{(0)}$ increase. The reason for such tendency is still unclear and requires more investigation in the future. However, independent of the precise value of the exponent, the important fact is that the protein-protein correlation function does not decay exponentially but rather has a fat (power-law) tail which enables interdependency between proteins binding at long distances from each other along the molecule.

### 2.7 Summary

According to our model, the ensemble of RNA secondary structures can be viewed as a medium for this long range interaction. When one protein or microRNA binds to the RNA, it also changes the ensemble, and the change of the ensemble transmits the effect of this binding partner to the other binding partner, thus resulting in an interplay between them. Following this concept, we have quantified this interplay, the cooperativity between two
Figure 2.9: Numerical calculation of protein-protein correlation $G$ on RNA polymers of 200 nucleotides. The calculation is based on the Vienna package and the quenched average is taken over 100,000 random sequences. Temperature is set to $T = 37^\circ C$. The parameters for protein binding are $c_1/K_{d,1} = c_2/K_{d,2} = 100$ with $e^{\beta \Delta F_k} \approx 0.00410$, which can be easily achieved by RNA binding proteins and are chosen to allow the proteins to compete with the RNA base pairing without completely outcompeting it. The length of the binding sites (footprint) is equal to 6 bases. Distance is expressed by the number of base pairs (bp) between the two binding sites. This log-log plot shows a regression to a power-law decay function $G(D) \sim 1/D^{0.9}$. 
binding proteins, at least for the simplest two-partner case. We considered coarse-grained models for RNA secondary structures, and discovered that a long-range cooperativity between single-stranded binding proteins (or microRNAs) on RNA, which expresses itself as a power-law protein-protein correlation function, occurs when a loop penalty is taken into account in the model.

The model has been well discussed in detail through analytical and numerical approaches. For the RNA in the high-temperature molten phase, we analytically derived a power-law correlation function, and verified this result numerically. In the low-temperature glass phase, we numerically calculated the cooperativity and discovered that this long-range phenomenon is strongly affected by protein concentrations, but occurs precisely at those biologically meaningful protein concentrations at which the RNA is neither fully saturated with proteins nor completely unoccupied. In biology, this cooperativity, thus can play an important role in combinational gene regulation on the post-transcriptional level. It is therefore worth to theoretically and experimentally investigate the phenomenon discovered here on the level of generic sequences for specific naturally occurring RNA molecules and their protein binding partners.

These consequences of the coarse-grained models are verified by the Vienna package, which is viewed as the state of the art for the numerical simulation of real RNA molecules. A length-dependent loop penalty has also been included as an advanced discussion for the issue of loop cost, and a similar long-range cooperativity remains. It is thus believed that the constant loop initiation penalty, the one we focused on in the very first discussion, is the critical factor for the occurrence of the long-range effect.
Chapter 3

Cooperativity in natural mRNAs between the 5’ and 3’ UTRs

In Chapter 2 we have focused on establishing that this RNA secondary structure mediated cooperativity is a long-range effect, and used simplified models of RNA folding to understand this behavior for random sequences. In this chapter we apply the mechanism to a human RNA database, focusing on the cooperativity between two binding sites at the 5’UTR and the 3’UTR respectively. We use the Vienna package \cite{Vienna} to numerically calculate this end-to-end cooperativity. We will show that this cooperativity (i) occurs for natural sequences, (ii) is of a biologically relevant order of magnitude of several kcal/mol, (iii) not only occurs between protein binding sites that are within 100nt or so along the RNA but also between the 5’UTR and the 3’UTR, and (iv) is determined by the UTR sequences with little effect from the coding regions.

3.1 Cooperativity in terms of free energy

We follow the strategy in the theoretical discussion in Chapter 2, considering only the simplest effects of the binding of the protein on the RNA secondary structure \cite{29,50}: (i) A bound protein prevents all nucleotides in its footprint from base pairing and (ii) binding of a protein to the RNA results in a free energy gain of $RT \ln(c/K_D)$, where $c$ is the concentration of the protein in solution, $K_D$ is the dissociation constant of the protein from the specific site in the RNA, which contains all the details of the interaction between the protein and the RNA, $R$ is the gas constant and $T$ is the temperature in Kelvin \cite{59}. We
also constrain ourselves to the case that there are only two proteins, $P_1$ and $P_2$, binding one binding site each on a given RNA molecule, which is the simplest system to investigate protein-protein cooperativity in [29]. Consequently, there are four possible RNA-protein complexes, distinguished by whether the first and/or the second protein binding site is occupied or not, which we label as $RNA$, $RNA^{(1)}$, $RNA^{(2)}$, and $RNA^{(12)}$ as illustrated in Fig. 2.1.

With the two proteins $P_1$ and $P_2$, all RNA-protein binding reactions can be expressed as

$$RNA + P_1 + P_2 \xleftrightarrow{\Delta G_{0 \rightarrow 1}} RNA^{(1)} + P_2$$

$$RNA^{(2)} + P_1 \xleftrightarrow{\Delta G_{2 \rightarrow 12}} RNA^{(12)}.$$  

For each of the four RNA-protein complexes, different sets of base pairs can participate in base pairing interactions. We calculate the free energies $G^{(0)}$, $G^{(1)}$, $G^{(2)}$, and $G^{(12)}$ of the secondary structures for each of the four complexes using the constraint folding capabilities of the Vienna package to exclude the nucleotides “hidden” in the footprints of bound proteins from base pairing. To obtain the free energies of the entire complexes the protein binding free energy $RT \ln(c_1/K_{D,1})$ has to be subtracted from $G^{(1)}$ and $G^{(12)}$ and the protein binding free energy $RT \ln(c_2/K_{D,2})$ of the second protein has to be subtracted from $G^{(2)}$ and $G^{(12)}$. The free energy differences $\Delta G$ of the four reactions are given by

$$\Delta G_{0 \rightarrow 1} = G^{(1)} - G^{(0)} - RT \ln(c_1/K_{D,1}), \quad (3.1a)$$

$$\Delta G_{0 \rightarrow 2} = G^{(2)} - G^{(0)} - RT \ln(c_2/K_{D,2}), \quad (3.1b)$$

$$\Delta G_{1 \rightarrow 12} = G^{(12)} - G^{(1)} - RT \ln(c_2/K_{D,2}), \quad (3.1c)$$

$$\Delta G_{2 \rightarrow 12} = G^{(12)} - G^{(2)} - RT \ln(c_1/K_{D,1}). \quad (3.1d)$$
The cooperativity between the two binding sites can be quantified by comparing the binding free energies $\Delta G$ of the reactions on opposite sides of this diagram, e.g., the difference in binding free energy $\Delta G_{0 \to 1}$ for the binding of protein $P_1$ in the absence of protein $P_2$ and the binding free energy $\Delta G_{2 \to 12}$ associated with binding of protein $P_1$ in the presence of protein $P_2$ (or vice versa with the two proteins interchanged which leads to the same result). This yields the cooperativity free energy

$$
\Delta \Delta G \equiv \Delta G_{0 \to 1} - \Delta G_{2 \to 12} = \Delta G_{0 \to 2} - \Delta G_{1 \to 12}
$$

which becomes independent of the protein concentrations $c_1$ and $c_2$. Therefore, for an arbitrary RNA sequence, once the two protein binding sites are assigned, the Vienna package can be used to calculate the structural free energies $G^{(0)}$, $G^{(1)}$, $G^{(2)}$, and $G^{(12)}$, from which the free energy difference $\Delta \Delta G$ quantifying the cooperativity can be deduced via Eq. (3.2). For a positive cooperativity, a protein already bound helps the other protein bind, leading to $\Delta G_{2 \to 12} < \Delta G_{0 \to 1}$ and thus a positive $\Delta \Delta G$; for a negative cooperativity, the situation is reversed and $\Delta \Delta G$ is negative.

### 3.2 Preparation of RNA sequences and protein binding sites

#### 3.2.1 Sequences

We select human mRNA sequences in NCBI’s RefSeq database [60] to study cooperativity between 5’ and 3’UTRs. As a comparison to gauge the influence of the length of the UTRs, we also investigate *Caenorhabditis elegans* mRNAs from the WormBase database [61], in which the UTRs are much shorter than those in human mRNAs.

Some of the sequences are deposited in these databases including their poly(A) tails. Since these tails, as homopolymers, contribute very little to the RNA secondary structure, and in order to treat all mRNAs consistently no matter if they were submitted to the databases with or without poly(A) tails, we discard all poly(A) tails. Specifically, we remove any runs of three or more consecutive adenines at the 3’ ends of the sequences in the
database. We are aware that in eliminating the poly(A) tails, we lose the ability of applying our results to the important class of poly(A) binding proteins (PABP) \cite{62-64}; however, since the poly(A) tails should not form secondary structures, PABPs are not expected to be subject to structure-mediated cooperativity, and thus are irrelevant in this study.

Since RNA secondary structure prediction scales as the third power of sequence length, computational resources constrain us to focus on the shorter mRNAs in the database. Dictated by computational feasibility, we thus choose the range from 500 to 1500 bases (counted after dropping the poly(A) tails). For human sequences, we select those sequences in which both 5' and 3'UTRs are longer than 110 bases to ensure that all our binding sites are always in the UTRs even for the greatest outside distance, $\max(D_{\text{out}}) = 200\text{nt}$, yielding a total of 2282 sequences. For \textit{C. elegans} sequences, we select the sequences with short UTRs by constraining the UTR lengths between 30 and 150 bases, yielding a total of 1277 sequences.

Since we want to identify which, if any, regions of the human mRNAs might have evolved to generate cooperativity, we also generate several sets of artificial RNA sequences based on our set of human sequences: (a) We shuffle each of the 2282 selected human sequences, generating an ensemble of random sequences with equal dinucleotide frequencies \cite{65}. (b) We shuffle the UTRs of each of the 2282 human sequences to generate an ensemble of random sequences with equal dinucleotide frequencies in the UTRs but with unchanged coding sequences (CDSs). (c) We select sequences that have UTR lengths in the range from 110nt to 400nt and replace their CDSs by the CDS of the single sequence NM004242 (300nt), rendering an ensemble of totally 1377 artificial sequences with a fixed CDS.

In order to show that our results are not limited to short mRNA sequences, we also selected four individual mRNA sequences with more than 2000 bases, namely PTP4A1, MARCKS, MYC, and PPP1R15B for verification purposes (it is computationally feasibly to study a few individual long mRNAs, but averaging over the entirety or even just a large number of these is not possible).
3.2.2 Locations of protein binding sites

We are especially interested in the cooperativity between the proteins bound on the 5’ and 3’UTRs close to the 5’ and 3’ ends of the mRNAs, respectively. Therefore, we assign the two protein binding sites of each mRNA in their 5’ and 3’ ends, and vary their distances from the ends of the RNAs (note that in our minimal model the binding site of each protein is specified explicitly, i.e., that we assume perfect sequence specificity for the sequence at whatever binding site we choose without actually specifying that sequence). The relative location of the two binding sites is quantified by the outside distance, $D_{out}$, which is defined in Fig. 3.1 as the sum $D_{out} = n_1 + n_2$ of the distances $n_1$ and $n_2$ of the two binding sites from their respective sequence ends. In order to avoid the very significant computational cost of varying the two distances $n_1$ and $n_2$ independently, we consider only the situation that the two binding sites are symmetric, $n_1 = n_2$, when we calculate the $\Delta\Delta G(D_{out})$ of an ensemble of mRNAs or of several long sequences; we do not expect any qualitative differences for asymmetric situations. In our studies involving microRNA binding sites described below, we choose the known microRNA binding site in the 3’UTR and vary the location of the putative protein binding site in the 5’UTR. The footprint of the proteins is set to be $l = 6$nt in agreement with typical footprint sizes of RNA binding proteins.

3.3 Biologically significant cooperativity in mRNA sequences

As described in Sec. 3.2, in order to demonstrate that RNA secondary structure mediated protein binding cooperativity is a generic effect providing cooperativity between proteins binding at the two ends of an RNA molecule, we select 2282 human mRNA sequences. We choose one protein binding site to be close to the 5’ end of each molecule and one close to the 3’ end and systematically vary their distance from their respective ends such that the sum $D_{out}$ of the two distances from the ends ranges from 0nt to 200nt (see Fig. 3.1). We calculate the cooperativity free energy $\Delta\Delta G$ for each “outside” distance $D_{out}$ and each of the 2282 sequences.
Figure 3.1: Definition of the outside distance, $D_{out}$, as the sum of the distances from each of the two protein binding sites to their corresponding ends in units of nt. Black dots represent bases and the two blocks represent the protein binding sites with footprint length $l$ (in this figure $l = 4\text{nt}$). In our numerical calculation we restrict ourselves to the symmetric case $n_1 = n_2$, and use a footprint size of $l = 6\text{nt}$.

3.3.1 Distance dependence of structure-mediated cooperativity

Fig. 3.2 shows the mean and standard deviation of the cooperativity free energy $\Delta \Delta G$ over all 2282 sequences as a function of the outside distance $D_{out}$ of the two protein binding sites. The mean cooperativity free energy remains within about 0.1kcal/mol of zero, more or less independent of the outside distance between the protein binding sites. That in principle leaves two possibilities, namely that either RNA secondary structure mediated cooperativity is a very small effect or that it is a significant effect but roughly equal numbers of sequences show positive as negative cooperativity. The standard deviation allows us to differentiate between these two scenarios: the fact that it starts out above 1kcal/mol when the two binding sites are very close to their respective ends and then falls to about 0.5kcal/mol for outside distances of 30nt and stays in this range for binding sites up to 100nt away from their respective ends (yielding a total outside distance of 200nt) implies that many sequences must have cooperativity free energies with absolute values of this order of magnitude.
Figure 3.2: The mean (circles) and standard deviation (diamonds) of the cooperativity free energy $\Delta \Delta G$ of 2282 human sequences, as a function of the outside distance $D_{\text{out}}$ between the two protein binding sites. The mean of the cooperativity free energy $\Delta \Delta G$ is close to zero for all outside distances, while its standard deviation is on the order of or greater than the thermal energy scale of $RT = 0.6 \text{ kcal/mol}$. The dashed vertical lines label two different outside distances $D_{\text{out}}$ for which the whole distribution of the cooperativity free energies are compared in Fig. 3.3. The maximum of the standard deviation of the cooperativity free energy occurs at $D_{\text{out}} = 8 \text{ nt}$. 
3.3.2 Structure-mediated cooperativity is significant for large numbers of mRNA molecules

In order to classify cooperativity free energies as significant or not, it is convenient to compare them to the thermal energy of $RT$, which is approximately 0.6kcal/mol at temperatures of 37°C. Free energy differences below this magnitude will not have a measurable effect on protein binding in the thermal environment of a cell. Free energy differences larger than $RT$ do result in significant protein binding cooperativity. The horizontal dashed lines in Fig. 3.2 indicate cooperativity free energy levels of $RT$ and $2RT$ and show that the standard deviation of the cooperativity free energy is on the order of $RT$ in the distance range of $30\text{nt} \lesssim D_{\text{out}} \lesssim 150\text{nt}$. Thus a good fraction of molecules must have cooperativity free energies larger than $RT$ even at these outside distances. This is illustrated in Fig. 3.3 for two specific outside distances, namely $D_{\text{out}} = 8\text{nt}$ where both proteins are very close to their respective ends and the standard deviation of the cooperativity free energy has its maximum, and $D_{\text{out}} = 150\text{nt}$, where the standard deviation of the cooperativity free energy starts to fall below $RT$. The figure illustrates how many sequences have a given cooperativity free energy in bin sizes of 0.2kcal/mol. As suggested by the small mean, the distributions of cooperativity free energies are symmetric with equal numbers of molecules showing positive as negative cooperativity. It can also be seen that the distribution at $D_{\text{out}} = 8\text{nt}$ is much wider than the distribution at $D_{\text{out}} = 150\text{nt}$ as expected from their standard deviations shown in Fig. 3.2. The inset of Fig. 3.3 shows the same distributions as the main figure but on a logarithmic scale that emphasizes the long tails of the distribution indicating some sequences with very large cooperativity free energies.

An additional way to quantify the significance of the cooperativity free energies is to count the number of molecules out of the 2282 tested that have a cooperativity free energy above a certain threshold for each outside distance $D_{\text{out}}$. Fig. 3.4 shows this data for thresholds of $RT$ and $2RT$. The data shows that when the proteins bind very close to their respective ends of the RNA molecule, about half of the totally 2282 investigated
Figure 3.3: The number of mRNA molecules with respect to given cooperativity free energies $\Delta \Delta G$ (in bin sizes of 0.2 kcal/mol) at outside distances between the protein binding sites of $D_{\text{out}} = 8$ (red solid line) and $D_{\text{out}} = 150$ (green dashed line). The inset shows the same data on a logarithmic scale. The two distributions are both centered at 0. The distribution for $D_{\text{out}} = 8$ (when the protein binding sites are close to their respective ends of the molecule) extends much wider than that for $D_{\text{out}} = 150$ when the protein binding sites are further away from the ends of the molecule. Vertical black dashed and dash-dotted lines label $|\Delta \Delta G| = RT$ and $|\Delta \Delta G| = 2RT$, respectively.
Figure 3.4: The number of mRNA sequences with biochemically relevant protein binding cooperativity among the 2282 investigated human mRNA sequences. The two data sets correspond to $|\Delta \Delta G|/k_B T > 1$ (circles) and $|\Delta \Delta G|/k_B T > 2$ (diamonds). The data shows that at any given distance $D_{out}$ a significant fraction of molecules has a significant cooperativity free energy.

sequences have $|\Delta \Delta G| > RT$ and a quarter have $|\Delta \Delta G| > 2RT$ (at $D_{out}=6$nt), implying that a significant cooperativity between two protein binding sites, of which one is at the 5’ end of a molecule and the other is at the 3’ end of a molecule, is a generic property among messenger RNAs. In addition, even at an outside distance of 200nt (each protein binds 100nt from its respective end of the RNA) still 10% of the 2282 messenger RNAs studied show a cooperativity free energy of at least $RT$ and more than a hundred show a cooperativity of at least $2RT$. 
3.3.3 Cooperativities in human mRNAs and random RNA sequences have different statistical characteristics

In our previous work we verified that structure mediated cooperativity is a generic property of all RNA molecules, stating that cooperativity exists in an ensemble of random RNA sequences [29]. In this study we investigate the cooperativity in human, and thus evolved, mRNAs. Thus, we are curious about what the similarities and differences between the cooperativities for a “random” and a “natural” sequence ensemble are.

To compare with the human mRNA sequences, we thus randomly permute nucleotides in each of the 2282 selected human sequences, and generate two ensembles of 2282 random sequences (see Materials and Methods): (a) random sequences generated by shuffling the entire sequence of each human mRNA such that the dinucleotide frequencies are identical to those of the original human mRNAs, and (b) random sequences generated by shuffling only the 5’ and 3’UTR sequences of each human mRNA while leaving the CDS unaffected. We investigate cooperativity in these two random sequence ensembles, and compare their statistical properties with those of the natural human sequences.

The standard deviations of $\Delta \Delta G(D_{out})$ of the two random ensembles and the human mRNA ensemble are shown in Fig. 3.5. We notice that all these three ensembles comprise many sequences with biologically relevant cooperativities, i.e. $\Delta \Delta G$ is on the order of or greater than 0.6kcal/mol. However, the standard deviation of the human mRNA ensemble has a large peak at small $D_{out}$, and then sharply decreases to merely around $RT$ at $D_{out} \geq 50$nt. Instead, the standard deviation of the completely shuffled random sequences smoothly decays in the entire investigated region, $0$nt $\leq D_{out} \leq 200$nt. The standard deviation of the UTR-shuffled sequences is in between those of the human mRNA and fully shuffled sequence ensembles.

The fraction of sequences that have $|\Delta \Delta G| > 2RT$ in the human mRNA and the two random sequence ensembles are shown in Fig. 3.6. Similar to the standard deviations, for the two random sequence ensembles, the fractions of sequences with strong cooperativities decrease smoothly as $D_{out}$ increases, whereas for human mRNAs the fraction decreases
sharply once $D_{out}$ becomes greater than that for the maximum at $D_{out} = 8$nt. Taken together, a human mRNA is likely to have strong cooperativity between the ends of the 5' and 3'UTR, while this cooperativity decreases rapidly as the two UTR binding sites are farther from their respective ends. On the other hand, random sequences perform power-law like, smoothly decaying cooperativity with respect to $D_{out}$. We thus conclude that human RNAs are different from random RNAs in terms of their cooperativity properties. While one might have expected stronger cooperativity in the evolved human sequences than in their randomly shuffled counterparts, we would like to point out that we, as discussed above, for computational reasons only look at symmetrically located binding sites of the two proteins. These will in general not correspond to the true protein binding sites on those natural mRNAs and it is conceivable that a possible design for cooperativity at true binding sites in fact leads to the suppression of cooperativity at the symmetrically located positions we interrogate.

### 3.3.4 Cooperativity is driven by the UTRs and not by the coding sequences

Since we have shown in the previous section that replacing the UTRs with shuffled sequences yields similar cooperativities as shuffling the entire sequence, we here ask the inverse question and retain the UTRs while replacing the CDSs of the ensemble of human mRNAs by one fixed CDS (see Materials and Methods). Fig. 3.7 shows the comparison of the standard deviations of $\Delta\Delta G(D_{out})$ of human mRNAs and the CDS-replaced sequences. The two ensembles of sequences show very similar profiles. This similarity suggests that the cooperativity between the 5' and 3'UTR binding sites is determined by the two UTR sequences, while the coding region has little effect.

### 3.3.5 Cooperativities in *C. elegans* mRNAs are weaker than those in human mRNAs

Compared to human mRNA sequences, *C. elegans* mRNAs comprise much shorter UTRs [60, 61]. As we have confirmed in the last section that the cooperativity is driven by the UTRs,
Figure 3.5: The standard deviations of the cooperativity free energy $\Delta \Delta G$ for different mRNA ensembles including 2282 sequences each, as functions of the distance $D_{out}$ between the protein binding sites. Statistics of four sequence ensembles are shown: natural human mRNAs (circles), dinucleotide-shuffled human sequences (diamonds), sequences with randomized UTRs (crosses), and natural C. elegans mRNAs (squares). All four sequence ensembles have standard deviations on the order of or greater than the thermal energy scale of $RT = 0.6\text{kcal/mol}$, implying that a biologically relevant $\Delta \Delta G$ is a general property for mRNA sequence ensembles. However, different sequence ensembles show different statistical characteristics.
Figure 3.6: The fractions of sequences that have $|\Delta \Delta G(D_{out})| > 2RT$ in different sequence ensembles: natural human mRNAs (circles), dinucleotide-shuffled human sequences (diamonds), sequences with randomized UTRs (crosses), and natural C. elegans mRNAs (squares). All four ensembles show maxima of the fraction at small $D_{out}$. However, their distance dependences are different.
Figure 3.7: The standard deviations of the cooperativity free energy $\Delta\Delta G$ of 2282 human mRNAs and 1377 CDS-replaced sequences. The two standard deviations are very similar to each other, implying that the cooperativity depends on UTRs instead of CDSs.
it is natural to investigate the cooperativities of an ensemble of sequences that have short UTRs. Thus, we choose an ensemble of 1277 C. elegans mRNAs of which both UTRs are in the range from 30 to 150 bases and the total sequence lengths are between 500 and 1500 bases. Their standard deviation of $\Delta\Delta G(D_{out})$ and the number of sequences having significant $\Delta\Delta G(D_{out})$ are plotted in Figs. 3.5 and 3.6, respectively. The two figures show two consistent consequences. First, the cooperativities in C. elegans mRNAs are also biologically relevant, i.e. $|\Delta\Delta G| \gtrsim RT$. Second, although cooperativities in C. elegans mRNAs are biologically relevant, they are significantly weaker than the cooperativities in human sequences. The latter result is additional evidence for the strong relationship between cooperativity and UTR sequences.

### 3.3.6 RNA structure mediated cooperativity between the sequence ends persists in long mRNA molecules

While computational complexity does not allow us to include mRNA molecules longer than 1500nt in our systematic study above, we selected four individual mRNA molecules with more than 2000nt to determine if our findings of significant cooperativity between 5’ and 3’ ends of an mRNA hold for longer molecules as well. Fig. 3.8 shows the cooperativity free energy of these four molecules as a function of outside distance $D_{out}$. It can be seen that all four chosen molecules have some outside distances $D_{out}$ at which $|\Delta\Delta G| > 2RT$ and even though for two of them (MYC and PPP1R15B) the binding sites with such large cooperativity free energy are very close to the ends, all molecules also have cooperativity free energies in excess of $1RT$ at some outside distance between 50nt and 150nt.

### 3.4 Human microRNA binding sites

A large number of microRNAs bind to the 3’UTRs of mRNA sequences to regulate their abundance [22]. At the same time, proteins are bound to the 5’UTRs, e.g., see [28,66]. Thus, in addition to the symmetric binding sites with $n_1 = n_2$, studied so far, we also investigate the cooperativities between these “natural” microRNAs binding at their experimentally
Figure 3.8: The cooperativity free energy of four randomly selected mRNA molecules longer than 2000nt as a function of outside distance $\Delta\Delta G(D_{\text{out}})$. Horizontal lines label the scales of $\pm RT$ and $\pm 2RT$. For all four molecules $\Delta\Delta G$ exceeds these energy scales and thus develops significant cooperativity at specific values of the outside distance $D_{\text{out}}$. 
verified sites in the 3'UTR and putative 5'UTR binding proteins.

### 3.4.1 Preparation and measure of the RBP-microRNA cooperativity

We fetch microRNA binding sites from microRNA.org [25] using a strict mirSVR score cutoff of $-1$ [67]. Again, we select only the mRNA sequences with total lengths from 500nt to 1500nt and UTR lengths above 110nt for computational reasons. We limit the distance between the microRNA binding site and the 3’ end of the mRNA to be less than 100 bases, i.e. $n_2 \leq 100$, because we want to focus on the cooperativity between microRNAs bound on the end and proteins bound at the beginning of the same mRNA sequence. The above-mentioned conditions yield 1173 microRNA-mRNA binding pairs, in which some of the mRNA sequences correspond to more than one microRNA binding site.

We investigate each such microRNA-mRNA pair independently, i.e. we consider only one specific microRNA interacting with its bound mRNA at a time.

The cooperativity is calculated between the microRNA and a protein binding site around the end of the 5’UTR. We set the protein footprint to be $l = 6$nt, and investigate all possible protein binding sites from exactly at the end ($n_1 = 0$nt) to 100 bases from the end of 5’UTR ($n_1 = 100$nt). For each microRNA binding site in a 3’UTR, we search through putative protein binding sites in the first 100 bases in the 5’UTR of the corresponding mRNA, and record the cooperativity free energy $\Delta\Delta G$ of the binding site with the greatest absolute value of the cooperativity free energy $\Delta\Delta G$, termed $\Delta\Delta G_{\text{max}}$.

As a comparison, multiple random ensembles of microRNA-mRNA binding pairs are generated. In each random ensemble the microRNA binding sites, which are located by their $n_2$ on their “naturally” bound mRNA, are randomly assigned to a different mRNA. The strongest cooperativities of the reassigned microRNA-mRNA pairs are recorded as a reference to compare the cooperativities of the natural microRNA-mRNA pairs to.
3.4.2 Human microRNA binding sites show significant negative cooperativities

After investigating the 1173 microRNA-mRNA binding pairs, we note that although the position of the putative protein binding site in the 5’UTR is selected by maximizing the absolute value of the cooperativity free energy, the chosen cooperativity free energy still can be either positive or negative. As a comparison, we randomly shuffle the positions of the microRNA binding sites, matching them with different mRNAs. The $\Delta \Delta G_{\text{max}}$ calculated from these randomly shuffled microRNA-mRNA binding pairs help us evaluate the significance of the characteristics discovered in the natural microRNA-mRNA pairs.

Fig. 3.9 shows the distribution of the $\Delta \Delta G_{\text{max}}$ in the 1173 microRNA-mRNA binding pairs in bin sizes of 0.2kcal/mol, with the grey regions representing the distribution of 10 ensembles of randomly shuffled binding pairs. The dark grey region for each cooperativity free energy bin designates the range of frequencies observed among the 10 ensembles while the light grey region indicates the regime of four standard deviations. The figure illustrates that for most cooperativity free energies, the ensemble of natural binding pairs does not show significant differences from the ensembles of randomly shuffled binding pairs. However, in the region of negative $\Delta \Delta G_{\text{max}}$, the natural binding pairs exhibit a characteristic difference from random fluctuations, namely a peak at $\Delta \Delta G_{\text{max}} \approx -3.8\text{kcal/mol} \approx -6RT$ labeled in Fig. 3.9. The number of microRNAs with cooperativities of this magnitude is greater than its expected frequency based on the 10 random sequence ensembles by more than four standard deviations. We conclude that this peak is statistically significant, suggesting that a group of microRNAs result in strongly negative cooperativites with respect to proteins bound to the 5’UTRs. Considering the role of microRNAs in silencing and suppression of gene regulation [22], the negative cooperativity we discover in the statistics of human microRNA-mRNA binding pairs is reasonable and, furthermore, implies that a microRNA can participate in combinatorial regulation with proteins binding far from its binding site.
Figure 3.9: The number of microRNA-mRNA binding pairs, for which the strongest cooperativity with respect to a binding site in the 5’UTR corresponds to a given value of $\Delta \Delta G_{\text{max}}$, in bins of 0.2 kcal/mol. While the number distribution of human binding pairs does not show significant differences compared to randomly shuffled binding pairs for most values of $\Delta \Delta G_{\text{max}}$, a peak at $\Delta \Delta G_{\text{max}} \approx -3.8$ kcal/mol suggest that there are a group of human sequences with strongly negative cooperativities between the 3’UTR-bound microRNA and putative 5’UTR-bound proteins.
3.5 Artificial evolution of highly cooperative sequences

In order to find out how strong cooperativities between the 3’UTR and the 5’UTR can possibly become, we do not restrict ourselves to natural sequences but rather search for the maximum of the cooperativity free energy $\Delta \Delta G$ among a huge number of possible sequences using an evolutionary algorithm. We choose five natural sequences, namely NM_006308, NM_022645, NM_001104548, NM_001244390, and NM_001272086, as starting sequences. We calculate the maximum over all $|\Delta \Delta G|$ for all possible pairs of protein binding sites. Considering the massive computational resources required for a search through all possible protein binding site pairs (on the order of $N^3 \times N^2 \sim N^5$ in which $N$ is the length of sequence), we compromise on merely searching protein binding sites in the region of 5nt $\leq n_1, n_2 \leq 25$nt, as we are more interested in the cooperativity between the ends of the two UTRs. In each evolutionary step, we randomly choose and mutate a nucleotide of the sequence, and recalculate the maximum of $|\Delta \Delta G|$ for all protein binding pairs. If the maximum is greater than that of the previous sequence, we keep the new sequence; otherwise we continue with the previous sequence. This evolutionary step is repeated 1,000 times.

In Fig. 3.10 we show evolution trajectories starting from five human sequences over 1000 mutation steps. At the end of the artificial evolution trajectories all five human mRNAs mutate to artificial sequences comprising very strong cooperativities. Three sequences NM_001104548, NM_006308, and NM_022645 eventually mutate to sequences with $|\Delta \Delta G_{\text{max}}| \approx 20RT$, and NM_001244390 and NM_001272086 have even higher final cooperativities of $|\Delta \Delta G_{\text{max}}| \approx 30RT$. This result shows that appropriately designed RNA molecules can support cooperativities much stronger than typically found in natural mRNA sequences. However, considering that most biochemical reactions occur at energy scales of a few $RT$, such large $\Delta \Delta G$ are actually unrealistic and inappropriate for biological functions as the strong interactions they would confer would not be reversible thus rendering regulation impossible.
Figure 3.10: Cooperativities along artificial evolutionary trajectories starting from five human sequences seeking to maximize the cooperativity free energy. Compared to the cooperativities in natural human sequences, which are on the order of $RT$, the artificial sequences generated by mutations support much stronger cooperativities of $|ΔΔG| > 10RT$. The original sequences for initiating the evolutions are (from top to bottom at the final evolutionary step of 1000): NM_001272086, NM_001244390, NM_001104548, NM_022645, and NM_006308.
3.6 Summary

In this chapter we have introduced a quantitative measure of RNA secondary structure based protein binding cooperativity in terms of free energy difference. We have demonstrated that significant cooperativity between protein binding sites at the two ends of an RNA molecule is a generic feature of RNA secondary structure formation. We have come to this conclusion by systematically studying the ensembles of human mRNAs, *C. elegans* mRNAs, and random sequences generated from human mRNAs, in which the sequence lengths are between 500nt and 1500nt, and verified the effect in selected longer human mRNAs.

The cooperativity free energy $\Delta \Delta G$ rises up to a biologically relevant value (i.e. $|\Delta \Delta G| > RT \approx 0.6$ kcal/mol) in more than 1000 among the 2282 human mRNAs we investigated. Our statistical investigation of these sequences shows that this cooperativity gets stronger as the two binding sites get closer to their respective ends, suggesting a significant “end-to-end” cooperativity between two proteins or microRNAs that bind on the opposing ends of an mRNA molecule. We interpret this observed cooperativity to imply that the two binding partners, albeit far from each other when measured in terms of nucleotides along the mRNA sequence, can actually be close to each other in physical and secondary structure space due to the intricate secondary structures of the mRNA molecule. In consequence, the binding of the two partners becomes interdependent, and this effect allows for combinatorical post-transcriptional regulation.

We compared the cooperativity free energies $\Delta \Delta G(D_{\text{out}})$ of human mRNAs and two ensembles of the random sequences generated from the human mRNAs. We discovered that all three sequence ensembles support biologically relevant cooperativity, $|\Delta \Delta G| > RT$ and have maximal cooperativities when the binding sites are relatively close to the ends of the UTRs. However, in human mRNAs the cooperativity $\Delta \Delta G$ decreases rapidly as the distance $D_{\text{out}}$ increases; on the other hand, the cooperativity $\Delta \Delta G(D_{\text{out}})$ of random sequences decays smoothly, as shown in Fig. 3.5. We view the difference as an evidence that human mRNAs have distinguishable cooperativity properties compared to random sequences and note that the fact that human sequences appear to show a weaker cooperativity than their random
counterparts can be a consequence of us only looking at symmetrically located putative binding sites that likely do not coincide with the true pairs of binding sites in human UTRs. Consistently with this, when we investigated experimentally annotated microRNA binding sites, we did find a group of microRNA binding sites with a strong cooperativity free energy around $\Delta \Delta G = -3.8 \text{kcal/mol} \approx -6RT$ that is statistically overrepresented compared to randomly permuted sites.

An intriguing phenomenon about RNA secondary structure induced cooperativity is its symmetry in terms of the sign of the cooperativity. The symmetry of the distribution of cooperativity free energies shown in Fig. 3.3 implies that the number of mRNAs with positive cooperativity is roughly equal to the number of those with negative cooperativity. Whether this symmetry is a generic property of RNA or if it is specific to the ensembles of mRNAs studied here, is an interesting topic for future investigations.

Another interesting fact we find about the cooperativity between the two ends of an RNA is that it appears exclusively determined by the 5’ and 3’ UTR sequences. Our investigation of C. elegans mRNAs shows a positive relationship between UTR length and cooperativity strength. Moreover, an ensemble of sequences generated by substituting a fixed CDS for the CDSs of human mRNAs shows very similar cooperativities to the ensemble of original human mRNAs. These two results support the picture that the 5’ and 3’ UTRs “communicate” with each other with little interference from the CDS.

In our study we first chose the protein binding sites to be symmetrically arranged, i.e., at equal distances from their respective ends of the mRNA molecule. Based on the properties of RNA secondary structure, we would expect that statistically asymmetric configurations of the protein binding site will behave the same as the symmetric configurations with the outside distance $D_{out}$ being the only relevant parameter. However, when studying the cooperativity free energy of a specific molecule, being able to vary the positions of both binding sites independently should dramatically increase the search space and thus enable even higher cooperativity free energies. Unfortunately, the computational cost of varying the two positions independently is prohibitive since each combination of binding sites requires a new folding of the entire molecule.
However, we were still able to investigate asymmetric configurations in some limited cases. In our study of microRNAs, the 3’UTR binding sites are fixed, and thus we were able to tune the 5’UTR binding sites and search through a range of different $n_1$ with respect to the asymmetric, fixed $n_2$. Also, when we looked for the maxima of cooperativity in artificially evolved RNA sequences, we constrained the calculation of $\Delta \Delta G$ in a small range of asymmetric $n_1$ and $n_2$, as a compromise given the limitations imposed by computational cost. In this case we found very significant cooperativity free energies up to $|\Delta \Delta G_{\text{max}}| \approx 30RT$.

An important aspect for future investigations is the inclusion of sequence specificity. For the purpose of this work, we assumed perfect sequence specificity and allowed the protein (or microRNA) to only bind at one pre-defined location along the mRNA. With techniques like RNAcompete \cite{36, 37} systematically measuring the affinities of many important RNA binding proteins to all possible binding sites, it should be possible to allow sequence dependent binding of two proteins to be incorporated into RNA secondary structure prediction and we will pursue this avenue in order to identify mRNAs and pairs of proteins that actually employ this mechanism of post-transcriptional regulation and could be verified experimentally. Lastly, while our study entirely focused on thermodynamical quantities, studying the kinetics of the interplay of RNA structure and protein binding will be important as well.
Chapter 4

SEQUENCE-DEPENDENT PROTEIN BINDING AFFINITY

In this chapter, we extend our model for RNA-protein binding, from the simplest model including only two binding sites to a model taking into account all sequence dependent binding motifs. Based on this model, we are able to investigate the interplay between a large number of sequence-dependent protein binding sites, instead of being limited to the simplest systems including merely two pre-defined protein binding sites. To calculate realistic RNA-protein binding phenomena, we exploit the sequence-dependent protein binding data from RNAcompete experiments [36, 37]. We will first describe the extended model for sequence-dependent protein binding sites, and how we apply the RNAcompete data to this model. We will show that the structure-mediated cooperativity discussed in previous chapters plays a crucial role in the interplay between these large numbers of RNA-binding proteins (RBPs) on RNA secondary structures. As an evidence, we compare our theoretical calculation with experiments, and verify that this cooperativity is necessary for describing the observed binding affinities of a protein HuR on several RNA sequences.

4.1 Sequence specificity of RNA-binding proteins

In previous chapters, we have verified the long-range cooperativity between RNA-binding proteins (RBPs), and we have discovered that it results in intriguing phenomena in human RNA sequences. In these theoretical and statistical investigations, we manipulated a simplified protein-binding system such that there are essentially two pre-defined protein
binding sites on an RNA molecule. However, in reality, an RBP recognizes its binding motifs, and binding to these specific short sequence regions in RNA molecules is highly preferable. Consequently, the RBP binding sites are sequence specified, instead of location specified as in our simplified model. Extending this idea of sequence specificity, an RBP can actually bind on an arbitrary region in an RNA, while the binding probability for different regions can be significantly different.

In this chapter, we take into account these sequence-specified RBP binding regions. For a given type of RBPs with a footprint \( l \)nt, there are \( 4^l \) different short sequence segments for its RNA binding. If the binding affinities of an RBP on each of these \( 4^l \) segments are known, we could then separate each long RNA sequence into these small \( l \)-mers, and construct a binding profile of this long RNA sequence with respect to this type of RBP. This profile is a result of the interplay between multiple RBPs binding on different regions of the RNA molecule, mediated by not only the competition between the overlapping binding sites, but also the ensemble of RNA secondary structures. We thus extend our simplified model of merely two binding sites to multiple binding sites, and are able to deal with the sequence specificity, instead of being constrained to the pre-defined RBP binding sites.

4.2 Model for multiple RBP binding sites on an RNA molecule

Here we describe a model in which an RNA molecule includes \( n \) protein binding sites with different binding affinities. We consider that in this system there is only one type of RBP, and thus only one protein concentration \( c \) is required in the model. For the number \( i \) binding site among these \( n \) binding sites, a bare (i.e. without considering the effect from RNA secondary structure) dissociation constant \( K^{(0)}_{d,i} \) is given. The partition function of this RNA-RBP system becomes

\[
Z[c] = Z_0 + \sum_i Z_i \frac{c}{K^{(0)}_{d,i}} + \sum_{i<j} Z_{ij} \frac{c}{K^{(0)}_{d,i} K^{(0)}_{d,j}} + \sum_{i<j<k} Z_{ijk} \frac{c}{K^{(0)}_{d,i} K^{(0)}_{d,j} K^{(0)}_{d,k}} + ..., \tag{4.1}
\]
in which \( Z_0 \) is the partition function of the RNA molecule without any bound RBP, and the limited partition functions, \( Z_i \), \( Z_{ij} \), \( Z_{ijk} \), and so on, take into account all RNA secondary structures in which all nucleotides in the protein binding sites labeled in their subscriptions are unpaired.

We then discuss the interplay between multiple protein binding sites and the RNA secondary structures on the basis of this partition function. We focus on the effective binding affinity of a type of RBP on an RNA sequence. As an RNA molecule can be bound by multiple RBPs, the general description based on the reaction

\[
RNA + RBP \leftrightarrow RNA \cdot RBP
\]  \hspace{1cm} (4.2)

is not valid. Instead, there is a chain of binding reactions,

\[
RNA + RBP^n \leftrightarrow RNA \cdot RBP + RBP^{n-1} \leftrightarrow RNA \cdot RBP^2 + RBP^{n-2} \leftrightarrow .... \]  \hspace{1cm} (4.3)

Thus, the effective binding affinity is a combinatorial effect of all RBP bindings. The binding probability is then defined as the probability that at least one RBP is bound on the RNA, which is given by the ratio between the partition function for all RBP-bound RNAs and the whole partition function of the RBP-RNA system, yielding

\[
P_{bound}[c] = \frac{Z[c] - Z_0}{Z[c]}, \hspace{1cm} (4.4)
\]

with respect to the given RBP concentration \( c \). In an experiment, the effective dissociation constant \( K_{d,eff} \) is measured as the concentration rendering \( P_{bound} = 50\% \), leading to

\[
\frac{Z[K_{d,eff}] - Z_0}{Z[K_{d,eff}]} = \frac{1}{2}, \hspace{1cm} (4.5)
\]

The observed dissociation constant between a given RNA sequence and a type of RBP can thus be numerically derived as the solution of Eq. (4.5).
4.3 The sequence-specified bare dissociation constants

In this section we describe how we exploit the data of RNAcompete experiments to generate the sequence-specified bare dissociation constants $K_{d,i}^{(0)}$ for each protein binding site $i$ [36,37]. We first briefly review the RNAcompete method, and then we describe how we derive $K_{d,i}^{(0)}$ from these experimental data.

4.3.1 RNAcompete experiment

RNAcompete is a method developed for rapid characterization of the binding specificities of RBPs. It first generates a pool of various RNA sequences and/or structures. This pool of RNAs then interact with RBPs in the pulldown reaction. In this reaction, the RNA concentration is much greater than that of RBPs and thus different RNA sequences “compete” with each other for interacting with RBPs, hence the name RNAcompete. In this competition, the probability that an RBP successfully binds to an RNA sequence is determined by its relative affinity, compared to all other RNA sequences.

The system is then precipitated and washed. After these processes, only the RBP-bound sequences remain. This population of precipitated RNA sequences is then compared with the original RNA pool in a microarray experiment. In the microarray process the RBP-bound RNAs are labeled with Cy5 and co-hybridized with the Cy3-labeled original RNA pool. The relative recovery of each RNA sequence in the pulldown RNA population as compared to the total pool signal is then measured as a log ratio of signal strength, serving as a measure of binding affinity and therefore sequence preference [36].

4.3.2 Derivation of the sequence-dependent binding affinities

The first published data of RNAcompete experiments focused on nine different RBPs, the footprint of which are $l = 7$nt [36]. Thus, there are a total $4^7 = 16384$ different $K_{d,i}^{(0)}$ that have to be determined by experiments.

The raw scores for the binding affinity of all 16384 7-mers (except the two sequences GCTCTTC and GAAGAGC needed for synthesizing the RNA pool) are in the form of
\[ R_i = \sinh^{-1}(x_i) - \sinh^{-1}(y_i) \]  

(4.6)

in which \( R_i \) is the score for 7-mer \( i \). The quantities \( y_i \) and \( x_i \) are the signal strengths (as measures of the concentrations) of 7-mer \( i \) in the two RNA pools before and after the pulldown process (i.e. the pool of the original and RBP-bound sequences), respectively. The use of inverse hyper sinusoidal function is for biostatistical purposes \[36, 68\], and this representation is asymptotic to the log ratio,

\[ R_i = \log(x_i/y_i), \]  

(4.7)

when \( x_i, y_i \to \infty \), i.e. signals are strong \[68\]. To maximally take advantage of the available experimental data, we consider that in an experiment such ideal conditions are always satisfied. Thus, in the following derivation, \( R_i \) directly represents the log relative signal strength between the two RNA pools.

There are still few parameters to be determined for deriving \( K_{d,i}^{(0)} \) from \( R_i \). The pulldown process is generated by 18.2\( \mu \)g RNAs, while the pool of original RNAs weighs 1.2\( \mu \)g \[36\]. Thus, the score \( R_i \) should be divided by \( \alpha = 18.2/1.2 \) for taking into account the signal enhancement resulting from this weight difference. The \( K_{d,i}^{(0)} \) can thus be calculated from the relation

\[ \frac{R_i}{\alpha} = \frac{c_{\text{RNA compete}}}{K_{d,i}^{(0)}}, \]  

(4.8)

in which \( c_{\text{RNA compete}} \) is the RBP concentration in equilibrium.

The RBP concentration in equilibrium becomes the other issue in the derivation. While the initial RBP concentration is recorded as 20nM \[36\], the total concentration of all RNA molecules is 75-fold larger than that of RBPs (1.5\( \mu \)M versus 20nM). Thus, the competition between the saturated RNA molecules leads to a much smaller RBP concentration in equilibrium. This equilibrium concentration is left as a fitting parameter in the comparison with the experimentally observed \( K_d \). In our derivation we rewrite the concentration as

\[ c_{\text{RNA compete}} = \gamma c_0, \]  

(4.9)

with \( c_0 = 20\text{nM} \) as the default concentration and \( \gamma \leq 1 \) as the fitting parameter.
Consequently, in the calculation of effective dissociation constants, we first derived

$$\frac{\gamma}{K_{d,i}^{(0)}} = \frac{R_i}{\alpha c_0},$$

(4.10)

and substitute $R_i/\alpha c_0$ for $1/K_{d,i}^{(0)}$ in Eq. 4.1. We then essentially calculate $c_{1/2,\gamma} = \gamma / K_{\text{d,eff}}$, which satisfies $P_{\text{bound}}[c_{1/2,\gamma}] = 1/2$. The ratio $\gamma$ is then determined in the comparison with experimentally observed $K_d$’s by regression, and will be applied back to determine the essential bare dissociation constants $K_{d,i}^{(0)}$.

4.4 Interplay between sequence-dependent protein bindings and RNA secondary structures

We focus on the interaction between the protein HuR and several RNA sequences. HuR is an RBP preferring AU-rich motifs [69] and controls multiple other RBPs in a translational network [70]. We fetch the experimentally observed $K_d$ for several sequences, as shown in Table 4.1 [1].

We apply two models, one including and the other excluding RNA secondary structures, for calculating the effective dissociation constants. In the model with RNA secondary structures, the partition function $Z[c]$ in Eq. (4.1) is calculated by a revised Vienna RNA package [40, 50]. In the model without RNA structure, all limited partition functions in Eq. (4.1) are equal to one, and $Z[c]$ can be calculated recursive by iterating the recursive equation,

$$Z_N[c] = Z_{N-1}[c] + Z_{N-l}[c] \frac{c}{K_{d,N-l+1}^{(0)}},$$

(4.11)

in which $l$ is the RBP footprint, and $Z_N[c]$ is the auxiliary partition function for the first $N$-nucleotides portion of the sequence, with respect to concentration $c$.

The theoretical $K_{d,eff}$’s derived from the two models are shown in Table 4.2, as well as the experimentally observed $K_d$ for each sequence. The comparison between the models and experiment is shown in Fig. 4.1. We note that the observed $K_d$’s are measured at $23.5^\circ C$ [1], whereas in RNAcompete experiments the temperature is $37^\circ C$ [36], and thus the $K_{d,i}^{(0)}$ applied to our models are for RNAs at $37^\circ C$. It has been discovered that the
Figure 4.1: Comparison between the experimentally observed $K_d$ and the $K_{d,\text{eff}}$ calculated by two theoretical models, for the HuR binding with respect to several AU-rich sequences listed in Table 4.1. The concentration parameter $\gamma$ is determined by the regression between experimental results and the model with RNA structure, yielding $\gamma = 0.674$ with $r = 0.89$. The model considering RNA structure is more consistent with the experiments.

The binding affinity of HuR can be significantly affected by temperature, up to a 4~10-fold change with respect to a 10°C difference [71]. We thus view Fig. 4.1 mainly as a qualitative comparison, while the quantitative difference is relatively tolerable.

Fig. 4.1 shows intriguing consequences of the competition between HuR-RNA bindings and RNA base pair bond formations. For large binding affinities (right hand side in the figure), both of the two models agree well with experiments, implying that protein binding prevails RNA base pair bond formation, and thus RNA structures are irrelevant. However, for marginal (middle in the figure) and small (left in the figure) binding affinities, the competition between HuR binding and base pair bond formation becomes critical, and the
Table 4.1: The sequences with measured $K_d$ with respect to HuR, from Meisner, et al. [1].

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
</table>
| Cox2     | UAUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUUAAUAUUUAAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUU
$K_{d,eff}$ derived from the model with RNA structures is more consistent with experiments.

This finding is reasonable. When HuR binding affinity is strong, the nucleotides in the footprints are more likely to be bound by HuR, instead of forming secondary structures with other nucleotides. Therefore, for those sequences with small $K_d$ (i.e. strong binding affinity), RNA secondary structures are unfavorable and thus less relevant. When the binding affinity is not strong, HuRs have to compete with other nucleotides to bind with the nucleotides in their footprints. To take into account this competition, RNA secondary structure can no longer be ignored.

The $K_{d,eff}$ can also be compared with the minimum $K_{d,i}^{(0)}$ (i.e. the strongest binding affinity) among each of the sequences, as a comparison between the picture of the single-binding reaction in Eq. 4.2 and that of the binding-chain reactions in Eq. 4.3. We discover that the minimum $K_{d,i}^{(0)}$ in most of the sequences is 4.51nM, which results from the motif UUAUUUA. The three exceptions are Cox-2, IL-$\beta$, and (CUUU)$_4$C, which have $(K_{d,i}^{(0)})_{\text{min}}$ equal to 7.92nM (UAAUUUA), 3.30nM (UUGUUUG), and 7.08nM (UUCUUUC), respectively. All these bare dissociation constants are much larger than the $K_{d,eff}$ calculated by the model without RNA structures, which takes into account the RNAs bound by multiple HuRs. This comparison verifies that the picture of binding-chain reactions shown in Eq. 4.3 is valid, while the single-binding picture in Eq. 4.2 is oversimplified.

However, the observed $K_d$’s in experiments are not always smaller than $(K_{d,i}^{(0)})_{\text{min}}$. For example, the observed $K_d$ of Cox-2 is 13.6nM, much greater than its minimum bare dissociation constant, $(K_{d,i}^{(0)})_{\text{min}} = 7.92nM$. Similar phenomena occur on sequences IL-2 and TNF$\alpha_{\text{mut}}$, of which $K_d > (K_{d,i}^{(0)})_{\text{min}}$. Some sequences, for example IL-4 and TNF$\alpha_{45}$, have $K_d$ slightly smaller than their $(K_{d,i}^{(0)})_{\text{min}}$, which can not be explained by multiple HuR bindings. It is worth to notice that all these sequences have marginal or small binding affinities, i.e. $K_d \sim 1nM$ or $K_d > 1nM$. For these sequences, as we already pointed out, the model with RNA structures generates $K_{d,eff}$ consistent with experimental results, as shown in Table 4.2 and Fig. 4.1. Thus, it is again verified that the competition between HuR bindings and base pair bond formations is important. In the binding-chain reactions, the
effect from RNA secondary structure is crucial, especially for those sequences with marginal or small binding affinities.

4.5 Summary

In this chapter we investigate the sequence specificity of RNA-protein binding. We have constructed a model for the interplay between multiple RBPs on RNA secondary structures. Based on the bare dissociation constants measured by RNAcompete experiments [36,37], we manipulated our model and essentially calculated the effective dissociation constant, $K_{d,eff}$, for the RBP-RNA interactions of the protein HuR and several AU-rich RNA sequences. We compared our calculation with the experimentally observed $K_d$ of these sequences, and received highly consistent results.

We have constructed two models for multiple RBP bindings: one includes RNA secondary structure, and thus both, RNA base pairings and HuR binding energy, contribute to the partition function; the other does not consider RNA structure at all, and HuR binding energy is the only contribution to the partition function. We have discovered that, although the two models are equally consistent with experimental data when $K_{d,eff}$ is small, the model including RNA secondary structure prevails for sequences with marginal or large $K_{d,eff}$.

Consequently, we have extended our model for RNA-protein binding, from the simplified model with only two pre-defined protein binding sites to a complete model including multiple sequence-specified protein binding motifs. We have applied our model to the interaction between HuR and RNA, and have shown that this model can be quantitatively powerful, once the required parameters, the bare dissociation constants $K_{d,i}^{(0)}$, are measured experimentally. As such data can already be supplied by RNAcompete experiments [36,37], we expect that our model can be applied to more complicated phenomena of RNA-RBP interactions. For example, there can be competitions between multiple types of RBPs on the same RNA molecule. It is thus interesting to figure out how these RBPs interact with each other, and what role RNA secondary structures play.
in these competitions.
Chapter 5

CONCLUSION

Triggered by the combinatorial effects in post-transcriptional regulation, in this thesis we have discussed the long-range interplay between the single-stranded RNA binding proteins, mediated by RNA secondary structures. We have proposed a mechanism for this long-range interdependency on the basis of fundamental principles in statistical physics, and have verified its crucial role in RNA-protein interactions.

Post-transcriptional gene regulation enhances the complexity of the flow of genetic information. Because of these regulatory processes, the information flow from mRNAs to proteins is far from merely decryption. Instead, a complicated network of genetic modifications and spatial-temporal controls determines the fate of each RNA molecule and thus its biological function. An important phenomenon in this network is the combinatorial gene regulation, in which multiple RNA-binding proteins (RBPs) and/or microRNAs bind on one RNA molecule to accomplish the regulation. Such a collaboration, or the cooperativity, is intriguing, and thus we investigated this biological phenomenon through an approach of theoretical physics.

Our strategy is based on the fundamental principles in statistical physics. We started from the simplest model including RBP-RBP cooperativity, in which there are two RBP binding sites on an RNA molecule. Working on this model, we proposed a mechanism for the cooperativity, which results from the ensemble of RNA secondary structures. Once an RBP binds on the RNA, some secondary structures are prohibited, and the ensemble of structures is changed, resulting a different binding probability with respect to the other RBP. In this
model, no specific protein binding detail is required, except the trivial assumption that, when an RBP is bound, all nucleotides in its footprint are forced to be unpaired. Therefore, the consequence of our model shall be robust and generic [29].

We investigated the simplified model theoretically, and discovered that this model supports a power-law correlation function between the binding probabilities of the two RBPs. As a power-law function has no characteristic length, this discovery implies a long-range cooperativity. We analytically derived this power-law correlation in a homopolymer RNA molten-phase model, and numerically calculated the same consequence in a heteropolymer RNA glass-phase model [29]. In addition to these simplified models, we manipulated the Vienna RNA package [40], and discovered the same power-law correlation. We thus concluded that this power-law phenomenon is generic, and our mechanism supports the long-range cooperativity in realistic RNA-protein interactions [29].

We have also applied our model to real RNA databases. We fetched RNA sequences of human and C. elegans, and investigated the cooperativity between one RBP on the 5’UTR and the other on the 3’UTR. We derived a characteristic difference of free energy differences, i.e. $\Delta \Delta G$, as a quantitative measure of the structure-mediated cooperativity between these two RBPs. We have discovered that, although these two RBPs are far away from each other on the RNA, they show biologically relevant cooperativity: The $\Delta \Delta G$ is generally greater than $RT$, and sometimes is much greater than this energy scale. Interestingly, the cooperativity gets stronger as the two RBPs get closer to their respective ends on the RNA. More investigations showed that this cooperativity strongly depends on the lengths and sequences of the 5’ and 3’UTRs, while the coding region has little effect [30].

As we are also interested in how strong the cooperativity between two UTRs can possibly become, we generated some artificial sequences by manipulating an evolutionary algorithm, and calculated their $\Delta \Delta G$’s. We succeeded to generate several sequences supporting cooperativities much stronger than those supported by most of the natural RNA sequences. Considering that the protein binding in post-transcriptional regulation is generally a dynamic process, it is reasonable that the natural RNAs support cooperativity with “marginal” strength [30].
We also investigated the cooperativity between human microRNAs and putative RBP binding sites. We discovered that, statistically, there is a negative cooperativity between a 3'UTR-binding human microRNA and a 5'UTR-binding RBP, which can be reasonably described as a consequence of the roles of microRNAs in gene repression \([30]\).

At last, we extended our discussion from the simplest model of only two binding sites to a model taking into account all sequence dependent binding motifs. This model allows us to investigate the interaction between a large number of sequence-dependent protein binding sites. We applied this model to calculate the effective dissociation constants of an RNA-binding protein HuR on several RNA sequences, with the help of the sequence-dependent protein binding data from RNAcompete experiments \([36, 37]\). We compared our calculation with the observed dissociation constants of the sequences, and discovered that RNA secondary structures play an important role in the HuR-RNA interaction, verifying that the structure-mediated cooperativity between RBPs is necessary for describing the observed HuR binding affinities.

Consequently, we proposed a robust and generic mechanism for the long-range cooperativity between RNA binding partners on the basis of statistical physics. In post-transcriptional gene regulation, various proteins and microRNAs bind on RNA molecules through different microscopic mechanism. However, our mechanism is not restricted to any of these microscopic details, and thus it is valid for a wide variety of RNA-protein interactions. Moreover, our mechanism does not require the two RBPs to physically contact each other. Instead, our mechanism describes the interdependency between two RBPs which bind on the same RNA but are far away from each other. Such long-range interdependency is an intriguing phenomenon and is often described by theories with complicated but restricted suppositions. However, our mechanism offers a simple but generic picture, and thus exposes an advantageous approach for this phenomenon.
Bibliography


[65] Altschul, S. F. and Erickson, B. W. “Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon


Appendix A

MOLten-PHASE partition FUNCTION
OF THE MODEL INCLUDING A
CONSTANT LOOP COST

In the molten phase, an RNA molecule can be viewed as a homopolymer, and the partition functions $Z_{(i,j)}$ and $Z_{b(i,j)}$ in Eq. (2.13) retard to functions of sequence length $j - i + 1$. Its partition function has been calculated before [48, 49, 54] but we include it here for the sake of completeness.

Due to the translational invariance, we can rewrite the partition functions as

$$Z_b(j - i + 1) = Z_{b(i,j)}, \text{ and } Z_0(j - i + 1) = Z_{(i,j)}.$$  (A.1)

The recursive relation in Eq. (2.13) is then rewritten as

$$Z_b(N+1) = qZ_b(N-1) + q\xi \left(Z_0(N-1) - Z_b(N-1)\right),$$  (A.2a)

$$Z_0(N+1) = Z_0(N) + \sum_{k=0}^{N-1} Z_0(k)Z_b(N-k+1).$$  (A.2b)

The first of these equations can be solved as

$$Z_0(N) = \frac{1}{q\xi}Z_b(N + 2) - \frac{1 - \xi}{\xi}Z_b(N),$$  (A.3)

and can then be used to replace all $Z_0$’s in the second equation, rendering a purely $Z_b$-dependent recursion. Applying $z$-transformation to this recursion with the initial conditions...
Z_b(0) = 0, Z_b(1) = 0, Z_b(2) = q\xi, \quad (A.4)
yields the equation for the z-space partition function \( \hat{Z}_b(z) \equiv \sum_{N=0}^{\infty} Z_b(N) z^{-N} \)
\[
(\frac{z^2}{q\xi} - \frac{1-\xi}{\xi}) \hat{Z}_b(z)^2 - \frac{1}{z} \left( \frac{z^2}{q\xi} - \frac{1-\xi}{\xi} \right) (z - 1) \hat{Z}_b(z) + 1 = 0. \quad (A.5)
\]
The limited partition function in z-space is thus given as
\[
\hat{Z}_b(z) = \frac{z-1}{2z} - \frac{1}{2} \sqrt{(z-1)^2 - \frac{4q\xi}{z^2 - q(1-\xi)}}, \quad (A.6)
\]
The z-space partition function \( \hat{Z}_0(z) \) can be derived by applying z-transformation to Eq. (A.2a), yielding
\[
\hat{Z}_0(z) = \left( \frac{z^2}{q\xi} - \frac{1-\xi}{\xi} \right) \hat{Z}_b(z), \quad (A.7)
\]
which leads to \( \hat{Z}_0(z) \) by substituting Eq. (A.6) for \( \hat{Z}_b(z) \), yielding
\[
\hat{Z}_0(z) = \frac{(z-1)(z^2 - q + q\xi)}{2q\xi z} - \frac{\sqrt{(z^2 - q + q\xi) [(z-1)^2(z^2 - q + q\xi) - 4q\xi z^2]}}{2q\xi z}. \quad (A.8)
\]
The partition function in real space is then derived by the inverse z-transformation,
\[
Z_0(N) = \frac{1}{2\pi i} \oint_C dz z^{N-1} \hat{Z}_0(z). \quad (A.9)
\]
As \( N \to \infty \), the leading order term of this contour integral, contributed by the topological defect with the largest real part of \( z \), determines its value. According to Eq. (A.8), \( \hat{Z}_0(z) \) has a simple pole at \( z_p = 0 \), and branch cuts at all \( z \) satisfying
\[
[z^2 - q + q\xi] [(z-1)^2(z^2 - q + q\xi) - 4q\xi z^2] \leq 0. \quad (A.10)
\]
The first component in Eq. (A.10) leads to two branch cut end points \( z_{\pm} = \pm \sqrt{q(1-\xi)} \).
Substituting these two values into the function of the second component,
\[
f(z) = (z-1)^2(z^2 - q + q\xi) - 4q\xi z^2 \quad (A.11)
\]
it can be found that \( f(z_{\pm}) < 0 \). Since \( f(\infty) > 0 \) is known, at least one of the branch cuts must extend along the real axis beyond \( z = z_+ \). Following the derivation in the appendix of Ref. [49], the large-\( N \) expression of \( G(N) \) is then obtained as

82
\[ Z_0(N) = A(q, \xi) N^{-3/2} z_c^N (q, \xi) [1 + O(N^{-1})], \quad (A.12) \]

where \( z_c \) is the solution of \( f(z_c) = 0 \) with the greatest real part, and the prefactor \( A(q, \xi) \) is given by

\[ A = \frac{\sqrt{z_c^2 z_c^2 - q + q\xi} f'(z_c)}{2\pi z_c q\xi} \Gamma \left( \frac{3}{2} \right). \quad (A.13) \]

To simplify the notation, we omit the subscript \( c \) in this manuscript and write down the asymptotic partition function, for example in Eq. (2.14), using \( z \) rather than \( z_c \).
Appendix B
Calculation of the molten-phase correlation function for the model including a constant loop cost to first order in loop cost

In this appendix, we will calculate the correlation function $g(D)$ between two protein binding sites at a distance $D$ from each other. To calculate the correlation function $g(D)$, we need to know the changed components $S_d$, $S^{\ast}_{dd}$, and $S^{(11)}_{dd}$ in the limited partition functions $Z_d$ and $Z_{dd}$. Since these are difficult to obtain exactly, we will only calculate their expansions in $(1 - \xi)$, thus taking the effects of a finite loop cost into account perturbatively. In addition, to exclude all boundary and finite-size effects, we consider the limit of an infinitely long molecule in which both footprints are far from the ends of the RNA, i.e., the limit of $N \gtrsim n_1 \approx n_2 \gg D \gg l \geq 1$. Here, we will aim to calculate the overall correlation function $g(D)$ to first order in $(1 - \xi)$ which will show its power law dependence on the distance $D$ between the binding sites. Appendix C will then demonstrate that this power law dependence is not changed by the second order term, which is a lot more difficult to obtain.

The limited partition functions $S_d$ and $S^{\ast}_{dd}$ appear in Eqs. (2.15) and (2.19) with prefactors of $(1 - \xi)$ while $S^{(11)}_{dd}$ occurs with a prefactor of $(1 - \xi)^2$. Thus, an expansion of the correlation function $g(D)$ to first order in $(1 - \xi)$ requires knowledge of $S_d$ and $S^{\ast}_{dd}$ to zeroth order in $(1 - \xi)$ but does not depend on $S^{(11)}_{dd}$. We will thus start by calculating the
Figure B.1: A one-to-one mapping from the changed structures to the structures including the certain bond. The footprint is between the \( n^{th} \) and the \((n+1)^{th}\). Red dots are the nucleotides forming the base pair stack which includes the footprint. The two types of structures are taken into account in Eq. (B.1), where the first summation is for the structures in (a) and the second one is for (b). All labels for the segment lengths in the plot do not count the red dots.

limited partition functions \( S_d \) and \( S_{dd}^* \) to zeroth order in \((1 - \xi)\).

**B.1 \( S_d \), the partition function for the changed structures in \( Z_d \)**

The partition function for the changed structures in \( Z_d(n,m) \), expressed as \( S_d(n,m) \), includes the structures shown in Fig. 2.3, i.e., those structures in \( Z_0(n+m) \) in which the \( n^{th} \) and the \((n+1)^{st}\) base are involved in a base stack of a stem. To calculate the contribution of all these structures we note that in all structures in which the \( n^{th} \) and the \((n+1)^{st}\) base are involved in a base stack the two consecutive base pairs can be contracted to a single base pair yielding a new structure on an RNA of \( n+m-2 \) bases, in which the stem is shortened by one stack but the topology of the structure is otherwise unchanged. This process is described in Fig. B.1. Thus, qualitatively \( S_d(n,m) \) is given as \( q \) times (to represent the one additional base pair) the partition function of all structures of an \( n+m-2 \) base RNA in which the \( n^{th} \) base is paired; this in turn is the partition function over all structures of an \( n+m-2 \) base RNA minus the partition function over all structures of an \( n+m-2 \) base RNA in which the \( n^{th} \) base is unpaired and the latter is in turn the partition function over all structures of an \( n+m-3 \) base RNA, i.e., we would expect
However, there are some subtleties missed in the above qualitative argument that require a more careful computation. To this end, we write \( S_d(n,m) \) explicitly as a sum over the position of the other side of the stack that the \( n^{th} \) and \( (n+1)^{st} \) base are included in and separate this sum into terms where the other side of the stack is to the left of \( n \) and terms where the other side of the stack is to the right of \( n \) as shown in Fig. B.1. This yields

\[
S_d(n,m) = q \sum_{k=2}^{n-1} Z_b(k) \tilde{Z}(n-k-1;m-1) + q \sum_{k=2}^{m-1} Z_b(k) \tilde{Z}(n-1;m-k-1),
\]

(B.1)

where \( \tilde{Z}(k_1;k_2) \) is for the outer part in Fig. B.1, taking into account all structures formed by the nucleotides in the two non-independent segments of lengths \( k_1 \) and \( k_2 \). Notice that \( \tilde{Z} \) is symmetric with respect to its two segments, i.e. \( \tilde{Z}(k_1;k_2) = \tilde{Z}(k_2;k_1) \).

For zero loop cost (\( \xi = 1 \)) we would simply have \( \tilde{Z}(k_1;k_2) = Z_0(k_1 + k_2) \). Since our goal is here to calculate only the zeroth-order term of \( S_d \), we can thus use

\[
\tilde{Z}(k_1;k_2) = Z_0(k_1 + k_2) + O((1 - \xi)).
\]

(B.2)

Similarly, Eq. (A.2a) yields

\[
Z_b(k) = q\xi Z_0(k-2) + q(1 - \xi) Z_b(k-2)
\]

\[
= q\xi Z_0(k-2) + O((1 - \xi)).
\]

(B.3)

Substituting the zeroth-order approximated \( \tilde{Z} \) and \( Z_b \) into Eq. (B.1), \( S_d \) can be approximated to the zeroth order in \( (1 - \xi) \) as
The summation term is exactly equivalent to a conditional partition function of an RNA molecule with \( n + m - 2 \) nucleotides which takes into account all structures in which the 1\(^{st} \) nucleotide is paired with another arbitrary nucleotide. This conditional partition function can be calculated by subtracting the partition function for the structures including an unpaired 1\(^{st} \) nucleotide from the whole partition function, leading to

\[
\sum_{k=2}^{n+m-2} Z_0(k) Z_0(n + m - 2 - k) = Z_0(n + m - 2) - Z_0(n + m - 3). \tag{B.5}
\]

Thus, the partition function for the changed structures, \( S_d \), is derived to the zeroth order in \( (1 - \xi) \) as

\[
S_d(n, m) = q [Z_0(n + m - 2) - Z_0(n + m - 3)] - q^2 \xi Z_0(n-2) Z_0(m-2) + O((1 - \xi)), \tag{B.6}
\]

which is nearly our naive expectation. In fact, in the relevant limit of \( n \approx m \approx N/2 \to \infty \), we can insert the asymptotic form Eq. (2.14) of \( Z_0 \) to obtain

\[
S_d(n, m) = q A \left[ \frac{z^{n+m-2}}{(n + m - 2)^{3/2}} - \frac{z^{n+m-3}}{(n + m - 3)^{3/2}} \right] - q^2 \xi A^2 \frac{z^{n+m-4}}{(n-2)^{3/2}(m-2)^{3/2}} + O \left( (1 - \xi), \frac{z^{n+m}}{N^{5/2}} \right) \tag{B.7}
\]

and find that the additional term on the second line is actually decaying with a power of
Figure B.2: The structures included in $S_{dd}^*$. The outer segment lengths $n_1 - 1$ and $n_2 - 1$ do not include the two nucleotides forming the outer bond. The partition function for the inner segment is given by $Z_b(D)$ and thus includes an algebraic component $D^{-3/2}$ for $D \gg 1$.

$N^{-3}$ compared to the power of $N^{-3/2}$ of the terms on the first line, which represent our naive expectation, and thus can be neglected (the $z^{n+m}$ behavior in the numerator is the same for all terms). This finally yields

$$S_d(n, m) = q Az^{n+m} \frac{z - 1}{z^3} N^{-3/2} + O \left( (1 - \xi), \frac{z^{n+m}}{N^{5/2}} \right) \quad \text{(B.8)}$$

### B.2 $S_{dd}^*$, contributions to $Z_{dd}$ when both footprints are in the same base stack

The structures included in $S_{dd}^*$, in which both footprints are in the same stack of a stem, are shown in Fig. B.2. Upon contracting the two stacked base pairs into one, these configurations exactly correspond to the configurations of an RNA of $n_1 + D + n_2 - 2$ bases in which base $n_1$ and base $n_1 + D - 2$ are paired. Thus, this partition function encodes the pairing probability for two bases of an RNA with distance $D - 2$ which is known to depend like a power law on the distance $D$.

Specifically, the partition function for the structures included in this quantity can be written as

$$S_{dd}^*(n_1, D, n_2) = q Z_b(D) \tilde{Z}(n_1 - 1; n_2 - 1). \quad \text{(B.9)}$$

Since we only need to know the zeroth order term of $S_{dd}^*$ in $(1 - \xi)$ we can substitute the zeroth order expansions Eqs. (B.2) and (B.3) of $\tilde{Z}$ and $Z_b$ respectively and obtain

$$S_{dd}^* = q^2 \xi Z_0(D - 2) Z_0(n_1 + n_2 - 2) + O((1 - \xi)). \quad \text{(B.10)}$$
Inserting the asymptotic form Eq. (2.14) for $Z_0$ finally yields

$$S_{dd}^*(n_1, D, n_2) = \frac{q^2 \xi}{z^4} A^2 \frac{2^{n_1+n_2+D}}{D^{3/2} N^{3/2}} + O \left( (1 - \xi), \frac{z^{n_1+D+n_2}}{N^{5/2}}, \frac{z^{n_1+D+n_2}}{D^{5/2}} \right). \quad (B.11)$$

This term explicitly contains the power law dependence on the distance $D$ between the protein binding sites.

### B.3 Correlation function

The molten-phase protein-protein correlation function, $g(D)$, is given by the first equality in Eq. (2.20) in terms of the limited partition functions $Z_d$ and $Z_{dd}$, and the protein-binding parameters $c_i$ and $K_i$ with $i = 1, 2$. As we have described at the beginning of this appendix, to exclude all boundary and finite-size effects, we consider the limit of an infinitely long molecule in which both footprints are far from the ends of the RNA, i.e., the limit of $N \gg n_1 \approx n_2 \gg D \gg l \geq 1$.

As a first step to calculating $g(D)$ we divide the numerator and the denominator of the first expression in Eq. (2.20) by $Z_0^2$ and get

$$g(D) = \frac{Z_0 Z_{dd} - Z_d Z_d}{Z_0^2} \frac{Z_d}{Z_0} = \frac{Z_{dd} - Z_d}{Z_0} \times \frac{Z_d}{Z_0} \left( 1 + \frac{c_1}{K_{d,1}} \frac{Z_d}{Z_0} + \frac{c_2}{K_{d,2}} \frac{Z_d}{Z_0} + \frac{c_1 c_2}{K_{d,1} K_{d,2}} \frac{Z_{dd}}{Z_0} \right) \equiv \frac{N}{D^2}. \quad (B.12)$$

Thus, the relevant quantities are the ratios $Z_d/Z_0$ and $Z_{dd}/Z_0$, which are obtained by dividing the results from sections B.1 and B.2 by the asymptotic form $Z_0(N) \approx A z^N/N^{3/2}$. It is easy to see that this division eliminates the exponential dependence on $z$ and cancels the $N^{-3/2}$ dependence of the results from sections B.1 and B.2.

We will first calculate the numerator and denominator of $g(D)$ separately to appropriate orders in $(1 - \xi)$, and then merge them together to derive $g(D)$ to the first order in $(1 - \xi)$. We are going to show that $g(D)$ decays algebraically as $D \gg 1$ and thus supports a long-range interaction between binding proteins. Appendix C will then demonstrate that this algebraic
behavior also holds up in the second order in \((1 - \xi)\).

The numerator \(N\) of \(g(D)\) is the difference between two products of two partition functions. Without loop cost, this difference, as shown in Eq. (2.12), leads to a residue in \(O(1/N^2)\) and thus converges to zero as \(N \to \infty\). In the model including a constant loop cost, however, the residue is \(O(1/N^0)\) and does not diminish in the large-\(N\) limit. The goal in this section is to calculate this finite residue, and confirm that this residue is a power-law function of \(D\) and therefore supports a long-range effect in the system.

With the help of the relations in Eqs. (2.15) and (2.19), the numerator in Eq. (B.12) is written as

\[
N = \frac{Z_{dd}(n_1, D, n_2)}{Z_0(N)} - \frac{Z_d(n_1, D + l + n_2)}{Z_0(N)} \frac{Z_d(n_1 + l + D, n_2)}{Z_0(N)}
\]

\[
= \left[ \frac{Z_0(N - 2l)}{Z_0(N)} - \left( \frac{Z_0(N - l)}{Z_0(N)} \right)^2 \right]
\]

\[
- (1 - \xi) \left[ \frac{S_d(n_1, D + n_2)}{Z_0(N)} - \frac{Z_0(N - l)}{Z_0(N)} \frac{S_d(n_1, D + l + n_2)}{Z_0(N)} \right]
\]

\[
- (1 - \xi) \left[ \frac{S_d(n_1 + D, n_2)}{Z_0(N)} - \frac{Z_0(N - l)}{Z_0(N)} \frac{S_d(n_1 + D + l, n_2)}{Z_0(N)} \right]
\]

\[
+ (1 - \xi)^2 \left[ \frac{S_{dd}^{(11)}}{Z_0(N)} - \frac{S_d(n_1, D + l + n_2)}{Z_0(N)} \frac{S_d(n_1 + l + D, n_2)}{Z_0(N)} \right]
\]

\[
+ \xi(1 - \xi) \frac{S_{dd}^*}{Z_0(N)}
\]

The first term is the same as for the model in the absence of a loop cost and thus leads to a residue proportional to \((l/N)^2\) which vanishes in the limit \(N \to \infty\). The forth term is of second order in \((1 - \xi)\) and can thus be ignored here. To calculate the second term, we substitute the asymptotic expressions Eqs. (2.14) and (B.8) for \(Z_0\) and \(S_d\), respectively, and obtain for the term in brackets

\[
\frac{S_d(n_1, D + n_2)}{Z_0(N)} - \frac{Z_0(N - l)}{Z_0(N)} \frac{S_d(n_1, D + l + n_2)}{Z_0(N)}
\]

\[
= q \frac{z - 1}{z^{3+2l}} \left[ \frac{N^{3/2}}{(N - 2l)^{3/2}} - \frac{N^3}{(N - l)^3} \right] + O(N^{-1})
\]

which also vanishes for \(N \to \infty\). Due to symmetry the same argument applies to the third term in Eq. (B.13).
Finally, we can substitute the asymptotic expansions Eqs. (2.14) and (B.11) of $Z_0$ and $S_{dd}^*$, respectively, into the last term to find the entire numerator as

$$
N = (1 - \xi) \frac{q^2 \xi^2 A}{z^{4+2l}} D^{-3/2} + O((1 - \xi)^2, D^{-5/2}, N^{-1}). \quad (B.15)
$$

Considering that this numerator of $g(D)$ has an explicit prefactor of $(1 - \xi)$, it is enough to calculate the denominator to zeroth order in $(1 - \xi)$. Including all the arguments of the limited partition functions, the denominator is given by (before taking the square)

$$
D = 1 + \frac{c_1}{K_{d,1}^{(0)}} \frac{Z_d(n_1, D+l+n_2)}{Z_0(N)} + \frac{c_2}{K_{d,2}^{(0)}} \frac{Z_d(n_1+l+D, n_2)}{Z_0(N)} + \frac{c_1 c_2}{K_{d,1}^{(0)} K_{d,2}^{(0)}} \frac{Z_{dd}(n_1, D, n_2)}{Z_0(N)}, \quad (B.16)
$$

Eqs. (2.15) and (2.19) explicitly show that in zeroth order in $(1 - \xi)$ $Z_d$ and $Z_{dd}$ can be replaced by $Z_0$ thus yielding

$$
D = 1 + \frac{c_1}{K_{d,1}^{(0)}} \frac{Z_0(N-l)}{Z_0(N)} + \frac{c_2}{K_{d,2}^{(0)}} \frac{Z_0(N-l)}{Z_0(N)} + \frac{c_1 c_2}{K_{d,1}^{(0)} K_{d,2}^{(0)}} \frac{Z_0(N-2l)}{Z_0(N)} + O((1 - \xi))
= 1 + \frac{c_1}{K_{d,1}^{(0)}} \frac{N^{3/2}}{(N-l)^{3/2} z^l} + \frac{c_2}{K_{d,2}^{(0)}} \frac{N^{3/2}}{(N-l)^{3/2} z^l} + \frac{c_1 c_2}{K_{d,1}^{(0)} K_{d,2}^{(0)}} \frac{N^{3/2}}{(N-2l)^{3/2} z^{2l}} + O((1 - \xi), N^{-1})
= \left(1 + \frac{c_1}{K_{d,1}^{(0)} z^l}\right) \left(1 + \frac{c_2}{K_{d,2}^{(0)} z^l}\right) + O((1 - \xi), N^{-1}). \quad (B.17)
$$

where we have again used the asymptotic expression Eq. (2.14) for $Z_0$.

The correlation function is then obtained by dividing Eq. (B.15) by the square of Eq. (B.17) and is thus given by

$$
g(D) = (1 - \xi) \frac{q^2 \xi^2 A}{z^{4+2l}} \left(1 + \frac{c_1}{K_{d,1}^{(0)} z^l}\right) \left(1 + \frac{c_2}{K_{d,2}^{(0)} z^l}\right) D^{-3/2} + O((1 - \xi), N^{-1}, D^{-5/2}). \quad (B.18)
$$

Consequently, once a finite loop cost is added into the RNA-folding model, a long-range correlation occurs between two binding partners on the RNA.
Appendix C

Calculation of the molten-phase correlation function for the model including a constant loop cost to second order in loop cost

In this appendix, we will calculate the second order terms in the expansion of the correlation function \( g(D) \) in \((1 - \xi)\), i.e., in the loop energy. Since the calculation of this term is quite involved, it is important to point out that the main result, the power law behavior of the correlation function \( g(D) \), already occurs in the first order in \((1 - \xi)\) as detailed in Appendix B. We do include the calculation here nevertheless for two reasons. First, it quantitatively improves the pre-factor of the power law when compared to numerical results at finite \((1 - \xi)\). Second, the fact that the second order term has the same power law dependence on the distance \( D \) as the first order term strengthens the argument that this power law behavior is not simply an artifact of the perturbative calculation. However, the reader content with only the expansion of the correlation function to first order in \((1 - \xi)\) may opt to skip this appendix.

C.1 \( S_d \), the partition function for the changed structures in \( Z_d \)

Calculating the correlation function \( g(D) \) to second order requires expanding the limited partition function \( S_d \) to first order in \((1 - \xi)\). Our starting point for this calculation will
be Eq. (B.1). To make progress, we need to know the partition function \( \tilde{Z} \) to first order in \((1 - \xi)\).

To find the expansion of \( \tilde{Z}(k_1; k_2) \), again two groups of secondary structures have to be distinguished in \( \tilde{Z}(k_1; k_2) \). Similar to the idea of calculating \( S_d(n, m) \), one group of structures contributes the same in both \( Z_0(k_1 + k_2) \) and \( \tilde{Z}(k_1; k_2) \), and the other contributes differently. The latter group includes two types of structures, whose differences in contribution between \( Z_0(k_1 + k_2) \) and \( \tilde{Z}(k_1; k_2) \) are shown in Fig. C.1. Thus, \( \tilde{Z}(k_1; k_2) \) can be expressed as \( Z_0(k_1 + k_2) \) plus a changed term resulting from the loop cost as

\[
\tilde{Z}(k_1; k_2) = Z_0(k_1 + k_2) + (1 - \xi) \left[ q \tilde{Z}(k_1 - 1; k_2 - 1) - S_d(k_1, k_2) \right], \quad (C.1)
\]

where the term including \( \tilde{Z}(k_1 - 1; k_2 - 1) \) contains the configurations in Fig. C.1(a) and the following term is for those in Fig. C.1(b). Substituting Eq. (C.1) into Eq. (B.1) rewrites \( S_d \) as

\[
S_d(n, m) = q \left( \sum_{k=2}^{n-1} + \sum_{k=2}^{m-1} \right) Z_b(k) Z_0(n + m - k - 2) \quad (C.2a)
\]

\[+ q^2 (1 - \xi) \sum_{k=2}^{n-2} Z_b(k) \tilde{Z}(n - k - 2; m - 2) \quad (C.2b)\]

\[+ q^2 (1 - \xi) \sum_{k=2}^{m-2} Z_b(k) \tilde{Z}(n - 2; m - k - 2) \quad (C.2c)\]

\[- q(1 - \xi) \sum_{k=2}^{n-2} Z_b(k) S_d(n - k - 1, m - 1) \quad (C.2d)\]

\[- q(1 - \xi) \sum_{k=2}^{m-2} Z_b(k) S_d(m - k - 1, n - 1). \quad (C.2e)\]

We now calculate each term in Eq. (C.2) to the first order in \((1 - \xi)\). Note, that all summations in these terms are multiplied by \((1 - \xi)\), except the first term (C.2a). Therefore, our calculation will be to the first order for the summation in (C.2a), and to the zeroth order for the remaining ones, i.e., terms (C.2b)-(C.2e).
We first notice that the combination of terms (C.2b) and (C.2c) has the same form as the expression of $S_d$ in Eq. (B.1), and thus can be written as $q(1 - \xi)S_d(n - 1, m - 1)$. Substituting the zeroth-order $S_d$ from Eq. (B.6) leads to the first-order expression for the combination of the two terms,

$$q^2(1 - \xi) \sum_{k=2}^{n-2} Z_b(k) \tilde{Z}(n - k - 2; m - 2) + q^2(1 - \xi) \sum_{k=2}^{m-2} Z_b(k) \tilde{Z}(n - 2; m - k - 2)$$

$$= q^2(1 - \xi)[Z_0(n + m - 4) - Z_0(n + m - 5)] - q^3\xi(1 - \xi)Z_0(n - 3)Z_0(m - 3) + O((1 - \xi)^2).$$

(C.3)

Next, we calculate the first one among the remaining three terms, i.e., term (C.2a). To evaluate this term to the first order in $(1 - \xi)$, it is necessary to first figure out the partition function $Z_b$ to first order. Iterating Eq. (B.3) once leads to the approximation,

$$Z_b(D) = q\xi Z_0(D - 2) + q^2\xi(1 - \xi)Z_0(D - 4) + O((1 - \xi)^2).$$

(C.4)

Substituting this approximation into the second summation in term (C.2a) yields the first order approximation.
\[ q \sum_{k=2}^{m-1} Z_b(k)Z_0(n + m - k - 2) \]
\[ = q^2 \xi \sum_{k=2}^{m-1} Z_0(k - 2)Z_0(n + m - k - 2) \]
\[ + q^3 \xi(1 - \xi) \sum_{k=4}^{m-1} Z_0(k - 4)Z_0(n + m - k - 2) \]
\[ + O((1 - \xi)^2). \]  

(C.5)

We first calculate the first term in Eq. (C.5). Applying the changing variable strategy similar to that in Eq. (B.4), this term becomes

\[ q^2 \xi \sum_{k=2}^{m-1} Z_0(k - 2)Z_0(n + m - k - 2) \]
\[ = q^2 \xi \sum_{k'=n+1}^{n+m-2} Z_0(n + m - k' - 2)Z_0(k' - 2) \]
\[ = q \sum_{k=n+1}^{n+m-2} Z_b(k)Z_0(n + m - k - 2) - q^3 \xi(1 - \xi) \sum_{k=n+1}^{n+m-2} Z_0(k - 4)Z_0(n+m-k-2) + O((1 - \xi)^2) \]

(C.6)

Substituting Eq. (C.6) into Eq. (C.5) results in a subtraction between two first-order summations and leads to

\[ q \sum_{k=2}^{m-1} Z_b(k)Z_0(n + m - k - 2) \]
\[ = q \sum_{k=n+1}^{n+m-2} Z_b(k)Z_0(n + m - k - 2) \]
\[ + q^3 \xi(1 - \xi) \left( \sum_{k=4}^{m-1} \sum_{k=1}^{m+1} - \sum_{k=4}^{m+1} \right) Z_0(k - 4)Z_0(n+m-k-2) + O((1 - \xi)^2) \]

(C.7)

Finally, substituting Eq. (C.7) into term (C.2a) yields
With the help of the equality in Eq. (B.5), Eq. (C.8) can be further simplified to

\[ g \left( \sum_{k=2}^{m-1} \sum_{k=2}^{m-2} Z(k)Z(n+4m-2-k) \right) + \sum_{k=2}^{m-1} Z(k)Z(n+m-k-2) \]

\[ + \sum_{k=2}^{m-1} Z(k)Z(n-k-1) \]

\[ + \sum_{k=2}^{m-1} Z(k)Z(n-k-2) \]

\[ + \sum_{k=2}^{m-1} Z(k)Z(n-k-3) \]

\[ + \sum_{k=2}^{m-1} Z(k)Z(n-k-4) \]

\[ + \sum_{k=2}^{m-1} Z(k)Z(n-k-5) \]

\[ + O(1 - \xi^2). \]

The first two terms in Eq. (C.10) are both in the form of the zeroth-order \( S_0 \) in Eq. (B.4) (considering \( Z_0(k) = Z_0(k) + O(1 - \xi^2) \)) and thus their combination can be rewritten as

\[ (1 - \xi)^2 \]

There are now two last terms, terms (C.2a) and (C.2b), remaining to be calculated. To this end, we substitute the zeroth-order \( S_0 \) in Eq. (B.6), yielding

\[ (1 - \xi)^2 \]

\[ (1 - \xi)^2 \]

\[ (1 - \xi)^2 \]

\[ (1 - \xi)^2 \]
\[-q(1 - \xi)S_d(n-1, m-1) + q(1 - \xi)S_d(n-2, m-1) + q^2(1 - \xi)Z_0(n-2)Z_0(m-3) + O((1 - \xi)^2)\]
\[= -q^2(1 - \xi)[Z_0(n + m - 4) - 2Z_0(n + m - 5) + Z_0(n + m - 6)]\]
\[+ q^3\xi(1 - \xi)Z_0(n - 3)Z_0(m - 3) + O((1 - \xi)^2)\]

(C.11)

The last two terms can be simplified by rewriting the summations with the help of the relation in Eq. (B.5), which yields
\[q^3\xi(1 - \xi)[Z_0(n - 3) - Z_0(n - 4)]Z_0(m - 3) + q^3\xi(1 - \xi)[Z_0(m - 3) - Z_0(m - 4)]Z_0(n - 3)\]
\[= q^3\xi(1 - \xi)[2Z_0(n - 3)Z_0(m - 3) + Z_0(n - 3)Z_0(m - 4) + Z_0(n - 4)Z_0(m - 3)].\]

(C.12)

Collecting Eqs. (C.11) and (C.12), we express the combination of terms (C.2d) and (C.2e) to the first order in \((1 - \xi)\) as
\[- q(1 - \xi)\sum_{k=2}^{n-2} Z_0(k)S_d(n - k - 1, m - 1) - q(1 - \xi)\sum_{k=2}^{m-2} Z_0(k)S_d(m - k - 1, n - 1)\]
\[= -q^2(1 - \xi)[Z_0(n + m - 4) - 2Z_0(n + m - 5) + Z_0(n + m - 6)]\]
\[+ q^3\xi(1 - \xi)[3Z_0(n - 3)Z_0(m - 3) + Z_0(n - 3)Z_0(m - 4) + Z_0(n - 4)Z_0(m - 3)]\]
\[+ O((1 - \xi)^2)\]

(C.13)

All five terms in Eq. (C.2) are now calculated to the first order in \((1 - \xi)\). We then collect these terms in Eqs. (C.3), (C.9), and (C.13) and discover the first order expression of \(S_d(n, m)\) as
\[S_d(n, m) = q[Z_0(n + m - 2) - Z_0(n + m - 3)] - q^2\xi Z_0(n - 2)Z_0(m - 2)\]
\[+ q^2(1 - \xi)[Z_0(n + m - 5) - Z_0(n + m - 6)]\]
\[+ q^3\xi(1 - \xi)[Z_0(n - 4)Z_0(m - 2) + Z_0(n - 2)Z_0(m - 4) - Z_0(n - 3)Z_0(m - 3)]\]
\[+ q^3\xi(1 - \xi)[Z_0(n - 3)Z_0(m - 4) + Z_0(n - 4)Z_0(m - 3)]\]
\[+ O((1 - \xi)^2),\]

(C.14)

where the first line is the zeroth-order term, exactly identical to the one derived in Eq. (B.6),
and all the subsequent terms are the first order term in \((1 - \xi)\).

In the limit \(N/2 \approx n \approx m \gg 1\), the asymptotic form of Eq. \((C.14)\) is given by inserting \(Z_0(N) \approx Az^N N^{-3/2}\). As before, all terms in which \(n\) and \(m\) are not arguments of the same \(Z_0\) decay as \(N^{-3}\) and can thus be neglected with respect to the \(N^{-3/2}\) dependence of the terms where \(n + m\) is the argument of one \(Z_0\). The asymptotic \(S_d\) is thus given by the remaining relevant terms as

\[
S_d(n, m) = qA z^{n+m} \frac{z - 1}{z^3} \frac{1}{N^{3/2}} + q^2 A (1 - \xi) z^{n+m} \frac{z - 1}{z^6} \frac{1}{N^{3/2}} + O \left( (1 - \xi)^2, \frac{z^{n+m}}{N^{5/2}} \right), \quad (C.15)
\]

### C.2 \(S^*_{dd}\), contributions to \(Z_{dd}\) when both footprints are in the same base stack

The starting point for the first order terms of the limited partition function \(S^*_{dd}\) is Eq. \((B.9)\). We obtain the expansion to first order in \((1 - \xi)\) by substituting the first order expansions of \(Z_b\) and \(\tilde{Z}\). The first order expansion of \(Z_b\) has already been given in Eq. \((C.4)\) and we obtain the first order expression of \(\tilde{Z}\) by substituting the zeroth-order \(S_d\) from Eq. \((B.6)\) into Eq. \((C.1)\) yielding

\[
\tilde{Z}(k_1; k_2) = Z_0(k_1 + k_2) + (1 - \xi) [q Z_0(k_1 + k_2 - 3) + q^2 \xi Z_0(k_1 - 2) Z_0(k_2 - 2)] + O((1 - \xi)^2). \quad (C.16)
\]

This substitution reveals the first order expansion of the limited partition function \(S^*_{dd}\) as

\[
S^*_{dd} = q^2 \xi Z_0(D - 2) Z_0(n_1 + n_2 - 2) + q^3 \xi (1 - \xi) Z_0(D - 4) Z_0(n_1 + n_2 - 2) + q^3 \xi (1 - \xi) Z_0(D - 2) Z_0(n_1 + n_2 - 5) + q^4 \xi^2 (1 - \xi) Z_0(D - 2) Z_0(n_1 - 3) Z_0(n_2 - 3) + O((1 - \xi)^2). \quad (C.17)
\]

In the limit \(n_1 \approx n_2 \approx N/2 \to \infty\), the last term can be dropped since it is of higher order in \(1/N\) than the others. Inserting the asymptotic form Eq. \((2.14)\) for \(Z_0\) finally yields

\[
S^*_{dd}(n_1, D, n_2) = \left( \frac{q^2 \xi}{z^3} + (1 - \xi) \frac{q^3 \xi (z - 1)}{z^6} \right) A^2 z^{n_1+n_2+D} \frac{D^{3/2} N^{3/2}}{D^{3/2} N^{3/2}} + O((1 - \xi)^2, z^{n_1+D+n_2 N^{-5/2}}, z^{n_1+D+n_2 D^{-5/2}}). \quad (C.18)
\]
C.3 $S_{dd}^{(11)}$, contributions to $Z_{dd}$ when both footprints are in stacks

The limited partition function $S_{dd}^{(11)}$ over all configurations in which both footprints are inserted into base stacks is multiplied in the expression for $Z_{dd}$ by $(1 - \xi)^2$. Therefore, it was not relevant when calculating the correlation function $g(D)$ to first order in $(1 - \xi)$ but needs to be considered to zeroth order in $(1 - \xi)$, now that we aim for the second order term of the correlation function $g(D)$.

Qualitatively, $S_{dd}^{(11)}$ can be estimated by a strategy similar to the one that yielded the naive expectation for $S_d$. That is, $S_{dd}^{(11)}$ can be roughly given as $q^2$ times the partition function for all structures of an $n_1 + D + n_2 - 4$ base RNA in which the $n_1^{th}$ and $(n_1 + D)^{th}$ bases are required to be paired (albeit not necessarily with each other), as described in Fig. C.2 for two examples. The latter structures can in turn be calculated by starting from the partition function over all structures of an $n_1 + D + n_2 - 4$ base RNA, subtracting all those in which the $n_1^{th}$ or the $(n_1 + D)^{th}$ base are unpaired and adding back the structures that were subtracted twice because both bases are unpaired. Thus, we would expect

$$S_{dd}^{(11)} \approx q^2[Z_0(n_1 + D + n_2 - 4) - 2Z_0(n_1 + D + n_2 - 5) + Z_0(n_1 + D + n_2 - 6)].$$ (C.19)

However, this naive estimation again has shortcomings. It is not always exactly the $n_1^{th}$ and $(n_1 + D)^{th}$ base which are paired with other bases and not even their distance is always exactly $D$ bases. Based on the configurations of the original structure with two inserted base pair stacks, the exact position of the two paired bases (in the structures where the inserted stacks are contracted to single bonds) can deviate from the $n_1^{th}$ and $(n_1 + D)^{th}$ by up to ±4. E.g., Fig. C.2 describes two configurations in which the two paired bases are at different positions and with different deviations. Just as in the case of $S_d(n, m)$ (see Eq. (B.6)) this leads to additional terms. However, while in the case of $S_d(n, m)$ these additional terms became irrelevant in the limit $n_1 \approx n_2 \approx N/2 \to \infty$, here some of the terms remain relevant and contain $Z_0(D)$, thus contributing to the power law dependence on $D$.

To explicitly derive the partition function $S_{dd}^{(11)}$ to the zeroth order in $(1 - \xi)$, we separate
Figure C.2: Two examples for the naive expectation for $S_{dd}^{(11)}$. In these two examples, the contractions from base pair stacks to single base pairs result in the same topology, whereas the explicit positions of the remaining base pairs and the distances between two inserted footprints are different. These small deviations from the $n_1^{th}$ and $(n_1 + D)^{th}$ base pairs are ignored in the naive estimation for $S_{dd}^{(11)}$. 
Figure C.3: Two configurations included in $S_{dd}^{(11)}$. The two footprints are in two consecutive stacks of the same stem, sharing a mutual bond. The partition function of these two configurations is very similar to $S_{dd}^*$, in which the two footprints are in the same stack.

The structures included in $S_{dd}^{(11)}$ into three types of configurations, described in Figs. B.2, C.3, and C.4, respectively. In all secondary structures described in these three figures, both footprints are in either different or the same base pair stack(s). However, they share the bonds of the base pair stack(s) in different ways.

In Fig. B.2, both footprints are in the same base stack of a stem. The partition function for this configuration is exactly $S_{dd}^*$, the zeroth order expansion of which has already been given in Eq. (B.10).

In Fig. C.3, the two footprints are in two consecutive stacks of the same stem, and thus share a mutual bond. Considering the similarity of Figs. C.3a and C.3b with B.2, their partition functions should be very similar to $S_{dd}^*$. In fact, the partition function for the structures in Fig. C.3a is given as

$$q^2Z_0(D - 1)\tilde{Z}(n_1 - 2; n_2 - 1) = q^3\xi Z_0(D - 3)Z_0(n_1 + n_2 - 3) + O((1 - \xi)), \quad (C.20)$$

where the zeroth-order approximations $\tilde{Z}(n_1 - 2; n_2 - 1) = Z_0(n_1 + n_2 - 3) + O((1 - \xi))$.
Figure C.4: The three types of structures included in \( S^{(11)}_{dd} - S^*_{dd} - S^{(11)}_{dd, mutual} \), i.e., when both footprints are in different stems or different base pair stacks of the same stem. Two footprints are inserted between the \( n_1^{th} \) and \( (n_1 + 1)^{st} \) and the \( (n_1 + D)^{th} \) and \( (n_1 + D + 1)^{st} \) nucleotide. All labels for segment lengths in the figures do not take into account the red dots, which form the bonds nearby the footprints.

(Eq. (B.2)) and \( Z_b(D - 1) = q\xi Z_0(D - 3) + O((1 - \xi)) \) (Eq. (B.3)) have been applied. The partition function for the structures in Fig. C.3b is obtained by exchanging the variables \( n_1 \) and \( n_2 \) which leads to the same result to zeroth order in \( (1 - \xi) \). Thus, the partition function for all structures including a mutual bond, which are described in Fig. C.3, is given as

\[
S^{(11)}_{dd, mutual} = 2q^3\xi Z_0(D - 3)Z_0(n_1 + n_2 - 3) + O((1 - \xi))
\]  

(C.21)
to the zeroth order in \( (1 - \xi) \).

The remaining structures in \( S^{(11)}_{dd} \) are described in Fig. C.4. The three Figs. C.4a, C.4b, and C.4c enumerate all structures in \( S^{(11)}_{dd} - S^*_{dd} - S^{(11)}_{dd, mutual} \) by the following strategy: all three possible configurations of the base pair stack comprising the first (left) footprint are shown separately in the three parts of the figure, and in each of these configurations all
possible configurations of the second (right) footprint are considered. Notice that since $S_{dd}^{(11)}$ has to be a symmetric function of $n_1$ and $n_2$, evaluating either the left or the right footprint as the “first” footprint makes no difference. We then define the partition functions for the structures considered in Figs. C.4a, C.4b, and C.4c as $S_{dd,1}^{(11)}$, $S_{dd,2}^{(11)}$, and $S_{dd,3}^{(11)}$ respectively, and calculate the partition function for the remaining structures as $S_{dd}^{(11)} - S_{dd}^* - S_{dd,mutual}^{(11)} = S_{dd,1}^{(11)} + S_{dd,2}^{(11)} + S_{dd,3}^{(11)}$.

The partition function for the structures included in Fig. C.4a is written down as

$$S_{dd,1}^{(11)} = q \sum_{k=2}^{n-1} Z_b(k) \tilde{S}_d(n_1 - k - 1; D - 1, n_2),$$  \hspace{1cm} (C.22)

where $\tilde{S}_d(n_1 - k - 1; D - 1, n_2)$ is the partition function for all structures formed by the nucleotides outside of the base pair stack containing the first footprint, with the condition that the $(n_1 + D)^{th}$ and $(n_1 + D + 1)^{st}$ nucleotide are one side of a base pair stack. We would like to calculate $\tilde{S}_d(n_1 - k - 1; D - 1, n_2)$ by simply removing the inserted stack and enclosed base pairs after the $(n_1 - k - 1)^{st}$ base since this would yield the quantity $S_d(n_1 + D - k - 2, n_2)$ which we have calculated before. However, there are two effects of removing the inserted stack. First, the removal will lead to replacements of a factor $\xi$ by 1 (or vice versa) in the loop or stack containing the removed section if loops are turned into stacks or vice versa. Since we are only interested in the zeroth order in $(1 - \xi)$ this can be ignored. Second, and more importantly, after removal of the left stack, the structures shown in Fig. C.5(b), in which the right insertion site is not part of a base stack and which are thus not part of $S_{dd}^{(11)}$ turn into the structures shown in Fig. C.5(a) which thus are included in $S_d(n_1 - k + D - 2, n_2)$. Thus, their contribution $qZ_0(D - 1)\tilde{Z}(n_1 - k - 2; n_2 - 1)$ needs to be subtracted yielding to zeroth order in $(1 - \xi)$

$$\tilde{S}_d(n_1 - k - 1; D - 1, n_2) = S_d(n_1 - k + D - 2, n_2) - q^2\xi Z_0(D - 3)Z_0(n_1 + n_2 - k - 3) + O((1 - \xi))$$

$$= q[Z_0(n_1 + n_2 + D - k - 4) - Z_0(n_1 + n_2 + D - k - 5)]$$

$$- q^2\xi Z_0(n_1 + D - k - 4)Z_0(n_2 - 2) - q^2\xi Z_0(D - 3)Z_0(n_1 + n_2 - k - 3)$$

$$+ O((1 - \xi))$$  \hspace{1cm} (C.23)

103
where we have used Eq. (B.6) in the second equality. Inserting this into Eq. (C.22) and using Eq. (B.3) for the zeroth order approximation of $Z_0(k)$ we get

$$S_{dd,1}^{(11)} = q^3 \xi \sum_{k=2}^{n_1-1} Z_0(k-2)Z_0(n_1+n_2+D-k-4)$$ (C.24a)

$$- q^3 \xi \sum_{k=2}^{n_1-1} Z_0(k-2)Z_0(n_1+n_2+D-k-5)$$ (C.24b)

$$- q^4 \xi^2 Z_0(n_2-2) \sum_{k=2}^{n_1-1} Z_0(k-2)Z_0(n_1+D-k-4)$$ (C.24c)

$$- q^4 \xi^2 Z_0(D-3) \sum_{k=2}^{n_1-2} Z_0(k-2)Z_0(n_1+n_2-k-3) + O((1 - \xi)).$$ (C.24d)

Notice that the last summation is up to $k = n_1 - 2$ instead of $k = n_1 - 1$ as for the other terms since the subtraction of the terms shown in Fig. C.5 is not necessary in the case $k = n_1 - 1$.

Similarly, the partition function for the structures in Fig. C.4b is written down as

$$S_{dd,2}^{(11)} = q \sum_{k=2}^{D-2} Z_0(k) \tilde{S}_d(n_1 - 1; D - k - 1, n_2).$$ (C.25)

Again, the two factors can be replaced by their zeroth order terms using Eqs. (B.3) and (C.23) yielding

$$S_{dd,2}^{(11)} = q^3 \xi \sum_{k=2}^{D-2} Z_0(k-2)Z_0(n_1+n_2+D-k-4)$$ (C.26a)

$$- q^3 \xi \sum_{k=2}^{D-2} Z_0(k-2)Z_0(n_1+n_2+D-k-5)$$ (C.26b)

$$- q^4 \xi^2 Z_0(n_2-2) \sum_{k=2}^{D-2} Z_0(k-2)Z_0(n_1+D-k-4)$$ (C.26c)

$$- q^4 \xi^2 Z_0(n_1+n_2-3) \sum_{k=2}^{D-3} Z_0(k-2)Z_0(D-k-3) + O((1 - \xi)).$$ (C.26d)

The partition function for the structures described in Fig. C.4c can be written down as
where $\tilde{S}_d$ is a partition function over the same structures as in $S_d$ with the only difference that the weights of structures in $\tilde{S}_d$ are calculated in the context of an enclosing base pair (the one from base $n_1 + 1$ to base $n_1 + D + k$ in Fig. C.4c) while the weights of the structures in $S_d$ are evaluated in an open context. The context of the enclosing base pair implies that the weights of nearly all structures get multiplied by $\xi$ for the outermost loop closed by that enclosing base pair with the exception of the structures in which the first and last base of the substrand described by $\tilde{S}_d$ are paired. The latter structures in turn contain an $\tilde{S}_d$ on a shortened sequence, i.e.

$$\tilde{S}_d(n, m) = \xi S_d(n, m) + (1 - \xi)q \tilde{S}_d(n - 1, m - 1).$$

To the zeroth order in $(1 - \xi)$ we may neglect the second term and thus find

$$S_{dd,3}^{(11)} = q^2 \sum_{k=2}^{n_2-1} \tilde{S}_d(D - 1, k - 1)\tilde{Z}(n_1 - 1; n_2 - k - 1),$$

Figure C.5: The structures which (a) contribute differently in $Z_0$ and $Z_d$ and thus included in $S_d$ but (b) contribute identically in the corresponding $\tilde{Z}$ and $\tilde{Z}_d$ and thus should not be taken into account in $\tilde{S}_d$. These structure have to be excluded when approximating $\tilde{S}_d$ as $S_d$. 

105
\[ S_{dd,3}^{(11)} = q^2 \xi \sum_{k=2}^{n_2-1} S_d(D-1, k-1) \tilde{Z}(n_1-1; n_2-k-1) + O((1 - \xi)) \]
\[ = q^3 \xi \sum_{k=2}^{n_2-1} Z_0(n_1 + n_2 - k - 2) Z_0(D + k - 4) \]
\[ - q^3 \xi \sum_{k=2}^{n_2-1} Z_0(n_1 + n_2 - k - 2) Z_0(D + k - 5) \]
\[ - q^4 \xi^2 Z_0(D-3) \sum_{k=3}^{n_2-1} Z_0(n_1 + n_2 - k - 2) Z_0(k-3) + O((1 - \xi)), \]
(C.29a)

where we have used the zeroth order expansions Eqs. (B.2) and (B.6) of \( \tilde{Z} \) and \( S_d \), respectively, in the second equality.

At this point, all partition functions for structures in \( S_{dd}^{(11)} - S_{dd}^* - S_{dd,mutual}^{(11)} \) have been written down to the zeroth order in \( (1 - \xi) \) in Eqs. (C.24), (C.26), and (C.29). The next task is then summing over all the terms in the three equations (a total of eleven terms, four in each of Eqs. (C.24) and (C.26), and three in Eq. (C.29)). This task is going to be accomplished by the following steps. First, the eleven terms will be combined into several subgroups and the summations in each of the subgroups will be evaluated and simplified separately. Finally, these results will be combined together into a final expression for \( S_{dd}^{(11)} - S_{dd}^* - S_{dd,mutual}^{(11)} \).

The first of these subgroups includes the terms (C.24a), (C.26a), and (C.29a). In order to combine these terms, we apply changes of summation variable to the summation in (C.26a)

\[ q^3 \xi \sum_{k=2}^{D-2} Z_0(k-2) Z_0(n_1 + n_2 + D - k - 4) = q^3 \xi \sum_{k' = n_1 + n_2 + D - 4}^{n_1+n_2} Z_0(n_1 + n_2 + D - k' - 4) Z_0(k' - 2) \]

and to the summation in (C.29a)

\[ q^3 \xi \sum_{k=2}^{n_2-1} Z_0(n_1 + n_2 - k - 2) Z_0(D + k - 4) = q^3 \xi \sum_{k' = n_1 + n_2 - 2}^{n_1+1} Z_0(k' - 2) Z_0(n_1 + n_2 + D - k' - 4), \]

which gives them the same form as the summation in (C.24a). Thus, these three terms can be combined to
\[ q^3 \xi \left( \sum_{k=2}^{n_1-1} + \sum_{k=n_1+1}^{n_1+n_2-2} + \sum_{k=n_1+n_2}^{n_1+n_2+D-4} \right) Z_0(k-2)Z_0(n_1+n_2+D-4-k) \]

\[ = q^2 \sum_{k=2}^{n_1+n_2+D-4} Z_b(k)Z_0(n_1+n_2+D-4-k) \]  \hspace{1cm} (C.30)

\[-q^3 \xi [Z_0(n_1-2)Z_0(n_2+D-4) + Z_0(n_1+n_2-3)Z_0(D-3)] + O((1-\xi)) \]

\[ = q^2 [Z_0(n_1+n_2+D-4) - Z_0(n_1+n_2+D-5)] \]

\[-q^3 \xi [Z_0(n_1-2)Z_0(n_2+D-4) + Z_0(n_1+n_2-3)Z_0(D-3)] + O((1-\xi)), \]

where we have used Eq. (B.3) in the first equality to replace \( q\xi Z_0(k-2) \) by \( Z_b(k) \) up to terms of order \( (1-\xi) \) and Eq. (B.5) in the second equality to express the summation as a simple combination of partition functions.

The second subgroup comprises the terms (C.24b), (C.26b), and (C.29b). Again, we apply a change of summation variable to the term (C.26b)

\[-q^3 \xi \sum_{k=2}^{D-2} Z_0(k-2)Z_0(n_1+n_2+D-k-5) = -q^3 \xi \sum_{k'=n_1+n_2+D-5}^{n_1+n_2-1} Z_0(n_1+n_2+D-k'-5)Z_0(k'-2) \]

and to the term (C.29b)

\[-q^3 \xi \sum_{k=2}^{n_2-1} Z_0(n_1+n_2-k-2)Z_0(D+k-5) = -q^3 \xi \sum_{k'=n_1+n_2-2}^{n_1+1} Z_0(k'-2)Z_0(n_1+n_2+D-k'-5) \]

such that again all three terms in the subgroup have the same form. Then, their combination can be simplified as

\[-q^3 \xi \left( \sum_{k=2}^{n_1-1} + \sum_{k=n_1+1}^{n_1+n_2-2} + \sum_{k=n_1+n_2}^{n_1+n_2+D-5} \right) Z_0(k-2)Z_0(n_1+n_2+D-5-k) \]

\[ = -q^2 \sum_{k=2}^{n_1+n_2+D-5} Z_b(k)Z_0(n_1+n_2+D-5-k) + q^3 \xi Z_0(n_1-2)Z_0(n_2+D-5) + O((1-\xi)) \]

\[ = -q^2 [Z_0(n_1+n_2+D-5) - Z_0(n_1+n_2+D-6)] + q^3 \xi Z_0(n_1-2)Z_0(n_2+D-5) + O((1-\xi)) \]  \hspace{1cm} (C.31)

using the same relations as above.

The third subgroup combines the terms (C.24c) and (C.26c). Upon applying the change of variables

107
\[-q^4\xi^2 Z_0(n_2-2) \sum_{k=2}^{D-2} Z_0(k-2) Z_0(n_1+D-k-4) = -q^4\xi^2 Z_0(n_2-2) \sum_{k'=n_1+D-4}^{n_1} Z_0(n_1+D-k'-4) Z_0(k'-2)\]

to the term (C.26c) it takes the same form as the term (C.24c) such that their combination can be simplified to

\[-q^4\xi^2 Z_0(n_2-2) \left( \sum_{k=2}^{n_1-1} + \sum_{k=n_1}^{n_1+D-4} \right) Z_0(k-2) Z_0(n_1 + D - k - 4)\]

\[= -q^4\xi Z_0(n_2-2) \sum_{k=2}^{n_1+D-4} Z_b(k) Z_0(n_1 + D - 4 - k) + O((1 - \xi)) \tag{C.32}\]

\[= -q^4\xi Z_0(n_2-2) [Z_0(n_1 + D - 4) - Z_0(n_1 + D - 5)] + O((1 - \xi)).\]

The fourth subgroup comprising terms (C.24d) and (C.29c) is similarly simplified through the change of summation variable

\[-q^4\xi^2 Z_0(D-3) \sum_{k=3}^{n_2-1} Z_0(n_1+n_2-k-2) Z_0(k-3) = -q^4\xi^2 Z_0(D-3) \sum_{k'=n_1+n_2-3}^{n_1+1} Z_0(k'-2) Z_0(n_1+n_2-k'-3),\]

applied to the term (C.29c) which renders it of the same form as the term (C.24d) and allows their combination into

\[-q^4\xi^2 Z_0(D-3) \left( \sum_{k=2}^{n_1-2} + \sum_{k=n_1+1}^{n_1+n_2-3} \right) Z_0(k-2) Z_0(n_1 + n_2 - 3 - k)\]

\[= -q^4\xi Z_0(D-3) \sum_{k=2}^{n_1+n_2-3} Z_b(k) Z_0(n_1 + n_2 - 3 - k) \tag{C.33}\]

\[+ q^4\xi^2 Z_0(D-3) [Z_0(n_1-3) Z_0(n_2-2) + Z_0(n_1-2) Z_0(n_2-3)] + O((1 - \xi))\]

\[= -q^4\xi Z_0(D-3) [Z_0(n_1 + n_2 - 3) - Z_0(n_1 + n_2 - 4)]\]

\[+ q^4\xi^2 Z_0(D-3) [Z_0(n_1-3) Z_0(n_2-2) + Z_0(n_1-2) Z_0(n_2-3)] + O((1 - \xi)).\]

The last of the eleven terms is term (C.26d). Using the same relations Eq. (B.3) and Eq. (B.5) as before, this term can be evaluated by itself as follows:
respectively, we finally obtain
\[
-q^4 \xi^2 Z_0(n_1 + n_2 - 3) \sum_{k=2}^{D-3} Z_0(k-2)Z_0(D-k-3)
\]
\[
= -q^3 \xi Z_0(n_1 + n_2 - 3) \sum_{k=2}^{D-3} Z_0(k)Z_0(D-k-3) + O((1 - \xi))
\]
\[
= -q^3 \xi Z_0(n_1 + n_2 - 3) [Z_0(D-3) - Z_0(D-4)] + O((1 - \xi)).
\] (C.34)

Collecting all five subgroups Eqs. (C.30), (C.31), (C.32), (C.33), and (C.34) and adding them to the zeroth order expressions Eqs. (B.10) and (C.21) for \( S_{dd}^{(1)} \) and \( S_{dd,\text{mutual}}^{(1)} \), respectively, we finally obtain
\[
S_{dd}^{(11)} = q^2 \xi Z_0(D-2)Z_0(n_1 + n_2 - 2) - q^3 \xi Z_0(D-3)Z_0(n_1 + n_2 - 3)
\]
\[
+ q^2 \left[ Z_0(n_1 + D + n_2 - 4) - 2Z_0(n_1 + D + n_2 - 5) + Z_0(n_1 + D + n_2 - 6) \right]
\]
\[
- q^3 \xi \left\{ Z_0(n_1 - 2) [Z_0(n_2 + D - 4) - Z_0(n_2 + D - 5)]
\]
\[
+ Z_0(n_2 - 2) [Z_0(n_1 + D - 4) - Z_0(n_1 + D - 5)] \right\}
\]
\[
+ q^3 \xi [Z_0(D-3)Z_0(n_1 + n_2 - 4) + Z_0(D-4)Z_0(n_1 + n_2 - 3)]
\]
\[
+ q^4 \xi^2 Z_0(D-3) [Z_0(n_1 - 3)Z_0(n_2 - 2) + Z_0(n_1 - 2)Z_0(n_2 - 3)] + O((1 - \xi)).
\] (C.35)

In the limit \( N/2 \approx n_1 \approx n_2 \gg D \gg 1 \), we can insert the asymptotic form Eq. (2.14) for \( Z_0 \). As before, any term in which \( n_1 \) and \( n_2 \) occur as arguments of different \( Z_0 \)’s depend on \( N \) as \( N^{-3} \) and can thus be neglected compared to the terms in which \( n_1 + n_2 \) is the argument of one \( Z_0 \) such that only the terms in the first, second, third, and sixth line contribute in the limit of large \( N \). Inserting the asymptotic form Eq. (2.14) for all \( Z_0 \) in these lines finally yields
\[
S_{dd}^{(11)} = \frac{\xi^2}{\delta^6} A \frac{(z - 1)^2}{N^{3/2}} z^{n_1+D+n_2} + \left( \frac{q^2 \xi}{\delta^4} - \frac{q^3 \xi}{\delta^6} + \frac{2q^3 \xi}{\delta^7} \right) A^2 \frac{1}{D^{3/2} N^{3/2}} z^{n_1+D+n_2}
\]
\[
+ O((1 - \xi), z^{n_1+D+n_2} D^{-5/2}, z^{n_1+D+n_2} N^{-5/2}).
\] (C.36)

We note that the first line of this result is precisely the asymptotic expansion of the naive expectation Eq. (C.35) while the second line has the same power law dependence on the distance \( D \) between the protein binding sites as the first order term of the correlation function \( g(D) \) calculated above.
C.4 Numerator of the correlation function

The numerator of the correlation function $g(D)$ is given by Eq. (B.13). We have already argued in Appendix B that the terms on the first line vanish in the limit of large $N$ through an argument that was independent of the expansion in $(1 - \xi)$. We also argued that the terms in the second and third line vanish in the limit of large $N$ to first order in $(1 - \xi)$. In principle, the terms on the second and third line could still yield a second order contribution in $(1 - \xi)$. However, using the first order expansion Eq. (C.15) of $S_d$, we find that the first order term of the asymptotic form of $S_d(n, m)$ is just $q/z^3$ times the zeroth order term. Therefore, the contributions of the differences in the second and third line to the second order in $(1 - \xi)$ must vanish in the limit $N \to \infty$ as well.

It is then clear that only the last two terms can contribute to the numerator $N$ of the correlation function $g(D)$ to second order in $(1 - \xi)$. Thus, the numerator of $g(D)$ in the limit of $N \gtrsim n_1 \approx n_2 \gg D \gg l \geq 1$ becomes

$$\frac{Z_{dd}}{Z_0} - \frac{Z_d}{Z_0} \times \frac{Z_d}{Z_0} = \xi (1 - \xi) \frac{S^*_{dd}}{Z_0(N)} + (1 - \xi)^2 \left[ \frac{S^{(11)}_{dd}}{Z_0(N)} - \frac{S_d}{Z_0(N)} \times \frac{S_d}{Z_0(N)} \right] + O((1 - \xi)^3, N^{-1}),$$  

(C.37)

in which the leading term is $O((1 - \xi))$. Dividing the asymptotic expressions for $S_d$, $S^*_{dd}$, and $S^{(11)}_{dd}$ calculated above (Eqs. (B.8), (C.18), and (C.36), respectively) by the asymptotic expression Eq. (2.14) for $Z_0$ yields

$$\frac{S^*_{dd}(n_1, D, n_2)}{Z_0(N)} = \frac{1}{z^{2l}} \left( \frac{q^2 \xi}{z^4} + (1 - \xi) \frac{q^3 \xi(z - 1)}{z^7} \right) \frac{A}{D^{3/2}} + O((1 - \xi)^2, N^{-1}, D^{-5/2})$$

$$\frac{S^{(11)}_{dd}(n_1, D, n_2)}{Z_0(N)} = q^2 (z - 1)^2 \frac{1}{z^{6+2l}} + \frac{1}{z^{2l}} \left( \frac{q^2 \xi}{z^4} - \frac{q^3 \xi(z - 2)}{z^7} \right) \frac{A}{D^{3/2}} + O((1 - \xi), N^{-1}, D^{-5/2})$$

$$\frac{S_d(n_1, D+1+n_2)}{Z_0(N)} = S_d(n_1 + l + D, n_2) = \frac{q(z - 1)}{z^{3+l}} + O((1 - \xi), N^{-1}, D^{-5/2}).$$  

(C.38)

in the limit of $N = n_1 + D + n_2 + 2l \gtrsim n_1 \approx n_2 \gg D \gg l \geq 1$. Substituting these fractions into Eq. (C.37) shows that the terms independent of $D$ cancel each other, thus yielding the asymptotic form of the numerator of $g(D)$ as
\[ N = \frac{A}{z^{2l}D^{3/2}} \left[ (1 - \xi) \frac{q^2 \xi^2}{z^4} + (1 - \xi)^2 \left( \frac{q^2 \xi}{z^4} + \frac{q^3 \xi}{z^5} \right) \right] + O((1 - \xi)^3, N^{-1}, D^{-5/2}). \]

### C.5 Denominator

The denominator of the correlation function \( g(D) \) is given by Eq. (B.16) and now has to be calculated to first order in \((1 - \xi)\). With the help of Eq. (2.15) and the asymptotic form of \( S_d \) and \( S_d^{*} \), the first ratio in Eq. (B.16) can be rewritten as

\[
\frac{Z_d(n_1, D + l + n_2)}{Z_0(N)} = \frac{Z_0(N - l)}{Z_0(N)} - (1 - \xi) \frac{S_d(n_1, D + l + n_2)}{Z_0(N)} \\
= \left( 1 - (1 - \xi) \frac{q(z - 1)}{z^3} \right) \frac{N^{3/2}}{z^l(N - l)^{3/2}} + O((1 - \xi)^2, N^{-1})
\]

By symmetry, the asymptotic form of the second ratio and the constant term and we have neglected other subleading terms in

\[
\frac{Z_{dd}(n_1, D, n_2)}{Z_0(N)} = \frac{Z_0(N - 2l)}{Z_0(N)} - (1 - \xi) \left[ \frac{S_d(n_1, D + n_2)}{Z_0(N)} + \frac{S_d(n_1 + D, n_2)}{Z_0(N)} \right] \\
- (1 - \xi)^2 S_{dd}^{(11)} \frac{S_{dd}}{Z_0(N)} + \xi (1 - \xi) \frac{S_{dd}^*}{Z_0(N)} \\
= \frac{1}{z^{2l}} \left[ 1 - \frac{2q(1 - \xi)(z - 1)}{z^3} + \frac{q^2 \xi^2(1 - \xi)}{z^4} \frac{A}{D^{3/2}} \right] + O((1 - \xi)^2, N^{-1}, D^{-5/2})
\]

using Eq. (2.19) in the first equality and the asymptotic expressions Eqs. (B.8) and (B.11) for \( S_d \) and \( S_d^{*} \), respectively, in the second equality.

Combining all four terms we find

\[
D = \left( 1 + \frac{c_1}{K_d^{(0)} z^l} \right) \left( 1 + \frac{c_2}{K_d^{(0)} z^l} \right) \\
- (1 - \xi) \left[ \frac{q(z - 1)}{z^{3+l}} \left( \frac{c_1}{K_d^{(0)} z^l} + \frac{c_2}{K_d^{(0)} z^l} \right) + \frac{2q(z - 1)}{z^{3+2l}} \frac{c_1 c_2}{K_d^{(0)} K_d^{(0)} z^{3+2l}} \right] + O((1 - \xi)^2, N^{-1}, D^{-3/2}).
\]

where we ignored the terms depending on the distance \( D \) between the binding sites since they are subleading to the constant term and we have neglected other subleading terms in
the distance $D$ in the numerator already as well.

**C.6 Correlation function**

Dividing Eq. (C.39) by the square of Eq. (C.42) yields the correlation function $g(D)$ with the overall shape

$$g(D) = (1 - \xi) \frac{A}{D^{3/2}} + O((1 - \xi)^3, N^{-1}, D^{-5/2})$$

where $A$ is in principle given by an explicit expression of the parameters $z$ and $A$ of the partition function $Z_0$, the loop cost $(1 - \xi)$ (up to first order), the concentrations $c_1$ and $c_2$ of the proteins, and the bare equilibrium constants $K_{d,1}^{(0)}$ and $K_{d,2}^{(0)}$ of the two binding sites. For small protein concentrations $c_i \ll K_{d,i}^{(0)} z^l$ this prefactor simplifies to

$$A_{\text{low } c} = A z^{2f} \left[ \frac{q^2 \xi^2}{z^4} + (1 - \xi) \left( \frac{q^2 \xi}{z^4} + \frac{q^3 \xi}{z^7} \right) \right].$$

and has to be evaluated numerically for arbitrary protein concentrations.
Appendix D

The stretching transition in molten phase driven by the linearized size-dependent loop cost

As described in Chapter 2.5, this stretching phase transition is identical to the transition in RNAs under tension, which has been well discussed in previous literatures [48, 57, 58]. We here include the calculation adapting the symbols we use for completeness.

The calculation is similar to that in Appendix A. In molten phase, the recursive equations in Eq. (2.23) reduce to their translationally invariant version,

\[
Z_b(N + 1) = q Z_b(N - 1) + q \xi (Z_m(N - 1) - Z_b(N - 1)) \tag{D.1a}
\]
\[
Z_m(N + 1) = \tilde{\xi} Z_m(N) + \sum_{k=0}^{N-1} Z_m(k) Z_b(N - k + 1) \tag{D.1b}
\]
\[
Z_0(N + 1) = Z(N) + \sum_{k=0}^{N-1} Z_0(k) Z_b(N - k + 1). \tag{D.1c}
\]

Substituting the relationship between $Z_b$ and $Z_m$ in Eq. (D.1a) and applying $z$-transformation to Eq. (D.1b) yield

\[
z^2 \left[ z^2 - q(1 - \xi) \right] \hat{Z}_b(z)^2 - (z - \tilde{\xi}) \left[ z^2 - q(1 - \xi) \right] \hat{Z}_b(z) + q \xi = 0, \tag{D.2}
\]

with the initial conditions $Z_b(0) = Z_b(1) = 0$, $Z_b(2) = q \xi$, and $Z_b(3) = q \xi \tilde{\xi}$. On the other hand, with the initial conditions $Z_0(0) = Z_0(1) = 0$, $Z_0(2) = 1 + q \xi$, $Z_0(3) = 1 + 2q \xi + q \xi \tilde{\xi}$,
applying $z$-transformation to Eq. (D.1c) yields

$$z\hat{Z}_0(z) - 1 = \hat{Z}_0(z) + z^2\hat{Z}_0(z)\hat{Z}_b(z).$$

(D.3)

Substituting the solution of $\hat{Z}_b(z)$ in Eq. (D.2) to Eq. (D.3) leads to the form of $\hat{Z}_0(z)$,

$$\hat{Z}_0(z) = \left( \frac{z - (2 - \tilde{\xi})}{2} + \sqrt{\frac{(z - \tilde{\xi})^2[z^2 - q(1 - \xi)]^2 - 4qz^2[z^2 - q(1 - \xi)]}{2[z^2 - q(1 - \xi)]}} \right)^{-1}.$$  (D.4)

The topological defect of $\hat{Z}_0$ with the greatest real part determines the form of $Z_0(N)$ in the inverse $z$-transformation: If there is a simple pole of which real part is greater than that of all branch cuts, the partition function would become an exponentially increasing function, $Z_0(z) \sim z^N$, instead of the branch-cut-leading form $z^N/N^{3/2}$ as in Appendix A. Below we derive the condition to have such simple-pole-leading partition function.

First, we consider the general condition for having a simple pole. A simple pole $z_p$ of satisfies $\hat{Z}_0^{-1}(z_p) = 0$, yielding

$$[z_p - (2 - \tilde{\xi})][z_p^2 - q(1 - \xi)] + \sqrt{(z_p - \tilde{\xi})^2[z_p^2 - q(1 - \xi)]^2 - 4qz^2[z_p^2 - q(1 - \xi)]} = 0.$$  (D.5)

This condition can be rewritten as

$$f(z_p) = 0,$$

with

$$f(z) = (z - 1)(1 - \tilde{\xi}) [z^2 - q(1 - \xi)] - q\xi z^2 = 0$$

under the constraint

$$[z_p - (2 - \tilde{\xi})][z_p^2 - q(1 - \xi)] < 0.$$  (D.6)

As $f(z)$ is a cubic function, we know there are three simple poles for $\hat{Z}_0(z)$, corresponding to the three solutions of $f(z) = 0$. By noticing

$$f(z \to \infty) > 0, f(\sqrt{q(1 - \xi)}) < 0, f(0) > 0, \text{ and } f(z \to -\infty) < 0,$$

we conclude that there are three simple poles, $z_p = z_1, z_2, z_3$, on the real axis, and they have the relationship

$$z_1 < 0 < z_2 < \sqrt{q(1 - \xi)} < z_3.$$  (D.7)
Next, we introduce the condition for branch cuts and examine its relationship with the greatest simple pole, $z_3$. All points $z$ at the branch cuts satisfy

$$(z - \tilde{\xi})^2[z^2 - q(1 - \xi)]^2 - 4q\xi z^2[z^2 - q(1 - \xi)] < 0.$$  \hfill (D.10)

Since we already know $z_3 > \sqrt{q(1 - \xi)}$, we discuss branch cuts in the region $z > \sqrt{q(1 - \xi)}$. The branch cut condition can thus be simplified as

$$g(z) \equiv (z - \tilde{\xi})^2[z^2 - q(1 - \xi)] - 4q\xi z^2 < 0.$$  \hfill (D.11)

The relationship between branch cuts and $z_3$ can be disclosed by substituting the simple pole condition $f(z_3) = 0$ into $g(z)$, leading to

$$g(z_3) = [z_3^2 - q(1 - \xi)](z_3 - 2 + \tilde{\xi})^2 > 0, \text{ for } z_3 > \sqrt{q(1 - \xi)},$$  \hfill (D.12)

which thus shows that $z_3$ is not in a branch cut.

We then examine whether there is any branch cut including points greater than $z_3$. We notice that $g(z)$ can be written as

$$g(z) = [z^2 - q(1 - \xi)][(z - \tilde{\xi})^2 - 4q\xi] - 4q^2\xi(1 - \xi),$$  \hfill (D.13)

thus clearly showing that it is a monotonically increasing function for all $z > \sqrt{q(1 - \xi)}$. Since in Eq (D.12) we have verified that $g(z_3) > 0$ and $z_3 > \sqrt{q(1 - \xi)}$, it is in certain that there is not any $z > z_3$ which satisfies the branch cut condition $g(z) < 0$. In conclusion, $z_3$ is greater than any of the $z$ in branch cuts.

Therefore, once the simple pole constraint in Eq. (D.7) is satisfied, the simple pole $z_3$ determines the form of the partition function $Z_0(N)$, and a stretching phase transition from $Z_0(N) \sim z^N/N^3$ to $Z_0(N) \sim z^N$ occurs. To satisfy the simple pole constraint and thus have a $z_3$ such that $\sqrt{q(1 - \xi)} < z_3 < 2 - \tilde{\xi}$, it is necessary to have

$$f(2 - \tilde{\xi}) = (1 - \tilde{\xi})^2[(2 - \tilde{\xi})^2 - q(1 - \xi)] - q\xi(2 - \tilde{\xi})^2 > 0.$$  \hfill (D.14)

The critical $q$ for the phase transition is thus read as

115
\[ q_c = \frac{(1 - \xi)^2(2 - \xi)^2}{(1 - \xi)(1 - \xi)^2 + \xi(2 - \xi)^2}. \] (D.15)