Multiscale Simulation and Analysis of Structured Ribonucleic Acids

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der Fakultät für Physik des
Karlsruher Instituts für Technologie

genehmigte

DISSERTATION

von

Dipl. Phys. Benjamin Lutz
aus Stuttgart

Tag der mündlichen Prüfung: 04.07.2014
Referent: Prof. Dr. Wolfgang Wenzel
Korreferent: Prof. Dr. Gerd Ulrich Nienhaus

Introduction

One of the most fundamental questions asked by humankind is the question where life originates from. Unraveling this mystery aims both at the initial, prehistorical origin of life and at the ongoing process of reproduction and evolution of new living organisms. Closely linked to these rather abstract conceptual questions are hundreds of years of macroscopic empirical efforts to preserve life by curing diseases and to decelerate physiological aging. The invention of an optical microscope in the seventeenth century by Antoni van Leeuwenhoek that could resolve bacteria, red blood cells, and spermatozoa gave access to observations of constituents that are important actors in today's understanding of the microscopic organization of life. From that time on biologists benefited from technological improvements of imaging methods that facilitated the direct observation of ever smaller objects in organisms. With a growing understanding of chemical reactions and structure, the link between biology and chemistry became apparent and created the scientific branches of biochemistry and structural biology. Biological processes could now be understood by interlinked sequences of chemical reactions and the constituents of life were recognized as biomolecular assemblies. At that point biophysical approaches also came into play: As well the development of advanced imaging techniques – meanwhile going far beyond the demands of lens grinders like van Leeuwenhoek – as the basic theoretical or computational description of biomolecular processes were and still are core competences of biophysicists and contribute to progress in the field of molecular biology. Today's research in the interdisciplinary fields of molecular biology, biochemistry, and biophysics is interested in how to influence the generation of proteins, i. e., how to promote or prevent their expression, and in understanding related biomolecular mechanisms. A pharmaceutically motivated subject of biophysical research is the investigation of chemical compounds that bind to protein domains. Docking experiments and simulations ultimately aim at finding targets for drugs in order to provoke desired physiological responses. Several regulatory mechanisms of gene expression have been identified and are subject to ongoing investigations. A more recent field of research related to biophysics is termed "bioinformatics" and employs methods of data mining, often based on statistical physics approaches, to extract information from the growing sequence databases. The inherent link between structure and function motivates the research of techniques that determine protein or ribonucleic acid (RNA) structure from a given sequence or a collection of evolutionarily related sequences. As mentioned before, a general concept in biological sciences are computational approaches that assist and complement experimental studies. The most successful technique in the context of simulation schemes of the last decades are molecular dynamics simulations. This approach combines parametrizations from empirical measurements and *ab-initio* quantum mechanical calculations to approximate the dynamics of molecular systems. Following the picture of van Leeuwenhoek's ground-breaking invention of a microscope, this class of techniques has been referred to as "computational microscope" [1]. The Nobel prize in chemistry of 2013 was awarded to three pioneers in the field of multiscale approaches in the context of biomolecular simulation that also utilize the molecular dynamics approach. The concept of multiscaling is based on the combination of coarse-grained elements at different levels of precision in the description of the modeled system. The different levels of precision can be descriptions based on quantum mechanics calculations, electrostatics, the generalization of attractive and repulsive forces and exclusion volumes. The different scales can also be realized by different geometrical representations of atom groups, molecular crowding, polymeric assemblies (quaternary structure), the three-dimensional fold (tertiary structure), canonical local substructural elements (secondary structure) or sequence information (primary structure). The work presented in this thesis addresses two issues in the context of multiscale analysis and simulation of RNAs: simulations of cotranscriptional riboswitch folding and advances to a novel approach of RNA structure prediction.

In the last decades the fundamental understanding of RNA has been drastically revolutionized and has shed light on the importance of RNA for gene expression. The traditional picture of RNA was that of a messenger (therefore “messenger RNA”, mRNA) that carries genetic information stored in deoxyribonucleic acid (DNA) sequences to the ribosomes for translation into proteins. Apart from being a mere transcript, various kinds of RNA have been discovered and the involved mechanisms are still investigated. Transfer RNA [2, 3], ribozymes [4] or riboswitches [5–9] are just some of the discovered classes of non-coding functional RNA. A possible role is the regulation of gene expression as it can be found in riboswitches. Representatives of this specialized class of structured RNA are contained in the non-coding region of messenger RNA preceding a gene and are able to perform conformational switching as a response to environmental conditions. The two-state switching affects transcription or translation of the downstream gene that is usually involved in metabolic functions according to chemical compounds, the “ligands”, that surround the switch. Riboswitches regulate metabolic processes via an interaction network that leads to the promotion or prevention of genetic expression. The initial decision of this interaction network, i. e., the conformational reaction to the presence of ligands, has to be made during or right after transcription which creates a competition between RNA synthesis, RNA folding and ligand binding. All three occur on the seconds time scale given the situation that ligand binding is not simply a surface docking but an entry in the interior binding pocket provided by the riboswitch. The substantial effort of experiments on biological systems and the given limits of experiments motivate the relevance of simulations that complement experiments. Indeed, the interplay of experiment and simulation is justified by the following limits in both approaches. Experiments are typically limited to comparably long timescales, often given by response times of apparatuses, and large length scales, often given by optical resolution limits and restrictions on the applicable techniques due to the soft matter nature of biomolecular systems. On the other hand, simulations are typically limited to short timescales due to sequential time integrations, and the struggle with large length scales since a larger system contains more particles which increases the numerical complexity. Therefore, simulations are able to reduce the total effort of studies by assisting experiments and enhance the understanding of processes by providing additional insight. In the case of riboswitch folding, simulations need to be based on a suitable model that is computationally tractable while reaching the seconds timescale. The standard approach would be to employ molecular dynamics simulations of the explicit system [10]. This simulation scheme is based on a molecular interaction potential determined by the topology of the investigated system. The potential terms feature simple, derivable mathematical forms and the constituent-specific force constants and quantities of equilibrium are compiled in look-up tables, often referred to as “force fields”. The first spatial derivative of the potential energy yields interatomic forces for Newton’s equations of motion that are solved by numerical time integration using a time step in the femtoseconds regime in order to be able to resolve interatomic vibrations. Reduced scalability due to communication between subdivisions of the system and long range interactions, such as Coulomb interactions, render potential energy evaluations at each time step computationally expensive for explicit systems. Therefore, this approach is currently restricted to simulated time spans in the milliseconds regime on highly specialized computer hardware [11]. In order to reach the seconds timescale required for RNA folding simulations it is necessary to employ an alternative approach: The native structure-based model [12–17] gives access to longer timescales while it allows to explore folding pathways towards the native folded state. The model is motivated by the assumption that the free energy landscape of a biomolecule in conformational space has been formed by evolution to be funnel-shaped towards its native state in order to guarantee experimentally observed folding times. In addition to the overall funnel characteristics of the energy landscape the principle of minimal frustration reduces kinetic traps on folding routes towards the native state. The gain of computationally tractable simulations comes at the price of energetic coarse-graining and a predefined “native” state at the energetic minimum. The native structure-based model is implemented by a potential energy that has its minimum at the native folded state by construction and incorporates native contacts that cause the desired cooperativity. The SMOG web-server [18] or the local installation eSBMTools [19] facilitate the automatized setup of native structure-based potential formulations that interface with standard software that provides the molecular dynamics time integration procedure.

The native conformations that are investigated in the course of this thesis are experimentally resolved tertiary structures of the sensing domains of two riboswitches (SAM-I [20] and *add* adenine [21]). The available structures are obtained by X-ray diffraction measurements of the compactly folded, ligand-bound state under stabilizing ion concentrations. The switching mechanisms of these two riboswitches are transcription termination by terminator hairpin formation (SAM) or translational repression by rendering the start codon inaccessible for the ribosome (adenine). The study presented in this thesis aims at improving the understanding of cotranscriptional riboswitch folding while keeping in mind that transcription and riboswitch folding are time-wise competing processes. Former studies of riboswitch folding by means of native structure-based model simulations focused on the characterization of free folding [22, 23]. I introduce a coarse-grained model for transcription that emulates the crowded environment of the RNA polymerase during transcription by imposing position restraints in form of an enclosing tube. The sequential transcription process itself is modeled by acting forces that extrude the stretched RNA strand out of the tube. Residues that have left RNA polymerase are released from the acting forces and are free to form secondary and tertiary contacts. The folding progress of helical substructures is analyzed by plotting the fraction of formed base pairs in helical substructures over the global folding progress of the riboswitch. This analysis allows the observation of characteristic folding pathways by distinguishing folding orders within the sensing region of the riboswitch. The computational approach facilitates the variation of transcription rates and explores their influence on the folding characteristics. Therefore it is possible to observe differences between free and cotranscriptional folding and between slow and fast transcription. As a main result of this study cotranscriptional folding is found to be transcription-rate limited in both investigated systems for physiologically relevant transcription rates [24]. This result is in robust agreement with a complementary computational study to which my results are compared.

The other scientific project discussed in this thesis is motivated by the rapidly growing number of RNA sequence families and their representatives in databases while experimental structure determination is still a demanding and complex procedure. As a consequence, the number of experimentally determined tertiary structures of RNA families is much smaller than the number of sequence families in databases. A striking attempt, therefore, is to explore the possibilities of deriving tertiary structure predictions from sequence alignments in databases. The procedure involves statistical physics approaches to find directly coupled coevolving nucleotides – the sequence building blocks of RNA – in sequence alignments, which indicate candidates for spatially close nucleotide pairs. The maximum entropy method yields the least constrained statistical model for the empiric single site and pair frequencies in sequence alignments. The model is solved by the technique of Lagrange multipliers that are found via independent-site and mean-field approximation. Direct coupling information can be calculated from the statistical model as a coupling score that ranks the predicted contacts. Therefore the described method is referred to as direct coupling analysis (DCA) [25]. The outcome of the analysis of an RNA family's sequence alignment from the database is a predicted contact map for that RNA family. The gained contact information can be transferred into a simplistic model similar to the native structure-based model, where quantities of equilibrium are not taken from the native fold but from knowledge-based look-up tables. The tables of bonded interactions and typical base pairing and stacking information (based on secondary structure) are derived from representative RNA structures. The geometric values of bonds, angles and planar dihedral angles are derived from histograms of geometric information in an experimentally measured, native structure. Values for proper dihedral angles are found in helical substructures and define the twist of helical regions in hairpin stem loops. Part of the helical substructures are also the non-bonded contacts that realize stacking interactions and base pairing. The predicted nucleotide-nucleotide contacts predicted by DCA are mapped onto a list of atom-atom contacts by an averaging procedure that introduces a cut-off condition for mean distance and standard deviation of a list of atom-atom contacts between the given nucleotides. The averaging is performed by the analysis of an existing list of typical representatives of these nucleotide-nucleotide contacts [26]. As a result, this knowledge-based model defines simulations that yield stable conformations that can be compared to known RNA structures. The comparisons are used as benchmarks for the assessment of the presented RNA structure prediction technique based on DCA contact predictions.

The **first chapter** compiles essential, basic information about ribonucleic acid (RNA) that is of relevance for the work presented in this thesis. The synthesis of RNA sequences by RNA polymerase (“transcription” process) is discussed. Furthermore, the interactions between RNA constituents (the “nucleotides”) realize the formation of canonical structure elements, such as hairpin stem loops. This chapter also introduces riboswitches as a subclass of structured RNA with their specific regulatory functions.

In the **second chapter** an overview over various aspects of the standard molecular dynamics simulation scheme [10] is given. The different steps – preparation of initial conditions, computing forces, computing energies, updating the conformation and output – are presented as the typical sequence of events. Thereby, the significance of force fields in the context of molecular dynamics simulations is discussed and different algorithms for numerical time integration are described. In order to realize ensembles with constant temperature or pressure, the systems need to be coupled to heat or pressure reservoirs which is implemented by various temperature and pressure coupling techniques. This chapter also discusses existing state-of-the-art software implementations of the molecular dynamics simulation scheme that are often accompanied by their own force field formulations.

Simplistic models in the context of biomolecular folding and dynamics are presented in the **third chapter**. In order to reach relevant timescales of RNA folding it is necessary to reduce the computational effort of molecular dynamics simulation. The presented approach is motivated by a basic theory of biopolymer folding. Levinthal proposed the picture of a multikinetic folding pathway along intermediate states to resolve the apparent paradox of finite folding times in proteins [27]. This picture was extended by energy landscape theory [28–30] that postulates funnel-shaped energy landscapes for foldable biopolymer sequences. The principle of minimal frustration introduces cooperativity that guides folding by a network of native interactions towards the native state of the biopolymer – first discussed by a simplistic hydrophobic-polar model [31] – and is implemented by the native structure-based model [12–17]. In the course of this thesis, this model is used to simulate cotranscriptional riboswitch folding in a multiscale setup and modified to incorporate tertiary contact predictions into RNA folding simulations. The last presented simplistic model is a kinetic Monte Carlo approach to RNA folding simulations [32] that provides comparable results to native structure-based model simulations as part of a collaboration.

An introduction to coevolutionary statistical sequence analysis is discussed in the **fourth chapter**. The presented method is searching for the least constrained statistical model that reproduces the empirical single-site and pair frequency counts in sequence alignments [25]. This statistical model disentangles direct from indirect coupling and correlation information can be ranked by its direct information scores. Directly coupled sequence units are indicative of spatial closeness in the structure associated with the according sequences, which renders predicted coupling with a high direct information score a good candidate for a structural contact. The method, referred to as direct coupling analysis (DCA), has been successfully employed to predict contact maps of proteins and protein complexes. The idea presented as part of the results of this thesis is to assess the possibilities of the applicability of direct coupling analysis in the field of structured RNAs.

Chapters five, **six**, and **seven** elaborate on the three projects that represent my scientific accomplishments in the context of multiscale simulation and analysis of structured ribonucleic acids. First, the software implementation of native structure-based models that facilitated the work presented in this thesis is presented. The project "eSBMTools" [19] was initiated and has been accompanied by me in the course of this thesis. Secondly, my study on cotranscriptional riboswitch folding is presented. I propose a multiscale representation of transcription realized by a native structure-based model and analyze folding pathways of two riboswitch aptamer regions. The outcome suggests that cotranscriptional riboswitch folding is transcription-rate limited which is also backed by the comparison with a kinetic Monte Carlo study by Michael Faber as part of a collaboration [24]. The third project in collaboration with Eleonora De Leonardis is intended to transfer an established statistical contact prediction model in the context of proteins (the direct coupling analysis, DCA [25]) to RNA contact prediction and combine it with coarse-grained simulations in order to advance to RNA tertiary structure prediction. The ranked contact predictions by DCA are dominated by canonical base pairing contacts but also consist of tertiary contacts. A modified formulation of the standard native structure-based model is employed to test the ability to get stable folds from predicted contact maps which motivates the application of the model in the context of tertiary structure prediction.

The **last chapter** summarizes the important aspects of the conducted investigations. Finally, it gives an outlook to further investigations that are motivated by the outcome of this thesis.

Contents

List of Figures

1 Chapter 1. Ribonucleic Acid

The first chapter provides an overview of the different aspects of ribonucleic acid (RNA) that are essential to this thesis. RNA plays a crucial role in gene expression, the process of creating proteins according to information stored in genes. This process starts with the transcription of sequence information contained in genes that consist of deoxyribonucleic acid (DNA). RNA as the transcription product is translated by the ribosome into the final protein. In this simplified picture of gene expression RNA fills the role of a messenger, therefore referred to as messenger RNA (mRNA). Meanwhile, a more detailed view of processes involving RNA has emerged that assigns various tasks to different kinds of RNA.

After presenting a compact overview of RNA, I discuss the biochemistry and biophysics of RNA synthesis, folding and ligand binding. Due to the biochemical features of RNA distinct characteristic structural elements can be formed during and after transcription and RNA is able to fold into a native conformation. In addition, RNA can form binding pockets that detect chemical compounds, often referred to as "ligands", with high specificity. All three processes – synthesis, folding and ligand binding – are interdependent processes that take place at comparable time scales.

The next section introduces the RNA polymerase (RNAP) as the cellular machinery that synthesizes RNA. The sequence of building blocks is read from DNA and a corresponding complementary strand of RNA is simultaneously synthesized in a step-by-step manner. This process is referred to as "transcription". RNAP is a complex polymeric protein that has to perform different tasks – DNA sensing, DNA melting, DNA read-out, RNA synthesis, RNA and DNA release – in a coordinated fashion.

The last section of this chapter describes riboswitches as representatives of structured mRNA in the untranslated region (UTR). Riboswitches are bistable structural switches that can terminate transcription or attenuate translation. The equilibrium between the two stable structural states is shifted by the presence of ligands, chemical compounds involved in biomolecular processes via binding.

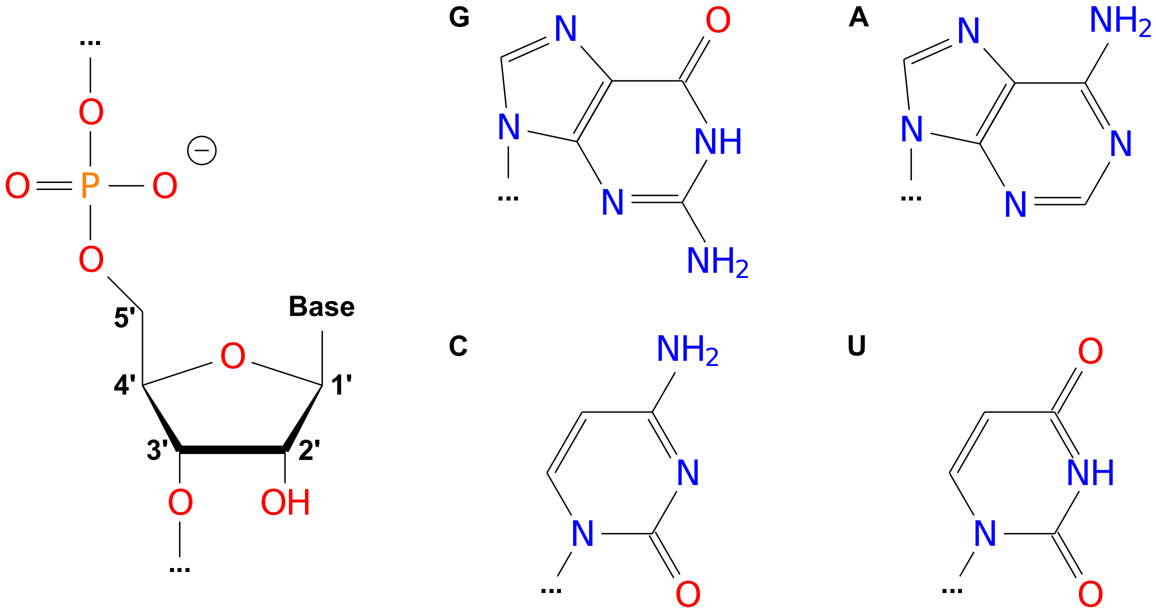


Figure 1.1.: Structural formulae of RNA. The backbone of nucleoside monophosphates (left) consists of a phosphate group and a pentose, a sugar ring with five carbon atoms, of which one is adjacent. The backbone is connected at the **1'** carbon atom in its pentose to one of the four organic bases guanine G, cytosine C, adenine A or uracil U via a glycosidic bond. The next residue along the sequence is attached to the oxygen atom at the **3'** end of the chain.

1.1. Overview

Gene expression is the process of transferring genetic information, stored in deoxyribonucleic acid (DNA) double helices that are assembled in chromosomes, to the production of respective proteins, the building blocks and machines of life. Genetic information is stored in linear sequences of DNA that consist of the deoxyribonucleotide building blocks deoxyadenosine, deoxythymidine, deoxyguanosine, and deoxycytidine. A unit of three consecutive deoxyribonucleotides that specify an amino acid, the building blocks of proteins, is called a “codon”. A simple picture of gene expression features two steps: During *transcription*, a corresponding strand of ribonucleic acid (RNA) is generated by the RNA polymerase (RNAP) as a sequence of nucleotides that match the respective deoxyribonucleotides in DNA. The building blocks (and their respective DNA counterparts) are adenosine (deoxythymidine), uridine (deoxyadenosine), guanosine (deoxycytidine), and cytidine (deoxyguanosine). Their chemical structures are shown in Fig. 1.1 and their composition and synthesis are discussed in more detail in Sec. 1.2. During *translation*, the sequence information contained in RNA is translated by the ribosome into proteins, where each codon of three consecutive nucleotides corresponds to an amino acid. This simplistic picture of gene expression has been more and more refined over the last decades and various roles of RNA, besides being a messenger (therefore: messenger RNA, mRNA), have been discovered [2–4, 33, 34]. A common characteristic among several types of RNA is the ability to form stable folds which are then referred to as structured RNA. Structured RNA can perform regulatory functions based on the inherent link between structure and function. A more detailed description of RNA and its structure, a discussion about the RNAP and an introduction to a specific class of structured RNA, relevant for the studies presented in this thesis, are given in the following sections.

1.2. Biochemistry and Biophysics

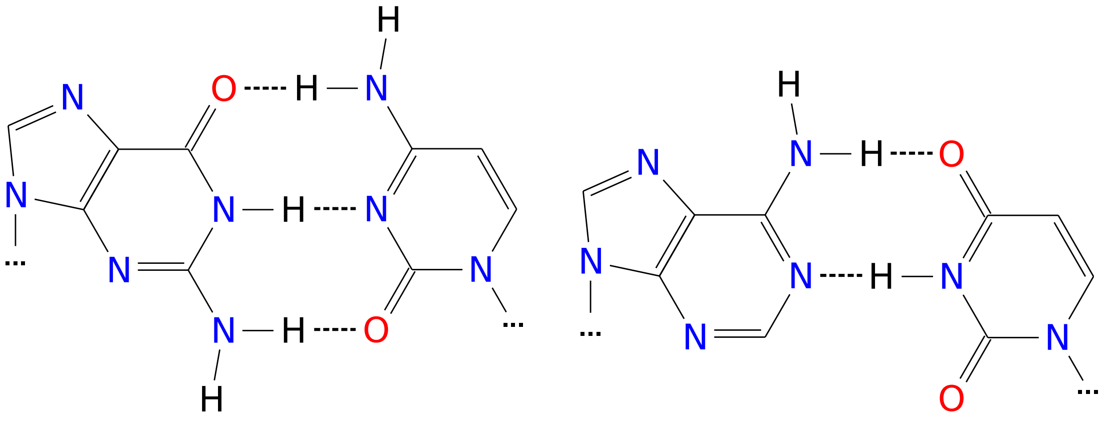
This section describes the biochemical and biophysical mechanisms that determine the synthesis and structure formation of ribonucleic acids.

1.2.1. Synthesis

RNA synthesis is a sequential loop process executed by the RNA polymerase (RNAP, described in Sec. 1.3) that elongates a strand of RNA according to a given complementary sequence of DNA. Transcription takes place in a complex of RNAP, DNA and nascent RNA that is called the “transcription bubble” [35]. An existing RNA chain RNA_n of length n is located with its 3’ end at the catalytic center in RNAP. Nucleoside triphosphate (NTP; nucleosides: adenosine, cytidine, guanosine or uridine triphosphate) as a generalized building block for RNA reaches the catalytic center via a channel from outside RNAP matching the present base in the DNA. Pyrophosphate ($\mathrm{P_2O_7^{4-}}$, PP_i) is cleaved from NTP and the remaining nucleoside monophosphate (NMP) is bound to the 3’ end of the existing RNA strand. Subsequently, PP_i is released and the strand translocated, extruding the grown nascent RNA_{n+1}. This circular process repeats until one of several sequence motifs [36], such as a terminator loop followed by a stretch of uridines, destabilizes the transcription bubble and the RNA as well as the DNA are released.

1.2.2. Structure

Structured RNA is able to form stable folds defined by secondary and tertiary structural elements. The secondary structure is defined by base pairs that form helical conformations. Base pairs are formed between the bases in side chains of RNA by hydrogen bonds. There are several possible realizations [37] of which the most stable ones are the “canonical” Watson-Crick (see Fig. 1.2) and Wobble base pairs (see Fig. 1.3). Depending on the type of base pair there are two or three hydrogen bonds realized between the involved bases. Secondary structure is realized via sequential pairings between bases that form a single stranded, two-dimensional assembly, as shown in Fig. 1.4. The dominant secondary structural element is the “hairpin” consisting of a helical stem that is realized by a single-stranded double helix and the connecting non-paired loop region. The compact arrangement is energetically favored due to stacking of the side chain bases that is physically explained by aligning dipole-dipole interactions between the rings of the bases. The other, related structural element is the non-local helix that pairs both ends of a large loop that contains one or more hairpins itself.



(a) Guanine and cytosine

(b) Adenine and uracil

Figure 1.2.: Structural formulae of Watson-Crick base pairing. The pairing between two bases of RNA is realized by hydrogen bonds. The canonical Watson-Crick pairing of guanine and cytosine (G-C) consists of three **(a)** and the pairing of adenine and uracil (A-U) of two hydrogen bonds **(b)**.

It assembles exactly like a local hairpin loop but has to overcome a higher entropic barrier because of the substantially increased loop size.

The three-dimensional arrangement of secondary structure elements that determines the geometric shape of RNA is referred to as the tertiary structure. The arrangement is fixed in place by tertiary contacts, i. e., single base pairs or non-canonically paired bases depending on relative orientation of bases to each other, as discussed in [37]. Common structural motifs in the tertiary structure are kink turns, where the strand is maximally bent to realize the other boundary conditions, or pseudoknots. Pseudoknots are two or more stem loops that feature connections to each others loop regions which results in a knot-like structure.

Quaternary structure denotes polymeric complexes, often consisting of RNA and proteins, such as ribosomal RNA [33], or RNA and bound ligands, as shown in Fig. 1.5. The bound conformations with ligands are stabilized by mediated tertiary and quaternary contacts in the vicinity of the ligand binding pocket. This binding mechanism is used by a subclass of messenger RNA, the riboswitches, that feature interior binding pockets and conformational switching in the presence of ligands, as discussed in Sec. 1.4.

1.3. RNA Polymerase

In this section the biomolecular machinery that sequesters the RNA, the RNA polymerase (RNAP), is introduced. RNA is transcribed from deoxyribonucleic acid (DNA) as a complementary strand by RNAP. Part of the information contained in RNA is subsequently translated by the ribosome to proteins. The residual regions in the sequence are referred to as the untranslated regions (UTR), as shown in Fig. 1.6.

The RNAP is attracted by promoter sequences along the DNA to which RNAP docks with its alpha and sigma subunits [38]. Subsequently, RNAP encompasses the double helix

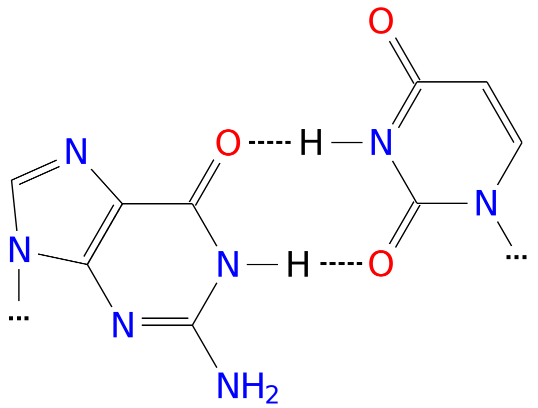
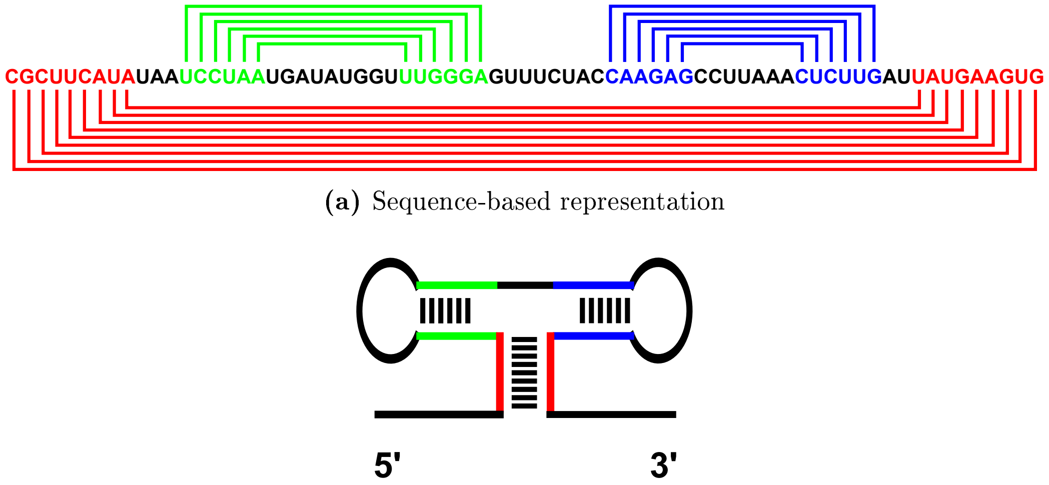


Figure 1.3.: Structural formula of Wobble base pairing. The non-Watson-Crick base pairing of guanine and uracil (G-U) consists of two hydrogen bonds.


CGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUG
5'
3'

(a) Sequence-based representation

(b) Schematic two-dimensional representation

Figure 1.4.: Secondary structure diagrams of an adenine sensing riboswitch [21] in its ligand-bound conformation. The formed base pairs are represented by lines connecting two bases. The local helices are colored in green and blue and the non-local helix tying up both ends of the riboswitch is colored in red. The two possible standard representations are shown here: the clear-cut sequence-based representation in **(a)** gives more detailed sequence information whereas the more graphic representation in **(b)** illustrates local and non-local helical substructures.
The figures are taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/)

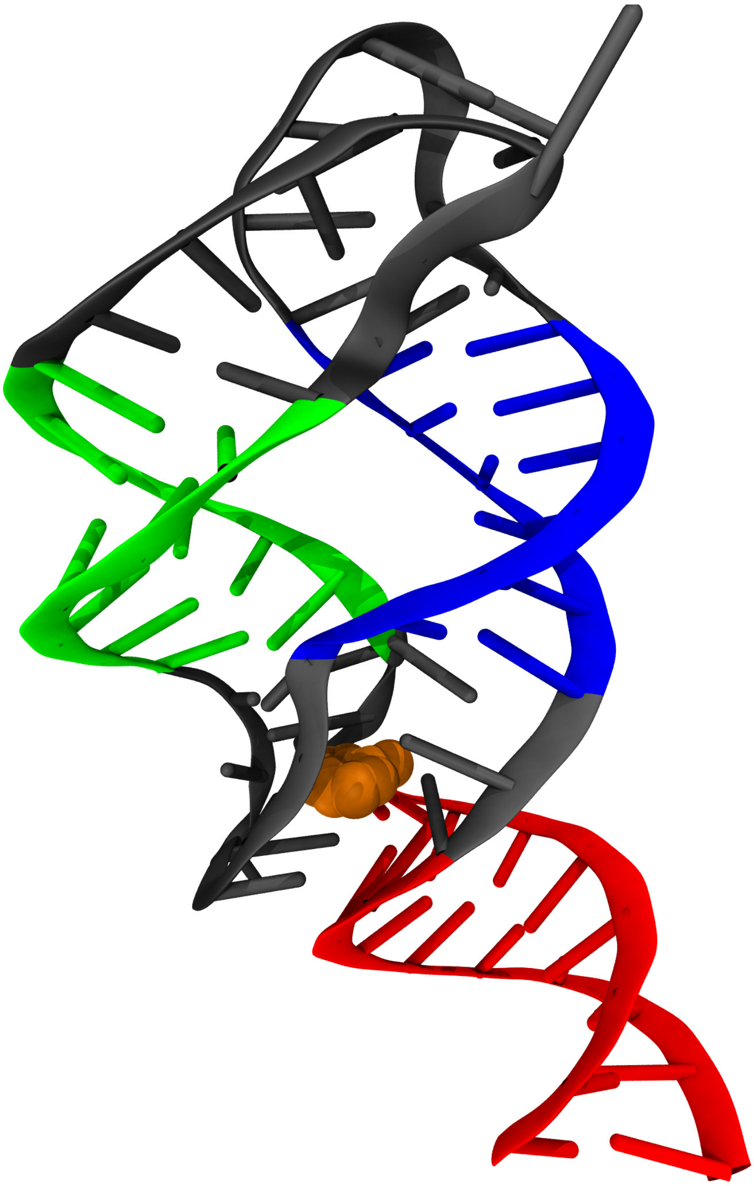


Figure 1.5.: Cartoon representation of tertiary RNA structure and a bound ligand. The picture depicts an *add* adenine riboswitch (PDB ID 1Y26 [21]). The local helices are colored in green and blue and the non-local terminal helix is colored in red. The bound ligand, adenine, is highlighted in orange and its binding pocket between the two coaxially stacked red and blue helices encloses the ligand. The ligand mediates stabilizing contacts that keep the ligand tightly bound and closed up once the ligand has entered the pocket.
The figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

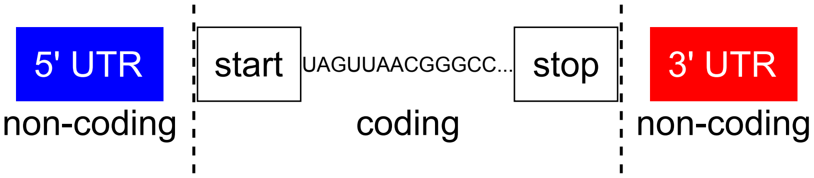


Figure 1.6.: Schematic representation of mRNA. The actual coding region of mRNA that represents the encoded protein sequence is encompassed by two non-coding UTRs. Riboswitches are part of the 5' UTR since they either terminate transcription as a structural response to the already transcribed aptamer conformation or they attenuate translation by preventing the ribosome to access the start codon AUG.

of DNA with two claw-like substructures, straightens out the double-helical DNA region in between and melts the involved DNA base pairs apart. The RNAP is moving along the isolated DNA sequence and successively reading the sequence information contained in the single strand. According to the isolated single strand of DNA, RNAP sequences a new strand of RNA at its catalytic center. The thereby formed complex of encompassing RNAP, melted DNA and emerging RNA is referred to as the transcription bubble. The growing nascent strand of RNA leaves the polymerase via the exit channel, another substructural unit formed by a flexible claw. As soon as the RNA leaves the exit channel of RNAP it is free to form secondary and tertiary contacts and fold. The positional restraints of the exit channel enforce sequential folding of substructural elements which has been shown in recent experimental [39] and computational [24] studies.

1.4. Riboswitches

Riboswitches are part of the UTR of mRNA at the 5' prime end, as seen in Fig. 1.6, and are able to modulate gene expression depending on the current environmental conditions [5–9]. They can react to metabolic compounds in their vicinity with high specificity. They consist of two structural subunits: aptamer region and expression platform. The aptamer region features a binding pocket for specific metabolites and reacts structurally to their presence. For example, the binding of a metabolite can promote the formation of a non-local helix. Contacts between the metabolite and the aptamer region mediate the stabilization of a compact formation that closes the binding pocket around the metabolite. As a response, the expression platform is able to fold into a conformation that terminates transcription or attenuates translation of the downstream gene. Transcription termination is often achieved by a hairpin stem loop followed by a uridine-rich sequence, which is, e. g., the mechanism in a SAM-I riboswitch [20]. Translation can be inhibited by obscuring the start sequence AUG from the ribosome inside a hairpin stem, as it is found in an *add* adenine riboswitch [21]. Since the synthesis (characterized by the transcription rate), folding (characterized by the folding rate) and ligand binding (characterized by the binding affinity) are interdependent processes on comparable time scales the investigation of cotranscriptional riboswitch folding is a topic of current research.

Their high specificity makes them a target for potential repression or promotion of metabolic mechanisms. As of today, research in this field is still limited to observing and determining riboswitch mechanisms instead of designing novel structures and interaction networks. A recent study [40] investigates, e. g., the biological implications of their discovered fluoride riboswitch. They report a class of fluoride riboswitches that boost the expression of proteins that counteract the toxic environmental influence of fluoride in bacteria.

1.4.1. Experimental Riboswitch Structures

Riboswitches are discovered by bioinformatics approaches that identify sequence motifs in RNA sequence data sets that are suspected to play a regulatory role, such as terminator hairpin stem loops. In the vicinity of such motifs ligand specific aptamer regions are likely to be found [41]. The identified suspect of such an aptamer region together with its expression platform is sequenced and exposed to a range of typical metabolites. The expression levels of a downstream gene can then be recorded to study the affinity and specificity of the riboswitch suspect. In case of a positive match, experimental techniques, such as X-ray crystallography or nuclear magnetic resonance (NMR), are used to resolve the riboswitch structure. In the last two decades several high-resolution crystal structures have been deposited in the RCBS Protein Data Bank (PDB) [42]. An exemplary subset of available structures is listed with their PDB IDs in parentheses:

- adenine (1Y26) [21]
- M-box (2QBZ) [43]
- glycine (3OWI) [44]
- lysine (3DIL) [45]
- FMN (3F2Q) [46]
- TPP (2HOJ) [8]
- SAM-I (2GIS) [20]
- c-di-GMP (3IRW) [47]
- fluoride (3VRS) [48]

The compilation of these structures will be motivated in Chap. 7 and referred to as a “gold standard” for the assessment of RNA contact predictions. Concrete studies about predicted contacts in simulations of riboswitches are presented in the same chapter for structures 1Y26, 2GIS, 3OWI and 3VRS. Furthermore, a computational analysis of cotranscriptional riboswitch folding is conducted in Chap. 6 for riboswitches 1Y26 and 2GIS.

2 Chapter 2. Molecular Dynamics

This chapter introduces molecular dynamics (MD) simulation as the established tool to model biopolymer dynamics. It is referred to as the "computational microscope" [1] since it has been proven to give access to beforehand unobservable processes, similar to the invention of the optical microscope by Antoni van Leeuwenhoek. The MD simulation scheme in its standard formulation generates the total potential energy of each particle in a system of interest from empirical look-up tables ("force fields"). The according forces on each particle are derived from the potential energy and Newton's equations of motion are solved by numerical integration of the equations over time: Starting from the initial conditions, positions and velocities of all particles are calculated for the next finite time step and forces are reevaluated. The desired temperature or pressure conditions are introduced via a wide range of possible coupling techniques.

First, I describe the general protocol within the MD simulation scheme. Acquiring initial conditions, generating the inter-atomic potentials and algorithms for numerical time integration of Newton's equations of motion are discussed. This section also covers various techniques to introduce temperature and pressure coupling to the model, such as Berendsen coupling or Langevin dynamics. In addition, the concept of steered molecular dynamics is presented. This concept introduces additional external forces to a system of interest by time dependent constraints.

At the end of this chapter I review several implementations of MD software and force fields. CHARMM, AMBER, NAMD, and GROMACS are the most widely used software packages that are architecture independent and offer a wide range of pre- and postprocessing tools in addition to their core functionality. GROMACS offers a very flexible interface for force field parametrizations which makes it favorable for simulations based on non-standard force field definitions. My studies are based on such a modified force field definition – a native structure-based model – that will be introduced in the next chapter as a simplistic theory within the MD frame work.

2.1. Molecular Dynamics Simulation Scheme

Molecular dynamics (MD) simulations are a subclass of the molecular mechanics concept that describes molecular systems by atomic particles under the influence of classical potential-energy functions [10]. Atoms are usually represented by spheres that interact with each other due to assigned properties, such as charge or the ability to share valence electrons, which is modeled by effective interaction functions. The simulations give access to experimentally inaccessible time scales and structural resolution levels. Therefore, MD simulations are an established tool to complement experiments and to assist investigations of biomolecular systems. The standard MD simulation scheme [49] consists of the following five steps:

1. Prepare initial conditions:

   The atom positions $\mathbf{r}_i$ and velocities $\mathbf{v}_i$ in a system of interest are acquired and the position dependent interaction potential $V(\{\mathbf{r}_i\})$ is determined accordingly.

   Positions are read from standardized coordinate files, velocities are generated due to thermal conditions and the interaction potential is calculated based on a combination of defined mathematical formulations and according parameter lists.

2. Compute forces:

   The force acting on an atom i is computed by

   $$\mathbf{F}_i = -\frac{\partial V}{\partial \mathbf{r}_i}\,, \tag{2.1}$$

   which is effectively a summation over non-bonded atom pair forces $\mathbf{F}_i = \sum_j \mathbf{F}_{ij}$ and bonded interaction forces. In addition, external forces or position constraints are evaluated.

3. Compute energies:

   Potential and kinetic energy values and the pressure tensor are computed. This evaluation is needed for temperature and pressure coupling techniques, as discussed in more detail in Sec. 2.1.4.

4. Update configuration:

   Newton's equations of motion

   $$\frac{\mathrm{d}^2\mathbf{r}_i}{\mathrm{d}t^2} = \frac{\mathbf{F}_i}{m_i} \tag{2.2}$$

   are solved numerically by time integrations with a finite time step Δt starting from the actual time t_0. Thereby, the coordinates of the system are updated to a later point in time $t_0 + \Delta t$. The employed integration scheme depends on the the chosen coupling techniques or the algorithm used to impose constraints.

5. Output system:

The positions, velocities, energies, temperature, pressure, etc., are evaluated, compiled and written to a file. Steps 2 - 5 are repeated n_t times until the desired period of time is covered: $t_{\text{final}} = n_t \cdot \Delta t$. The sequence of system properties at each time step is referred to as "trajectory".

2.1.1. Molecular Dynamics Potential

According to the Born-Oppenheimer approximation, the potential energy function can be a pairwise additive function of atom coordinates. The standard formulation of the MD potential [10, 49] reads as

$$\begin{aligned} V(\{r, R, \theta, \chi, \phi, r_{ij}\}) = & \sum_{\text{bonds}} K_{\text{b}}(r - r_0)^2 \\ & + \sum_{\text{Urey-Bradley}} K_{\text{UB}}(R - R_0)^2 \\ & + \sum_{\text{angles}} K_{\text{a}}(\theta - \theta_0)^2 \\ & + \sum_{\text{dihedrals, impropers}} K_{\text{i}}(\chi - \chi_0)^2 \\ & + \sum_{\text{dihedrals, propers}} K_{\text{d}}\left[1 - \cos(M(\phi - \phi_0))\right] \\ & + \sum_{\substack{\text{non-bonded} \\ i,j}} \left[K_{\text{c}} \left[\left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon_r \epsilon_0 r_{ij}} \right] . \end{aligned}$$

In this potential the zero indexed quantities $r_0, R_0, \theta_0, \chi_0, \phi_0, \sigma_{ij}^0$ represent the equilibrium values. The bonded interactions – bonds, Urey-Bradly bonds, angles and dihedral (or torsional) angles, as depicted in Fig. 2.1 – are expressed in a harmonic approximation. The proper torsional angles are represented by a periodic potential with a possible multiplicity M. Non-bonded interactions are introduced by Lennard-Jones potential terms and Coulomb interactions, where q_i, q_j are charges, ϵ_0 is the electric permittivity constant and ϵ_r is the relative permittivity of the surrounding medium. The force constants K_{b},K_{UB},K_{a},K_{i},K_{d},K_{c} describe the strengths of the respective interactions.

The collection of parameters (equilibrium values, force constants, charges, permittivities, etc.) and the functional shapes of the involved potential terms are referred to as a "force field". The cataloged values are experimentally measured or derived from quantum mechanical calculations. The basic assumption of the parametrization is that parameters derived from studies based on partial subsystems can be transferred and assembled to larger molecular systems.

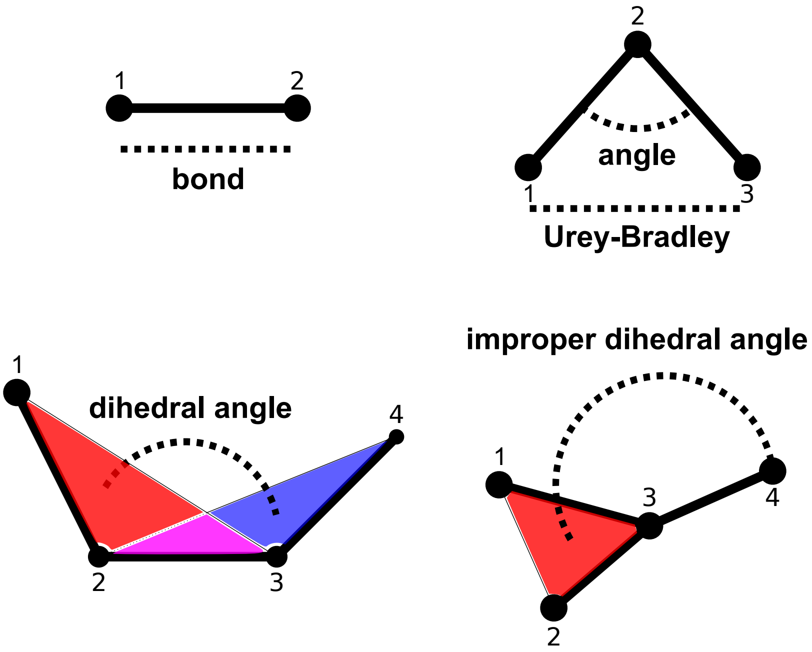


Figure 2.1.: Pictograms of bonded interactions relevant for MD potentials. The simplest interaction is the bond, a (1,2) interaction characterized by the distance between two atoms (top left). Angle and Urey-Bradley interaction are (1,3) interactions (top right). The angle is measured by the span between two line segments and Urey-Bradley interaction by the distance between the first and third atom. There are two kinds of (1,4) interactions: proper dihedral angle (bottom left) and improper or planar dihedral angle (bottom right). The proper dihedral angle is defined by the angle between the two planes given by the first three (1-2-3) and the last three (2-3-4) atoms of a sequence of four atoms. The improper dihedral angle is defined by the angle between a plane given by three (1-2-3) atoms and a fourth (4) atom of four atoms that form an intersection. The term planar dihedral angel is also used since this angle can guarantee the planarity of rings.

2.1.2. Velocity Generation

The initial generation of the particle velocities v_i under the given thermal condition T follows the Boltzmann velocity distribution

$$p(v_i) = 4\pi \left(\frac{m_i}{2k_\mathrm{B}T}\right)^{3/2} v_i^2 \exp\left(-\frac{m_i v_i^2}{2k_\mathrm{B}T}\right) , \tag{2.3}$$

where $p(v_i)$ is the probability distribution to find particle i at velocity v_i; m_i are the particle masses and k_B is Boltzmann's constant.

2.1.3. Time Integration

Newton's equations of motion that are derived from the interatomic potential need to be integrated numerically in time to gain the spatial trajectories of the particles $\mathbf{r}_i(t)$. The numerical integration scheme is a finite difference method that introduces a finite time step

$$\Delta t = \frac{t_\mathrm{final}}{n_t} , \tag{2.4}$$

where n_t is the number of time steps to reach time t_final. The time step is usually chosen in the order of femtoseconds to be able to resolve interatomic vibrations [10].

The desired integration scheme is required to be time reversible, i. e., after n integrations forward in time, n integrations backward in time should yield the same initial state. Apart from that, the integrator needs to be symplectic, i. e., it guarantees the conservation of the total energy in the system. A first possible algorithm that meets this requirements is the simple Verlet algorithm [50]. The updated coordinates are calculated by

$$\mathbf{r}(t+\Delta t) = 2 \cdot \mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \frac{\mathrm{d}^2\mathbf{r}(t)}{\mathrm{d}t^2}\Delta t^2 + \mathcal{O}(\Delta t^4) . \tag{2.5}$$

A variant of the Verlet integrator is the "leapfrog integrator" [51]. This scheme updates as well the coordinates as the velocities at among each other shifted times $t + \Delta t$ and $t + \frac{1}{2}\Delta t$:

$$\mathbf{v}(t+\frac{1}{2}\Delta t) = \mathbf{v}(t-\frac{1}{2}\Delta t) + \frac{\Delta t}{m}\mathbf{F}(t) , \tag{2.6}$$

$$\mathbf{r}(t+\Delta t) = \mathbf{r}(t) + \Delta t \cdot \mathbf{v}(t+\frac{1}{2}\Delta t) , \tag{2.7}$$

which results in the same trajectories for corresponding initial conditions as the original Verlet algorithm:

$$\mathbf{r}(t+\Delta t) = 2 \cdot \mathbf{r}(t) - \mathbf{r}(t-\Delta t) + \frac{1}{m}\mathbf{F}(t)\Delta t^2 + \mathcal{O}(\Delta t^4) . \tag{2.8}$$

The time step Δt needs in general to be chosen sufficiently small to resolve all interatomic motions. The fastest motions present in molecular systems are vibrational modes that oscillate with a cycle duration in the femtoseconds range (10^{-15} seconds). This makes the MD simulation scheme a challenging, inherently sequential, numerical procedure that has to overcome several orders of magnitude to reach relevant time scales for biological processes. Recent simulations on specialized hardware reach the millisecond regime [11] that is enough to study folding of small protein systems. For example, RNA folding has typical folding times in the seconds range and exceeds these limits by three orders of magnitude.

2.1.4. Temperature and Pressure Coupling

In the case of an uncoupled MD simulation the computational results would represent a microcanonical (NVE) ensemble. The experimental setup, however, usually resembles either a canonical (NVT) or an isobaric-isothermal ensemble (NpT). To realize such an ensemble one couples the simulations to a heat bath or a pressure reservoir [49]. The kinetic energy of a system is given by

$$E_{\mathrm{kin}} = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 \,, \tag{2.9}$$

where v_i are the particle velocities. The kinetic energy is connected to the temperature T by

$$\frac{1}{2} N_{\mathrm{DoF}} k_{\mathrm{B}} T = E_{\mathrm{kin}} \,, \tag{2.10}$$

where N_{DoF} denotes the number of degrees of freedom in the system. The number of degrees of freedom of a system with N particles is given by

$$N_{\mathrm{DoF}} = 3N - N_{\mathrm{c}} - 3 \,, \tag{2.11}$$

where N_{c} represents the number of position constraints in the system.

Berendsen temperature coupling introduces weak coupling to an external heat bath with temperature T_0 in first-order kinetics described by [52]

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau} \,, \tag{2.12}$$

where τ is the time constant of an exponentially decaying temperature deviation. The time-dependent scaling-factor for velocities at every n_{TC} steps is

$$\lambda = \left[1 + \frac{n_{\mathrm{TC}} \Delta T}{\tau_T} \left(\frac{T_0}{T(t - \frac{1}{2}\Delta t)} - 1 \right) \right]^2 \,, \tag{2.13}$$

where τ_T needs to be redefined. For a given constant τ in Eq. (2.12) and the heat capacity at constant volume C_V, τ_T is introduced by

$$\tau = \frac{2 \cdot C_V \cdot \tau_T}{N_{\mathrm{DoF}} \cdot k_{\mathrm{B}}} \tag{2.14}$$

in order to redistribute the energy changes due to velocity rescaling between kinetic *and* potential energy. This thermostat realizes a first order decay of temperature deviations without oscillations and generates an ensemble that differs from the canonical ensemble by an error that decays with the system size. Alternative algorithms to the Berendsen thermostat that reduce this error are the improved velocity rescaling method [53] or the Nosé-Hoover coupling algorithm [54].

For a formulation of pressure coupling the kinetic energy can be written as a tensor

$$\underline{\mathbf{E}}_{\mathrm{kin}} = \frac{1}{2}\sum_{i}^{N} m_i \mathbf{v}_i \otimes \mathbf{v}_i \tag{2.15}$$

that allows us to formulate the pressure tensor

$$\underline{\mathbf{p}} = \frac{2}{V}(\underline{\mathbf{E}}_{\mathrm{kin}} - \underline{\mathbf{\Xi}}) \,, \tag{2.16}$$

where V is the volume of the computational box and $\underline{\mathbf{\Xi}}$ is the virial tensor

$$\underline{\mathbf{\Xi}} = -\frac{1}{2}\sum_{i<j} \mathbf{r}_{ij} \otimes \mathbf{F}_{ij} \,. \tag{2.17}$$

In case of an isotropic system the scalar pressure p can be used for pressure coupling calculations and is calculated by

$$p = \mathrm{Tr}(\underline{\mathbf{p}})/3 \,. \tag{2.18}$$

The Berendsen pressure coupling [52] introduces a scaling matrix $\underline{\mathbf{s}}$ given by

$$s_{ij} = \delta_{ij} - \frac{n_{\mathrm{PC}}\Delta t}{3\tau_p}\beta_{ij}(p_{ij}^0 - p_{ij}(t)) \,, \tag{2.19}$$

which rescales the simulated volume of the system with isothermal compressibility β_{ij} every n_{PC} steps in order to adjust the actual pressure p_{ij} to the desired pressure p_{ij}^0. Furthermore, the *Kronecker* symbol is defined as

$$\delta_{ij} = \begin{cases} 1 & \text{for } i = j \,, \\ 0 & \text{otherwise} \,. \end{cases} \tag{2.20}$$

The coupling then is, as in the temperature coupling algorithm, introduced via a first-order decay relation

$$\frac{\mathrm{d}\underline{\mathbf{p}}}{\mathrm{d}t} = \frac{\underline{\mathbf{p}}_0 - \underline{\mathbf{p}}}{\tau_p} \,. \tag{2.21}$$

2.1.5. Stochastic Dynamics

Stochastic or velocity Langevin dynamics offers a possibility to introduce temperature coupling to the equation of motion [49]. The modified equation of motion reads as

$$m_i \frac{\mathrm{d}^2 \mathbf{r}_i}{\mathrm{d}t^2} = -m_i \phi \frac{\mathrm{d}\mathbf{r}_i}{\mathrm{d}t} + \mathbf{F}_i(\mathbf{r}) + \boldsymbol{\eta}_i(t) \,, \tag{2.22}$$

where ϕ is a friction parameter, and the noise term $\boldsymbol{\eta}_i(t)$ has a Gaussian probability distribution for which we get a correlation function following the fluctuation-dissipation theorem

$$\langle \eta_i(t)\eta_j(t')\rangle = 2m_i\phi \cdot k_\mathrm{B}T \cdot \delta_{ij}\delta(t-t')\,. \tag{2.23}$$

Therein, δ_{ij} is the *Kronecker* symbol as introduced in Eq. (2.20) and the respective function is defined as

$$\delta(t-t') = \begin{cases} 1 & \text{for } t = t'\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.24}$$

If $1/\phi$ is large compared to the used time step in a simulation, Langevin dynamics becomes effectively a standard MD formulation with stochastic temperature coupling. The time integration is performed with a modified leap-frog algorithm with third order accuracy in $\Delta t\phi$ [55].

2.1.6. Steered Molecular Dynamics

It is possible to introduce external forces to the simulations. They typically represent the numerical analogon to, e. g., single-molecule atomic force microscopy or optical trap experiments, or are a computational method to guide a system along a desired reaction coordinate. The latter technique is often referred to as umbrella sampling and drives the dynamics of a system to a predefined final state while conserving the possibility to calculate the free energy landscape based on the well defined external forces. Constant force pulling emulates the experimental setup of, e. g., experiments with optical traps [39] or atomic force microscopy [56]. The third possibility are constant velocity simulations where fixed interatomic constraints are increased with a constant rate. The standard method to introduce constraints in MD simulations is offered by the SHAKE algorithm [57]. This algorithm is a modified velocity Verlet method that imposes a limit on the acceptable deviations of a set of atom-atom distances.

2.2. Implementations

The implementation of MD software consist of two major parts: The compilation of suitable force fields as lists of experimentally determined or *de-novo* calculated interaction parameters and the collection of computational machinery that conducts simulations under user-defined conditions. Due to the huge amount of work behind such implementations the field is dominated by a few international research groups. Usually, these research groups provide both their own force field library and their MD software tools. The force field libraries are often complemented by modifications of theoretical groups that adjust the parametrizations to reproduce dynamical behavior of specific systems. This section gives a compact overview over existing implementations.

2.2.1. Force Fields

The oldest still actively maintained and most established force fields are AMBER [58] and CHARMM [59]. A newer derivative of CHARMM is the GROMOS force field [60]. A more recent force field is the coarse-grained MARTINI force field [61].

The usual workflow in force field development starts from protein parametrizations that are then adapted towards nucleic acids (RNA, DNA), lipids, and other organic compounds. Due to the multitude of possible choices for formulations of the parametrization the actual parameter values are usually not comparable between different force fields.

2.2.2. Molecular Dynamics Software

There exists a wide range of MD simulation software that cover multi-purpose suites and more specialized tools. Almost all of them have in common that they are architecture independent, which enables the users to run them as well on their desktop PC for prototyping as on high performance computing facilities for productive runs. The used file formats, albeit there exists no general standard, are usually interconvertible. This allows the combination of workflow steps between tools in different software suites. While some of them are developed together with a distinct force field they usually offer interfaces to alternative force field definitions which fosters the comparability of parametrizations.

The force fields AMBER and CHARMM provide their own MD simulation software packages (AMBER [62], CHARMM [63]). AMBER exhibits a modular concept with a suit of various tools, and CHARMM provides a more holistic approach combining all functionalities in a compact single application. Both implementations are tailored for the use on high performance computing facilities due to the general demand of MD simulations. A stand-alone software package without in-house force field development is the NAMD package [64] that offers its own well integrated visualization tool VMD [65]. NAMD is also highly optimized for high performance computation resources. GROMACS [66] is a very customizable software package that is fully released under open source, creative commons license. It allows a very flexible adjustment of force field definitions and is therefore best suited for modified formulations of the MD approach. It also offers optimizations for the use on desktop PCs, which makes it an excellent choice especially for simplistic biomolecular folding models, as introduced in Sec. 3.3 in the next chapter.

3 Chapter 3. Simplistic Models

As shown in the previous chapter, the applicability of standard molecular dynamics simulations is limited to time scales in the order of milliseconds. This chapter introduces simplistic models that are capable of reproducing the dynamics of biomolecules with reduced computational effort.

First, I motivate the assumption of a funneled energy landscape for biopolymer chains. The derivations made from the thermodynamics and kinetics of such systems lead to a discussion of biopolymer models in analogy to spin glass theory.

The next section covers concrete examples of simplistic models in the context of native structure-based simulations. First, the hydrophobic-polar (HP) model is introduced as a simple lattice model. The model introduces a highly frustrated energy landscape based on hydrophilic-hydrophobic interactions. An extension to the model, the HP+ model, grants energetic benefits for the formation of native contacts. The incorporation of native contact information reduces the frustration and creates a funnel-shaped energy landscape.

As a further step, an off-lattice model, in this thesis regarded as the *native (tertiary) structure-based model (SBM), is presented and discussed. Its potential energy has its minimum at the native conformation resulting in an overall funneled free energy landscape biased towards the native state. The introduced energetics with an all-(heavy)atom resolution are homogeneous for all atom types. The dynamics simulations in this model can be realized in a standard molecular dynamics integration scheme. Originating from protein folding models SBMs are motivated for RNA systems. This model will be the integral tool for tackling biophysical challenges throughout my studies presented in this thesis.*

In the last section, a native secondary structure-based model for RNA systems in a Monte Carlo implementation is shown. The kinetic Monte Carlo model features empirical energetics for the opening and closing of base pairs. This method is used as a complementary model for comparisons to some of my results in Chap. 6.

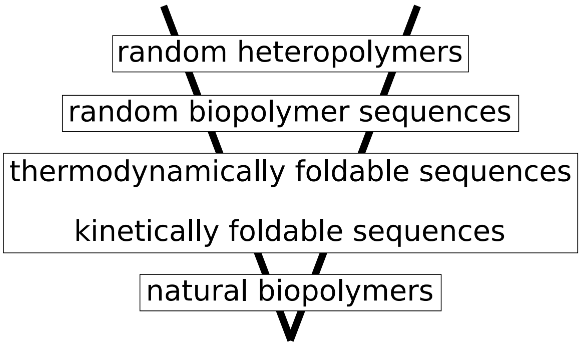

random heteropolymers
random biopolymer sequences
thermodynamically foldable sequences
kinetically foldable sequences
natural biopolymers

Figure 3.1.: Schematic representation of the number of sequences sorted by polymer classifications in the context of foldability. The class of random heteropolymers comprises the countable infinite number of random biopolymer sequences. A subclass of the mass of hypothetically constructible biopolymers are the sequences that are thermodynamically and kinetically foldable. This class determines the range for biomolecular design and contains the remaining, comparable small class of naturally occurring biopolymers. This section tries to shed light on the theoretical characterization of the class of thermodynamically and kinetically foldable sequences which allows then to model natural biopolymers.

3.1. Biopolymer Folding Thermodynamics and Kinetics

Biomolecules such as proteins and nucleic acids (RNA or DNA) fulfill myriads of important tasks that create and maintain life in cells of all organisms. These biomolecular polymers are sequences of distinct building blocks that fold into characteristic conformations depending on biophysical interactions, e. g., hydrophobic forces, hydrogen bonds, or more distinctively discussed RNA interactions in Sec. 1.2.2. Their actual functionality is determined by their structure and less unambiguously by their sequence. The connection between structure and function is thereby an established dogma of biomolecular sciences. In the context of this dogma and in combination with the rapidly growing number of measured biomolecular sequences [67, 68] it is very desirable to derive structure from biopolymer sequence alone. A formal description of the folding process can be found in the theory of polymer physics. In the context of heteropolymers the general challenge of biopolymers is their apparent thermodynamic and kinetic foldability, as indicated in Fig. 3.1. The theoretical approach needs, therefore, to reproduce basic thermodynamic and kinetic features of biopolymer folding behavior.

A simple but compelling estimation for the folding time of a biopolymer was performed by Levinthal who argued that the folding time of a rather small protein would exceed the lifetime of the universe [27] – Levinthal's paradox. In combination with direct measurements of protein folding times, at that time by Anfinsen [69], the resolution of this paradox was motivated in a quantitative manner. Levinthal himself proposed the following idea: Since the number of possible conformations in biomolecules times the typical dwell time exceeds the realistic folding times of polymers, folding needs to follow a folding pathway along distinct intermediate states that was formed under evolutionary pressure. The folding pathway picture by Levinthal was replaced by a more refined perspective of biopolymer folding that is based on energy landscape theory [70]. Therein, an energy landscape is defined as the Helmholtz free energy as a function of the conformational space of the biopolymer. This quantitative theoretical approach follows the lines of spin glass theory in combination with a random energy model (REM) [71, 72]. The apparent experimentally determined folding characteristics concur with the perspective of a funneled energy landscapes for proteins [28–30]. The theoretical development over the last three decades has characterized the biophysical view of biopolymer folding and has motivated the justification of native structure-based models [12–17].

The following derivation describes a statistical approach to protein folding and introduces the free energy and entropy of a system by means of statistical mechanics [12, 13]. By the choice of an order parameter, such as the similarity ratio to the native structure Q, it is possible to parametrize the energy landscape of a biopolymer and introduce strata with average energies $\bar{E}(Q)$. The central limit theorem for the energy distribution within a stratum results in:

$$P(E) = \frac{1}{\sqrt{2\pi\Delta E(Q)^2}} \cdot \exp\left(-\frac{\left(E - \bar{E}(Q)\right)^2}{2\Delta E(Q)^2} \right), \tag{3.1}$$

where ΔE is the variance (or the “roughness” of the energy landscape within a stratum), $\bar{E}(Q)$ is the mean of the distribution and Q is the similarity ratio to the native structure ranging from 0 for the completely unfolded polymer to 1 for the native folded state. The number of possible conformations is given by

$$\Omega = \gamma^N, \tag{3.2}$$

where γ is the number of possible states that can be occupied by a sequence unit (or “residue”) and N is the sequence length. Depending on the structural accuracy of the model γ may vary between 2 and more than 10. For a given nativeness Q of the polymer state the number of conformations reads

$$\Omega(Q) = \gamma^{\star N(1-Q)}, \tag{3.3}$$

where $\gamma^\star$ indicates an optional foreknowledge that reduces the number of possible states that can be occupied. The number of possible conformations depends strongly on the nativeness and reaches 1 in the case of a completely folded polymer. From this we can calculate the entropy of the biopolymer

$$S_0(Q) = k_\mathrm{B} \log(\Omega(Q)). \tag{3.4}$$

This formulation of the entropy reaches 0 for nativeness 1 in which case the system is supposed to occupy a defined, single state.

We introduce an energy dependence to the number of accessible conformations by multiplying the energy probability in Eq. (3.1) with the number of conformations in Eq. (3.3):

$$\Omega(E, Q) = P(E) \cdot \Omega(Q). \tag{3.5}$$

The energy dependence arises from the fact that for a given temperature T not all high energy states can be occupied. As a next step, we can derive the probability density at thermal equilibrium to find a polymer with energy E and the nativeness Q

$$p(E,Q) = \frac{1}{Z} \cdot \Omega(E,Q) \cdot e^{-E/k_\mathrm{B}T}\,, \tag{3.6}$$

where Z is the partition function that normalizes the probability function and k_B is Boltzmann's constant. By maximizing this expression we gain the most probable energy at

$$\hat{E}(Q) = \bar{E}(Q) - \frac{\Delta E(Q)^2}{k_\mathrm{B}T} \tag{3.7}$$

and following Eq. (3.5) we gain the number of occupied states as

$$\Omega\left(\hat{E}(Q),Q\right) \overset{(3.4)(3.1)}{=} \exp\left(\frac{S_0(Q)}{k_\mathrm{B}} - \frac{\Delta E(Q)^2}{2(k_\mathrm{B}T)^2}\right)\,. \tag{3.8}$$

Following Eq. (3.4) the entropy of the most probable energy state at a given nativeness becomes

$$S\left(\hat{E}(Q),Q\right) = S_0(Q) - \frac{\Delta E(Q)^2}{2k_\mathrm{B}T^2}\,. \tag{3.9}$$

The most probable energy $\hat{E}(Q)$ and the entropy of the system $S\left(\hat{E}(Q),Q\right)$ are counteracting thermodynamic observables that dominate the folding process. The maximization of entropy is linked to the tendency to increase the number of disordered configurations. The minimization of energy demands a decrease in the conformational degrees of freedom of the system. This motivates the consideration of the free energy at given temperature T and nativeness Q

$$\begin{aligned} F(Q) &= \hat{E}(Q) - TS\left(\hat{E}(Q),Q\right) \\ &= \bar{E}(Q) - \frac{\Delta E(Q)^2}{2k_\mathrm{B}T} - TS_0(Q)\,. \end{aligned} \tag{3.10}$$

For high temperatures, the free energy exhibits a single minimum at a small nativeness (close to 0), whereas for low temperatures it features a single minimum at a large nativeness (close to 1). In the intermediate temperature range the free energy has two minima that have equal thermodynamic weight at the folding temperature. These characteristics represent a two-state folding behavior.

The theoretical approach to biopolymer folding kinetics exhibits four crucial differences in contrast to standard transition state theory [73]:

1. The influence of a surrounding solvent has to be taken into account.
2. Entropic effects are important due to the conformational uncertainty of the intermediate and unfolded states.
3. The choice of a reaction coordinate is less clear.

4. The effective diffusion coefficient may dependent drastically on the reaction coordinate.

The gradient of the free energy function represents the tendency of the system to change its nativeness Q. The direction is determined by the general aim to minimize the free energy. Transition state theory describes a process by overcoming a transition state that prevents a direct conversion from an initial state towards the final state. The barrier height of this transition state corresponds to a transition rate that is characterized by its diffusion constant. If $\bar{t}(Q)$ stands for the average lifetime of a microstate in the system of interest the bottleneck can be found at the nativeness $\hat{Q}_{\mathrm{kin}}$ that maximizes the lifetime. This lifetime can be identified with the folding time t_{f}:

$$t_{\mathrm{f}} = \bar{t}(\hat{Q}_{\mathrm{kin}}) \,. \tag{3.11}$$

The roughness of a energy landscape $\Delta E(Q)$ influences the lifetime of its microstates. The lifetime is found to follow a Ferry law [74] for sufficiently high temperatures and reads:

$$\bar{t}(Q) = t_0 \cdot e^{(\Delta E(Q)/k_{\mathrm{B}}T)^2} \,. \tag{3.12}$$

Below the "glass transition" temperature

$$T_{\mathrm{g}} = \left(\frac{\Delta E(Q)^2}{2k_{\mathrm{B}} S_0(Q)} \right)^{1/2} \,, \tag{3.13}$$

at which the system runs out of entropy due to the lack of alternate conformation states, the lifetime becomes

$$\bar{t} = t_0 \cdot e^{S_0(Q)/k_{\mathrm{B}}} \,, \tag{3.14}$$

which resembles a "search" time in Levinthal's original point of view. Therefore, the relation between folding temperature and glass transition temperature determines whether a polymer will be able to fold in limited time or not.

Energy landscape theory has been discussed as an analytical description of biopolymer folding. Levinthal's paradox that assumes a completely flat energy surface with a single hole that represents the native conformation can be resolved by the concept of cooperativity that is observed in biomolecular systems. Cooperative phase transitions, similar to crystallization processes, can be understood as a funnel-shaped surface in the energy landscape. Due to the huge number of conformational states under geometric constraints, this surface is expected to be frustrated (in analogy to spin glasses) and rough. The experimental observations of protein folding give rise to the assumption that evolutionary pressure formed the energy landscapes of reliably folding polymers to be smooth or the underlying interaction network to be minimally frustrated. The theoretical investigation of these systems identifies as well thermodynamic as kinetic aspects involved in the assessment of the foldability of a biopolymer. The principle of minimal frustration introduces a suitable gradient in the free energy landscape that creates a variety of folding paths towards the native state. Without glass transition the system is able to achieve the necessary entropy loss. In other words: The free energy profile can allow folding at limited time scales but in case of a glass transition before sufficient folding, the folding channels will be kinetically inaccessible.

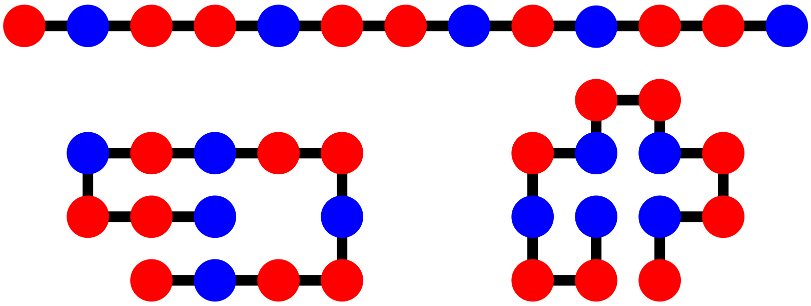


Figure 3.2.: Schematic illustration of the HP model (sequence and conformations taken from [31]). Top: A modeled protein sequence of H (blue) and P (red) building blocks represents the hydrophobic (H) and hydrophilic (P) sidechains. H blocks have the tendency to form contacts with each other in order to achieve a compact fold. Bottom: H-H contacts gain an energetic benefit compared to H-P or P-P contacts. The conformation on the left is kinetically trapped since it has to break two energetically favored contacts to proceed towards the native conformation. The conformation on the right represents the native fold at the energetic minimum of the system with five H-H contacts.

3.2. Hydrophobic-polar Protein Folding Model

The hydrophobic-polar protein folding (HP) model introduces biophysical interactions between the surrounding solvent (usually water-based solutions) and the hydrophobic (H) and polar or hydrophilic (P) side chains of proteins in a simplistic lattice model [31]. The system is represented by a rectangular lattice that can accommodate a sequence of H or P building blocks, as it is shown in Fig. 3.2 by a two-dimensional representation of the model. The energetic scoring function features a negative energy for each H-H contact and zero energy for H-P or P-P contacts. An energy minimization, therefore, maximizes the number of contacts of H blocks with each other. By means of these assumptions it is possible to investigate the kinetics of biomolecular folding directly and reproduce characteristic experimental plots like melting curves and Chevron plots [31]. This model also facilitates the investigation of influences of native interactions on the folding process: An extended formulation, the HP+ model, incorporates native interactions as additional contact energy scores that promote the native conformation. Thereby, kinetic traps in the folding process that are present in the original HP model are reduced, the corresponding funneled energy landscape of the sequence is smoothened and frustration is minimized. These findings – labeled with "principle of minimal frustration" [14, 29] – motivated the derivation of simplistic models that implement minimally frustrated funneled energy landscapes. In the following section structure-based models (SBMs) as off-lattice examples that follow the principle of minimal frustration are introduced.

3.3. Native Structure-based Model

The principle of minimal frustration [14, 29] can also be introduced in off-lattice, atomistic models. They are deducted from thermodynamic models of protein folding and introduce native interaction networks to a simplistic energy parametrization in order to create a funnel-shaped, minimally frustrated free energy landscape with the native conformation as its minimum. These models are referred to as Gō-like models or native structure-based models (SBMs) [12–17]. SBMs are designed to remove thermodynamic and kinetic traps and to smoothen out the energy landscape without abandoning global characteristics. SBMs have been validated by comparison with experimental measurements [75] in their originally developed context of protein folding simulations. The application of SBMs in the field of structured RNA folding is justified due to observations that shed light on the existance of native interaction networks that dominate the respective folding processes [76–78]. Accordingly, there are quite recent studies of RNA folding by means of SBM simulations [22, 23, 79]. A comprehensive overview over the current status of SBM techniques in their wide range of applications has been recently published [80].

There exist a web-based implementation [18] and a local implementation [19] that generate simulation input files for SBM simulations. SBM simulations can then be performed by standard MD software packages that allow a flexible modification of the used force field simulation. GROMACS [66] is such a software suite and its interplay with an SBM implementation is presented in Chap. 5.

3.3.1. Native Contact Information

The native contact information that needs to be incorporated in the SBM is determined by a variety of definitions, two of which are a simple cut-off contact map or a “shadow map” [81]. The cut-off definition regards all atom-atom pairs that are closer to each other than a certain distance threshold as contacts. The shadow map takes, in addition to an upper limit for the considered atom distances, a possible shadowing of inter-atomic contacts by closer atoms in between into account. Two representations of a standard cut-off contact map are depicted in Fig. 3.3. The natural reaction coordinate in the context of SBMs is the Q value – the total or normalized number of formed native contacts. The applicability of the Q value as a reaction coordinate for protein folding simulations has been recently investigated and reported in detail [82].

3.3.2. Potential Energy

The actual implementation of SBMs depends on the functional form and the parametrization of the corresponding potentials. In principle there exist various formulations of which two examples are given to illustrate the general approach: an all-(heavy)atom formulation and a possible coarse-grained formulation of the potential energy.

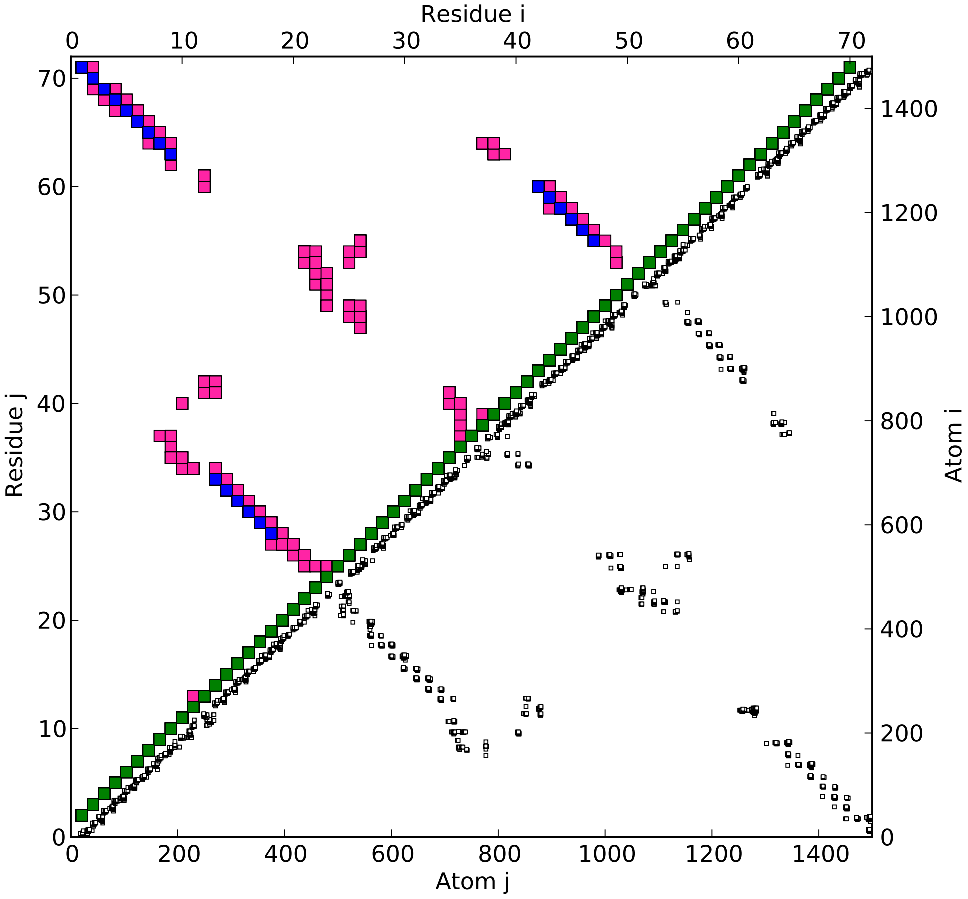


Figure 3.3.: Two possible contact map representations of an *add* adenine riboswitch (PDB ID 1Y26 [21]). In the lower right half of the diagram each atom-atom contact is marked which yields a *contact map*. The upper left half of the diagram depicts a clearer view of the contact map. The contacts are reduced to residue-residue contacts and color coded as stacking contacts along the sequence (green), helical base pairing contacts (blue), as discussed in Sec. 1.2.2 and general contacts highlighted in magenta.

All-atom

An all-atom formulation of an SBM potential [22, 83] features most of the standard terms in an MD potential, as discussed in Sec. 2.1:

$$
\begin{aligned}
V(\{r,\theta,\chi,\phi,r_{ij}\}) = & \sum_{\text{bonds}} K_{\mathrm{b}}(r-r_0)^2 \\
& + \sum_{\text{angles}} K_{\mathrm{a}}(\theta-\theta_0)^2 \\
& + \sum_{\text{dihedrals, impropers}} K_{\mathrm{i}}(\chi-\chi_0)^2 \\
& + \sum_{\text{dihedrals, backbone}} K_{\mathrm{d}}^{(\mathrm{bb})}\left[[1-\cos(\phi-\phi_0)]+\frac{1}{2}\left[1-\cos(3\cdot(\phi-\phi_0))\right]\right] \\
& + \sum_{\text{dihedrals, sidechain}} K_{\mathrm{d}}^{(\mathrm{sc})}\left[[1-\cos(\phi-\phi_0)]+\frac{1}{2}\left[1-\cos(3\cdot(\phi-\phi_0))\right]\right] \\
& + \sum_{\text{contacts}} K_{\mathrm{c}}\left[\left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{12}-2\cdot\left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{6}\right] \\
& + \sum_{\text{non-native contacts}} K_{\mathrm{nc}}\left(\frac{\tilde{\sigma}}{r_{ij}}\right)^{12} .
\end{aligned}
\tag{3.15}
$$

In this potential the zero indexed quantities $r_0, \theta_0, \chi_0, \phi_0, \sigma_{ij}^0$ represent the minimum of the funneled energy landscape that is equal to the native conformation and $\tilde{\sigma}$ is a global exclusion radius. Accordingly, the potential has its minimum at the native conformation. The bonded interactions – bonds, angles and dihedral (or torsional) angles – are represented in an usual fashion: Bonds, angles and improper dihedral angles are modeled in a harmonic approximation, as seen in Fig. 3.4. The proper torsional angles are represented by a periodic potential with a global minimum and two local minima (Fig. 3.5). Key to the SBM potential are the native contacts that are introduced via Lennard-Jones potential terms between two non-bonded atoms (Fig. 3.6). The contact information comprises an implicit modeling of the otherwise neglected electrostatic interactions and involved solvents. All non-native atom pairs are represented by repulsive terms.

The force constants K_{b},K_{a},K_{i},K_{d},K_{c},K_{nc} are homogeneous in the standard formulation of SBMs, i. e., they do not depend on atom or residue types. Possible modifications are discussed in the description of a software implementation that I present in Chap. 5. The values of the force constants are cataloged for bonds, angles and improper dihedral angles. For proper dihedral angles and contacts, a ratio between the total energy available in proper dihedral angles and contacts is introduced. The sum of total proper dihedral energy and contact energy is normalized to the number of atoms in the system. A more detailed description of the still homogeneous but relatively to each other defined force contacts for these two interactions is given in Sec. A.1. As a result, the involved time and temperature scales are given in unphysical, reduced units. In order to derive natural time and temperature

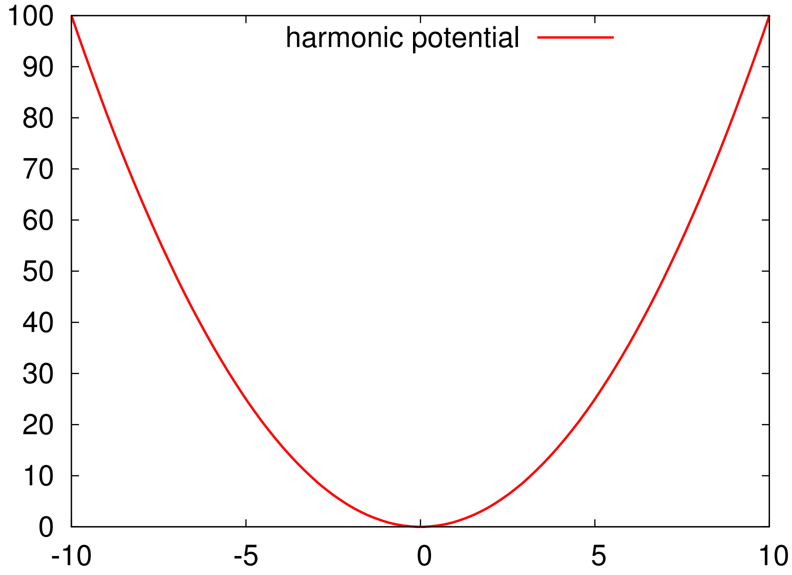


Figure 3.4.: Visualization of a harmonic potential. The standard representation of bonded interactions in molecular dynamics simulations is the harmonic approximation. It follows a Hookean force law for displacements out of the position of rest.

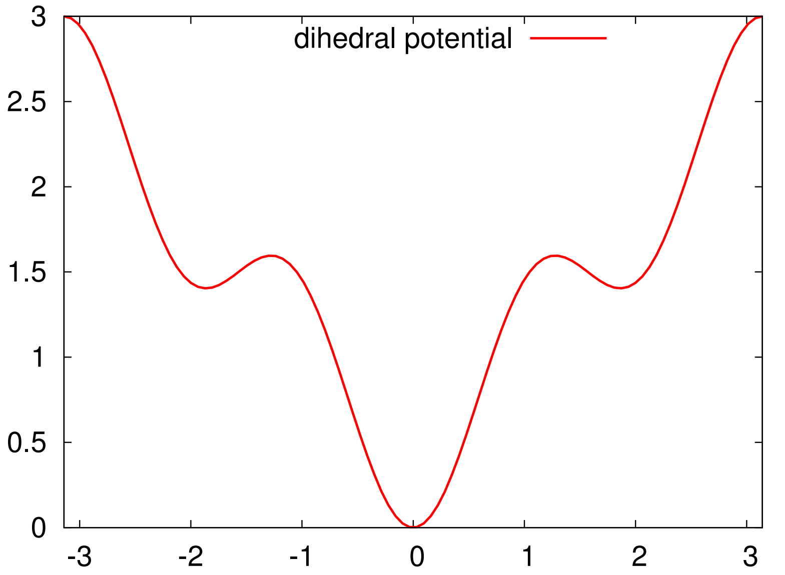


Figure 3.5.: Visualization of a dihedral angle potential. The dihedral angle potential is a 2π-periodic potential that has a global minimum at the native value (here 0) and two local minima that allow the structure to occupy isomeric conformations.

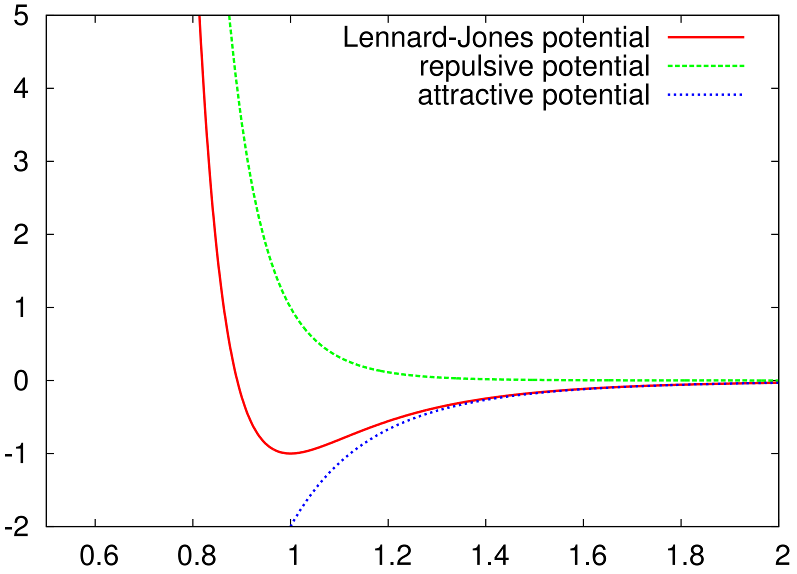


Figure 3.6.: Visualization of a Lennard-Jones Potential. The potential is the sum of a repulsive $1/r^{12}$ term and an attractive $1/r^6$ term. At distance 1 the potential well has its minimum at a depth of -1 in this example.

scales from SBM simulations they need to be introduced via comparisons of characteristic observables. Two strategies to introduce physical units to a system of interest are presented in Sec. 6.2.5.

C$_\alpha$ Coarse Graining

The second example introduces the concept of coarse-graining into the framework of SBM potential construction. The complexity of a system is reduced by aggregating the amino acids of a protein into single beads at the positions of their C$_\alpha$ atoms. The respective potential energy function can be formulated as [75]

$$\begin{aligned}V(\{r,\theta,\chi,\phi,r_{ij}\}) = &\sum_{\text{bonds}} K_{\text{b}}(r-r_0)^2 \\ &+ \sum_{\text{angles}} K_{\text{a}}(\theta-\theta_0)^2 \\ &+ \sum_{\text{dihedrals}} K_{\text{d}}\left[[1-\cos(\phi-\phi_0)] + \frac{1}{2}\left[1-\cos(3\cdot(\phi-\phi_0))\right]\right] \\ &+ \sum_{\text{contacts}} K_{\text{c}}\left[5\cdot\left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{12} - 6\cdot\left(\frac{\sigma_{ij}^0}{r_{ij}}\right)^{10}\right] \\ &+ \sum_{\text{non-native contacts}} K_{\text{nc}}\left(\frac{\tilde{\sigma}}{r_{ij}}\right)^{12} . \qquad (3.16)\end{aligned}$$

In this particular formulation the regular Lennard-Jones potential terms are modified to a 10-12 form. This empirical modification models the screening effects of the remaining atoms in a amino acid around the C_α atom.

3.4. Kinetic Monte Carlo Method

A complementary native secondary structure-based approach is the kinetic Monte Carlo (MC) method that has recently been presented for RNA folding simulations [32]. In contrast to the SBM approach the kinetic MC method uses an empirical energy parametrization but yields no atomistically resolved trajectories.

RNA is represented by a sequence of bases $b_1, ..., b_N$ where $b_i =$ A, C, G or U (see Chap. 1). A set of base pairs (b_i, b_j) defines the secondary structure of a given RNA sequence. Similar to native tertiary structure-based approaches native structural information is incorporated into the RNA model. A list of contacts that are closed in the native secondary structure is provided and the base pairing interactions restricted to those listed. The total free energy is then approximated by the sum of all structural motifs

$$G_{\text{tot}} = \sum_{\text{all base pairs}} G_{\text{base pair}} + \sum_{\text{all loops}} G_{\text{loop}} . \qquad (3.17)$$

In general, the formation of a base pair is energetically preferred ($\Delta G_{\text{base pair}} > 0$) and closing a loop is penalized ($\Delta G_{\text{loop}} < 0$) due to the entropic cost. The free energy benefits are calculated based on empirical models and parameters that are established in the field of RNA secondary structure prediction [84]. The closure of base pairs is parametrized depending on potential base stacking or single base mismatches. The formation of a hairpin loop,e. g., is parametrized by

$$G_{\text{hairpin loop}}(l,(b_i,b_{i+l+1})) = G_{\text{init}}(l) + G_{\text{base pair}}(b_i,b_j) + G_{\text{oligo C}}(l) , \qquad (3.18)$$

where the loop of length l is positioned between bases b_i and b_{i+l+1}, $G_{\text{init}}(l)$ are cataloged values and $G_{\text{oligo C}}(l)$ is a potential penalty for pure cytosine loops. Similarly, there exist parametrizations in combination with cataloged values for short loops, bulges and non-local loops connecting multiple helices.

A Monte Carlo simulation scheme based on Metropolis rates is used to introduce dynamics. The basic moves are the random closing and opening of single native base pairs within the sequence. A move is accepted with a probability that is determined by the Boltzmann weight function

$$p = \exp\left(-\Delta G/k_B T\right) \, , \tag{3.19}$$

where ΔG is the free energy difference based on secondary structure formation before and after the proposed move. In case of $\Delta G \leq 0$, moves are always accepted.

4 Chapter 4.

Coevolutionary Statistical Analysis

The connection between structure and function of biopolymer chains and the structural similarity of chains with high sequence identity ("homology") is the consequence of evolutionary pressure towards a desired function. Mutations within a class or "family" of homologous chains, therefore, need to be compensated to guarantee structural integrity. Experimental techniques allow the recording of genetic information for tens of thousands of different species in a family while the structural resolution of biological systems is a very demanding task. Sequence information can be filtered for specific functional elements and compiled in multiple sequence alignments (MSA) for such elements, available in open databases. Coevolutionary methods are able to detect correlated mutations within such alignments and give contact predictions for sequence elements in their chains. Contact investigations based on coevolutionary algorithms aim at structure predictions which can ultimately shed light on the function of biomolecular systems.

First, the basic concepts of sequence analysis by statistical methods are introduced. Mutual information as a measure of correlation in coevolutionary methods is introduced. The relevance of mutual information for coevolutionary statistical analyses is discussed and its importance in the context of RNA is highlighted. Mutual information is able to predict secondary structural elements very reliably but fails to do so for general tertiary contacts.

The next section discusses direct coupling analysis (DCA) as an improved approach to disentangle direct and indirect correlations between two residues. Since indirect evolutionary coupling is a common motif but obscures the prediction of spatial contacts, DCA is able to yield improved predictions. A Potts model represents an ansatz for a probabilistic model that maximizes entropy while satisfying the condition that the model depicts empirically observed frequency counts of an MSA. Solutions to this problem can be found by assuming small couplings that lead to a mean-field approximation. From the probabilistic model a gauge independent score, the direct information (DI), can be derived that ranks the list of predicted site pairs.

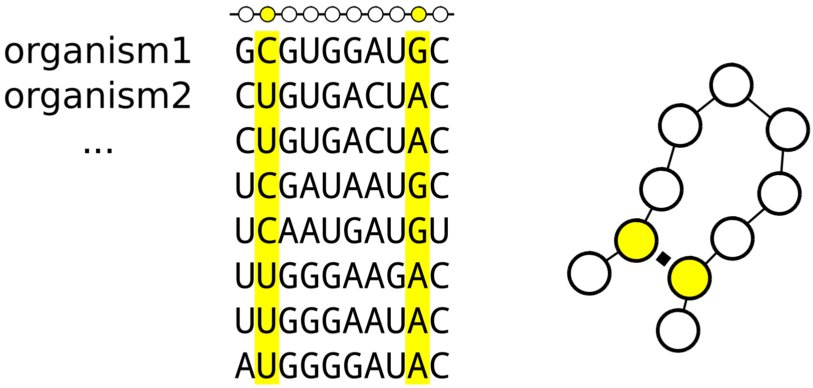


Figure 4.1.: Exemplary MSA of RNA and the connection between correlation and spatial closeness. Commutating columns in the alignment are highlighted in yellow. The example illustrates directly coupled mutations in contrast to indirectly coupled mutations that influence each other via one or more mediating mutation steps in between. The direct correlation can be indicative of spatial closeness of the two involved sequence elements.

4.1. Basic Concepts

Coevolutionary methods are based on multiple sequence alignments (MSAs) that contain sequence information of a given characteristic sequence motif that is found in many different organisms. The collective of a characteristic sequence motif is called a “family” and its member sequences are aligned to each other. Data sets of these families are stored in databases, such as the RFAM sequence database [68]. A fictive multiple sequence alignment of RNA and the implication of commutating columns on spatial closeness are depicted in Fig. 4.1: Destabilizing mutations within sequence contacts in a protein or an RNA strand are supposed to be compensated by commutations in the course of evolution in order to conserve functional structure elements. Therefore, directly correlated mutations can be linked to spatial closeness in biopolymers. At the same time, sequence databases are growing rapidly due to improved automated sequencing techniques [67, 68]. A striking idea in this context is to develop or enhance biomolecular structure prediction protocols based on easily accessible sequence information. These protocols are based on algorithms in the field of statistical methods that discover mutation correlations and with them spatial contacts [85, 86]. This section introduces the formal mathematical framework that is necessary to describe the methods, as it can be found in [25]. The basic idea of mutual information approaches and their applications and limits in the context of RNA structure predictions are discussed. In the next section, a more advanced method, the direct coupling analysis (DCA), that has been successfully employed for protein structure prediction is described.

4.1.1. Multiple Sequence Alignments and Frequency Counts

RNA families can be represented in databases by a multiple sequence alignment (MSA). The various representatives of a given family are depicted as lines containing sequences in a standard one-letter code for nucleotides or amino acids. These sequences are aligned so that sequence identity is maximized and the over-all consensus secondary structure is conserved as much as possible. There exist various algorithms that demand these conditions: Needleman-Wunsch [87] and Smith-Waterman [88] algorithms are established dynamic programming techniques to align sequences. Both algorithms evaluate similarity matrices that allow the insertion of gaps and they differ from each other by the employed gap penalty scheme, which makes the Smith-Waterman algorithm favor local alignments compared to the Needleman-Wunsch algorithm. Consensus secondary structures can be determined by a variety of methods based on dynamic programming approaches, such as the Nussinov algorithm [89], or statistical correlation approaches [90, 91].

In order to introduce a mathematical formalism to analyze the statistical properties of an alignment, a matrix $\underline{\mathrm{A}}$ that represents an MSA can be defined. $\underline{\mathrm{A}}$ is filled with entries $1, 2, \ldots, q$ representing the building blocks of sequences where q is the size of the "alphabet" of the biomolecular system, e. g., 5 for RNA (4 nucleotides G, C, A, U and 1 gap) or 21 for proteins (20 amino acids and 1 gap). The shape of $\underline{\mathrm{A}}$ is given by

$$\underline{\mathrm{A}} = (A_i^a), \qquad i = 1, \ldots, L, \quad a = 1, \ldots, M\,, \tag{4.1}$$

where L is the number of columns that represent the residues of the MSA (length of aligned biopolymer sequences), and M is the number of rows in the MSA (number of biopolymer representatives in the aligned family).

We define the single-site and pair frequency counts as

$$f_i(A) := \frac{1}{M} \sum_{a=1}^{M} \delta_{A,A_i^a}\,, \tag{4.2}$$

$$f_{ij}(A,B) := \frac{1}{M} \sum_{a=1}^{M} \delta_{A,A_i^a} \delta_{B,A_i^a}\,, \tag{4.3}$$

where $1 \leq i,j \leq L, 1 \leq A,B \leq q$ and δ represents the *Kronecker* symbol in Eq. (2.20). Frequency counts are able to represent the statistical properties of the MSA if they are based on independent samples. The usual content of sequence databases, however, has a heavy sampling bias. This bias is caused by the general problem of phylogenetic relations between species and by the choice of sampled species themselves due to research focuses. The influence of the similarities in the aligned sequences can be corrected by a reweighting scheme [92]. For a given sequence $A^a = (A_1^a, \ldots, A_L^a)$, the number of similar sequences $A^b = (A_1^b, \ldots, A_L^b)$ can be defined by

$$m^a := \left| \{ b \mid 1 \leq b \leq M, \mathrm{seqid}(A^a, A^b) \geq xL \} \right|\,, \tag{4.4}$$

where x is the the similarity ratio of a sequence with length L. From this we can deduce the effective number of independent sequences

$$M_{\mathrm{eff}} = \sum_{a=1}^{M} \frac{1}{m^a}\,. \tag{4.5}$$

Single-site and pair frequency counts can be redefined accordingly by

$$f_i(A) := \frac{1}{\lambda + M_{\text{eff}}} \left(\frac{\lambda}{q} + \sum_{a=1}^{M} \delta_{A,A_i^a} \right) , \tag{4.6}$$

$$f_{ij}(A,B) := \frac{1}{\lambda + M_{\text{eff}}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^{M} \delta_{A,A_i^a} \delta_{B,A_i^a} \right) , \tag{4.7}$$

where we introduce the pseudo-count λ [93]. The updated frequency counts compensate for insufficiently equally distributed sampling in the data set. The pseudo-count λ is usually chosen in the order of M_{eff} and can therefore be defined by

$$\lambda := l \cdot M_{\text{eff}}, , \tag{4.8}$$

where $l \approx 1$ [25].

4.1.2. Mutual Information

A standard measure of mutual dependence between two MSA columns is the mutual information MI_{ij} defined as

$$\text{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \left(\frac{f_{ij}(A,B)}{f_i(A) f_j(B)} \right) . \tag{4.9}$$

Mutual information equals zero if and only if columns i and j are independent of each other, which corresponds to $f_{ij}(A,B) = f_i(A) \cdot f_j(B) \quad \forall A,B$, and is positive otherwise. Mutual information is not able to distinguish between indirect and direct correlation and is therefore applicable only to a limited extent for detecting spatial contacts. Correlations could be mediated indirectly via one or more mutating building blocks that would give an MI signal but are not in the spatial vicinity of each other. The signal in RNA alignments for secondary structure contacts (Watson-Crick or Wobble base pairs) is sufficiently prominent to use mutual information as prediction score [90, 91]. Tertiary contacts, i. e., all non-local contacts apart from canonical base pairing, have a low signal that is indistinguishable from underground noise. Therefore it is necessary to disentangle direct from indirect correlations in order to raise the signal of direct tertiary contacts above the noise level.

4.2. Direct Coupling Analysis

To overcome the shortcoming of not being able to distinguish between direct and indirect inter-column correlations in the MSA, the direct coupling analysis (DCA) disentangles these two kinds of correlations [25]. DCA is an implementation of inverse statistical mechanics that describes a formalism that is able to derive model parameters (in equivalent terms: fields and couplings) from observables (in equivalent terms: magnetizations, order parameters, empirical samples, etc.). The maximum entropy *ansatz* is a concrete approach to solve this q-state Potts model, where q is the size of the sequence alphabet. In this section, the maximum-entropy model is introduced and solved under independent-site and mean-field approximations, as it can be found in [25]. Direct information (DI) can then be calculated as the new improved scalar score for spatial contact predictions in biopolymers.

4.2.1. Maximum-entropy Ansatz

In contrast to simple mutual information calculations direct couplings can be discovered by the more demanding and involved procedure of inferring a statistical model $P(A_1, \dots, A_L)$ for all biopolymer sequences in a family's MSA. We require a global model that can reproduce the empirical single site $f_i(A_i)$ and pair frequency counts $f_i(A_i, A_j)$:

$$P_i(A_i) = \sum_{\{A_k|k \neq i\}} P(A_1, \dots, A_L) \overset{!}{=} f_i(A_i) , \tag{4.10}$$

$$P_{ij}(A_i, A_j) = \sum_{\{A_k|k \neq i,j\}} P(A_1, \dots, A_L) \overset{!}{=} f_i(A_i, A_j) . \tag{4.11}$$

In addition, the model that satisfies Eqs. (4.10) and (4.11) should be most general and therefore least constrained, which can be determined by maximizing the entropy

$$S = - \sum_{\{A_i|i=1,\dots,L\}} P(A_1, \dots, A_L) \ln \left(P(A_1, \dots, A_L) \right) . \tag{4.12}$$

The text-book solution to this optimization problem follows the Lagrange formalism [94] and reads

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp \left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right) , \tag{4.13}$$

where $h_i(A)$ and $e_{ij}(A, B)$ are the respective Lagrange multipliers. The local fields $h_i(A)$ represent the local biases for the sequence units and the coupling strengths $e_{ij}(A, B)$ their statistical coupling.

It is helpful for a compact formulation of the formalism to introduce the partition function

$$Z = \sum_{\{A_i|i=1,\dots,L\}} \exp \left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right) \tag{4.14}$$

that contains the local fields and couplings of Eq. (4.13) and the Hamiltonian

$$H = - \sum_{1 \leq i<j \leq L} e_{ij}(A_i, A_j) - \sum_{i=1}^{L} h_i(A_i) , \tag{4.15}$$

which enables us to formulate the probabilistic model as

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp \left(-H \right) . \tag{4.16}$$

The computational challenge of the determination of the marginals $P_i(A)$ and $P_{ij}(A,B)$ roots in the fact that it requires the summation over all alignments A_i in the partition function in Eq. (4.14).

Here it can be pointed out that the marginals $P_i(A)$ and $P_{ij}(A,B)$ can be analytically derived from the partition function by

$$\frac{\partial \ln(Z)}{\partial h_i(A)} = -P_i(A)\,, \tag{4.17}$$

$$\frac{\partial^2 \ln(Z)}{\partial h_i(A)\partial h_j(B)} = -P_{ij}(A,B) + P_i(A)P_j(B)\,. \tag{4.18}$$

Accordingly, we can introduce the connected correlations

$$C_{ij}(A,B) := P_{ij}(A,B) - P_i(A)P_j(B)\,, \tag{4.19}$$

where indices $i,j \in \{1,\ldots,L\}$ and $A,B \in \{1,\ldots,q-1\}$. The limitation of A,B to $\{1,\ldots,q-1\}$ removes linear dependencies due to the normalization of the two-site marginals P_{ij} and makes C_{ij} invertible. This makes $C_{ij}(A,B)$ a $L(q-1)\times L(q-1)$-dimensional matrix where the pairs (i,A) and (j,B) represent joint single indices.

4.2.2. Gauge Invariance

The number of parameters in the statistical model in (4.13) is $\binom{N}{2}q^2 + Nq$. This number is not the number of independent variables and can be corrected to $\binom{N}{2}(q-1)^2 + N(q-1)$ for an independent set of parameters. Therefore, the following conditions can be applied:

$$e_{ij}(A,q) = e_{ij}(q,A) = h_i(q) = 0\,, \tag{4.20}$$

which means that all couplings and biases are determined with respect to state q. Thereby, the number of variables corresponds to the number of constraints, which renders the solution of the maximum-entropy model unique.

4.2.3. Small-coupling Expansion

The explicit calculation of the partition function Z is computationally expensive and renders the calculation for typical sequence lengths of biopolymers unfeasible. In order to reduce the computational effort the algorithm is based on a small-coupling expansion [95, 96]. The perturbed coupled Hamiltonian needs to be expanded for small perturbations around the unperturbed case and can in general be introduced as

$$H(\alpha) = -\alpha \sum_{1\le i<j\le L} e_{ij}(A_i,A_j) - \sum_{i=1}^{L} h_i(A_i)\,, \tag{4.21}$$

where α parameterizes the perturbation, ranging from 0 for independent variables and 1 for the original model. The Gibbs potential

$$-\,G(\alpha) = \ln\left(\sum_{\{A_i|i=1,\ldots,L\}} e^{-H(\alpha)}\right) - \sum_{i=1}^{L}\sum_{B=1}^{q-1} h_i(B)P_i(B) \tag{4.22}$$

can be introduced as the Legendre transform of the free energy $F = -\ln Z$. In contrast to the free energy that depends on the fields $h_i(A)$ and the couplings $e_{ij}(A, B)$, the Gibbs potential depends only on the couplings and the single-site marginals $P_i(A)$:

$$G(\alpha) = G\left(\{\alpha e_{ij}(A,B)\}_{1\leq i<j\leq L}^{A,B=1,\ldots,q-1}, \{P_i(A)\}_{i=1,\ldots,L}^{A=1,\ldots,q-1}\right) . \tag{4.23}$$

From the Gibbs potential the fields can be derived via Legendre transformation rules by

$$h_i(A) = \frac{\partial G(\alpha)}{\partial P_i(A)} \tag{4.24}$$

and the inverted connected couplings by

$$\left(C^{-1}\right)_{ij}(A,B) = \frac{\partial h_i(A)}{\partial P_j(B)} = \frac{\partial^2 G(\alpha)}{\partial P_i(A) \partial P_j(B)} . \tag{4.25}$$

The restriction introduced in Eq. (4.19) renders matrix C_{ij} invertible in case of a finite pseudo-count λ. Therefore, the two-site marginal distribution P_{ij} can be calculated by differentiating the Gibbs potential twice with respect to single-site marginals at i and j and by inverting matrix C_{ij}, following Eq. (4.19). To this end, it is also necessary to determine an expression for the Gibbs potential, which will be done in a first order Taylor expansion given by

$$G(\alpha) = G(0) + \left.\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} \alpha + \mathcal{O}(\alpha^2) . \tag{4.26}$$

4.2.4. Independent-site and Mean-field Approximation

The first summand contains the Gibbs potential for $\alpha = 0$ (independent-site approximation). Without coupling, the Gibbs potential represents the negative entropy of an ensemble that consists of L uncoupled "spins" (with q states) $A_1, ..., A_L$ with given marginals $P_i(A_i)$. In this case the potential can be written as

$$\begin{aligned} G(0) =& \sum_{i=1}^{L} \sum_{A=1}^{q} P_i(A) \ln(P_i(A)) & (4.27) \\ =& \sum_{i=1}^{L} \sum_{A=1}^{q-1} P_i(A) \ln(P_i(A)) \\ &+ \sum_{i=1}^{L} \left(1 - \sum_{A=1}^{q-1} P_i(A)\right) \ln \left(1 - \sum_{A=1}^{q-1} P_i(A)\right) , & (4.28) \end{aligned}$$

where the last summand introduces the chosen gauge.

The next step is to derive an expression for the first derivative with respect to α. From the definition of Gibbs potential in Eq. (4.22) we can derive

$$\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha} = -\frac{\mathrm{d}}{\mathrm{d}\alpha}\ln(Z(\alpha)) - \sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{\mathrm{d}h_i(A}{\mathrm{d}\alpha}P_i(A) \tag{4.29}$$

$$= -\sum_{\{A_i\}}\left(\sum_{i<j}e_{ij}(A_i,A_j) + \sum_i \frac{\mathrm{d}h_i(A)}{\mathrm{d}\alpha}\right)\frac{e^{-H}(\alpha)}{Z(\alpha)}$$

$$-\sum_{i=1}^{L}\sum_{A=1}^{q-1}\frac{\mathrm{d}h_i(A}{\mathrm{d}\alpha}P_i(A) \tag{4.30}$$

$$= -\left\langle \sum_{i<j}e_{ij}(A_i,A_j)\right\rangle_{\alpha}\ , \tag{4.31}$$

which represents the average of the coupling term in the Hamiltonian. The evaluation of this expression for $\alpha = 0$ yields the factorized joint distribution of all variables over the single-site marginals

$$\left.\frac{\mathrm{d}G(\alpha)}{\mathrm{d}\alpha}\right|_{\alpha=0} = -\sum_{i<j}\sum_{A,B}e_{ij}(A,B)P_i(A)P_j(B)\,. \tag{4.32}$$

Inserting Eqs. (4.27) and (4.32) in the first-order approximation of the Gibbs potential yields self-consistent mean-field equations for the local fields

$$\frac{P_i(A)}{P_i(q)} = \exp\left(h_i(A) + \sum_{\{j|j\neq i\}}\sum_{B=1}^{q-1}e_{ij}(A,B)P_j(B)\right) \tag{4.33}$$

and the inverse of the connected correlation matrix as

$$\left.\left(C^{-1}\right)_{ij}(A,B)\right|_{\alpha=0} = \begin{cases} -e_{ij}(A,B) & \text{for } i\neq j \\ \frac{\delta_{A,B}}{P_i(A)} + \frac{1}{P_i(q)} & \text{for } i = j \end{cases}\ . \tag{4.34}$$

With this we only need to invert the connected correlation matrix that can be determined from the empiric single-site and pair frequency counts

$$C_{ij}^{\mathrm{emp}}(A,B) = f_{ij}(A,B) - f_i(A)f_j(B) \tag{4.35}$$

in order to calculate the couplings $e_{ij}(A,B)$.

4.2.5. Direct Information

In order to be able to rank the pair correlations based on the calculated coupling matrices $e_{ij}(A,B)$ we need to calculate scalar values from the respective matrices. A gauge independent quantity is the direct information (DI), as introduced in [92]:

$$\mathrm{DI}_{ij} = \sum_{A,B=1}^{q}P_{ij}^{\mathrm{dir}}(A,B)\ln\left(\frac{P_{ij}^{\mathrm{dir}}(A,B)}{f_i(A)f_j(B)}\right)\ , \tag{4.36}$$

where we assume a two-site model for a pair i, j

$$P_{ij}^{\mathrm{dir}}(A,B) = \frac{1}{Z_{ij}} \exp\left(e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\right) . \tag{4.37}$$

The local fields $\tilde{h}_i(A), \tilde{h}_j(B)$ need to be calculated by identifying the corresponding marginal distributions with the empirical single-site frequency counts

$$f_i(A) = \sum_{B=1}^{q} P_{ij}^{\mathrm{dir}}(A,B) \quad , \quad f_j(B) = \sum_{A=1}^{q} P_{ij}^{\mathrm{dir}}(A,B) \, . \tag{4.38}$$

The presented approach [25] disentangles direct and indirect correlations that are found in biomolecular sequence alignments, where direct correlations are able to indicate spatial contacts. In order to achieve this disentanglement it is necessary to find a complete, least constrained statistical model that describes a given empirical data set, i. e., single-site and pair frequency counts. The maximum-entropy model is such a model and its standard solution can be found with help of the Lagrange formalism. By introducing respective Lagrange multipliers a Hamiltonian that contains local fields and couplings is set up. The formulation is similar to a two-state Ising model but is generalized instead to a q-state Potts model where q is the size of the biomolecular system's sequence alphabet. In order to solve the inverse Potts model it would be necessary to calculate the partition function explicitly which is computationally demanding and exceeds available computational resources for typical sizes of biomolecular systems of interest. Three approximative steps ensure a less demanding procedure: A small-coupling expansion is applied to the Gibbs potential by expanding it in a Taylor series for small couplings. Independent-site and mean-field approximations yield the zeroth and first order approximation for the potential and give a relation between couplings and an inverted connected correlation matrix that can be calculated from empiric single-site and pair frequency counts. Therefore, the dominant computational effort becomes a matrix inversion instead of explicit summations. A calculated coupling matrix for a given sequence unit pair can be condensed into a scalar quantity, the "direct information" that serves as a prediction score for the sequence unit pair. This method has been successfully employed in protein contact prediction [25, 92, 97] and promises applicability in the field of RNA contact prediction.

5 Chapter 5.

eSBMTools: Native Structure-based Model Software Package

This chapter introduces a Python implementation of native (tertiary) structure-based models that was initiated by me and developed in cooperation with several members of our research group. The software package is under constant development and released under GNU General Public License version 3.0 at `http://sourceforge.net/projects/esbmtools`*. The chapter is based on our publication [19].*

Molecular dynamics (MD) simulations are able to complement experimental measurements to gain more detailed insights into the structure and function of biomolecular systems at experimentally inaccessible time or length scales, as discussed in Chap. 2. Limitations of the standard MD simulation scheme are overcome by simplistic biopolymer models, such as native structure-based models (SBM) that are based on energy landscape theory and the principle of minimal frustration, as introduced in detail in Chap. 3. The computational effort for protein and RNA folding simulations is reduced by structural and energetic coarse-graining. A reduced complexity enables the design of workflows that require extensive sampling over simulations of biophysical dynamics.

In order to perform such automatized studies it is desirable to utilize a flexible implementation of SBMs that interfaces with an established simulation software. I present the software package eSBMTools ("enhanced SBM tools") that provides functionality to set up, modify and evaluate SBMs for biomolecular systems. The software package is released open source and written in Python that is architecture independent and allows the installation on desktop or high performance computing systems. Its modules are tailored for the use with the molecular dynamics simulation program GROMACS [66].

First, the structure of the package – preprocessing and postprocessing – and the range of functionalities to assist the practical use of SBMs is described in more detail. Eventually, concrete benefits for use cases presented in this thesis are discussed.

5.1. Motivation and Overview

In the course of this thesis I started and was since in charge of the development of a Python [98] software package called “enhanced native structure-based modeling tools” (eSBMTools) [19]. The toolkit is tailored to the MD simulation software GROMACS [66]. It interfaces with its standard input file formats, some of the evaluation tools provided by GROMACS and the output files generated by simulations.

Energy landscape theory and the principle of minimal frustration are based on the necessity that biomolecular evolution results in an effective tendency of a foldable biopolymer to have a energy landscape that is funnel-shaped where its global minimum being the native, folded state, as discussed in Chap. 3. This model justifies finite folding times and resolves Levinthal’s paradoxon. Native structure-based models (SBMs) realize an idealized, minimally frustrated folding funnel by means of a simplistic interaction potential [14, 83] that allows MD simulations to reach biologically relevant time scales. The corresponding potential is constructed from contact information of the native fold. The number of realized native contacts in a conformation, the Q value, is at the same time the natural choice of a reaction coordinate. The Q value is, therefore, used as a reaction coordinate to investigate folding pathways or as an order parameter for thermodynamic evaluations, such as the weighted histogram analysis method (WHAM) [99].

The SBM potential can be reformulated with structural coarse-graining techniques. Protein sequences can be represented by a chain of beads at the positions of the C_α atoms in amino acids. Concrete formulations of the SBM in an all-atom [83] or a C_α [14] formulation have been established. eSBMTools provides both formulations and takes the native conformation from a given Protein Data Bank (PDB) [42] input file. Appropriate input files for simulations by the MD simulation suite GROMACS are generated.

As a result, the implementation of SBM force fields facilitates the setup of folding simulations that yield folding transitions on standard desktop computers. These simulations are able to reach effective time scales of seconds and allow extensive sampling in case of RNAs that consist of about 100 nucleotides on multipurpose high performance computing systems. In contrast to existing web-based solutions, such as the SMOG-server [18] that allow a straightforward setup of standard SBMs, eSBMTools equips the user in addition to standard setup routines with a wide range of extensions. The customization of an existing SBM in Python scripts allows the setup of automatized workflows. The workflows can vary the standard formulation by, e.g., adding additional, non-native contacts, introducing novel ligand or constituent topologies, or manipulating force field parameters or they can scan global simulation parameters. The need for this range of functionalities can be motivated by a number of published scenarios: The function of biomolecular machinery, such as the ribosome, often requires structural transitions between different conformations [100]. The addition of non-native contacts to the SBM has been successfully implemented and investigated [101]. The SBM’s structural flexibility allows deformations of the overall structure as a possible response to such non-native interactions. Competing and alternative conformations in dual (or even multi)-basin models [102–104] are the investigated results of respectively introduced deformations. Recent studies have also discovered the potential of contacts predicted by statistical physics methods to predict complexes [105], protein structures [106], active conformations [97] or transmembrane proteins [107]. Another example for a general application of the described use cases is the prediction of exited states during ribosome translocation by including contact information derived from experimental measurements [108]. eSBMTools has already been successfully applied in projects outside my personal research focus in a broad study about folding pathways in proteins [109] and for modeling experimental markers in protein folding simulations [unpublished data by Ines Reinartz].

In the following the implementation of the model is described in detail. eSBMTools' structure, its interfaces to the GROMACS software package and the range of functions is outlined. In the last section, the implications of eSBMTools for conducted studies in the course of my research are discussed.

5.2. eSBMTools: Implementation of the Model

eSBMTools is organized in 13 modules that can be imported into custom Python projects. This architecture allows the flexible assembly of functionalities for advanced workflows in the context of SBM simulations. The modules accept input files and create output files that meet the formats that are defined by tools of the standard GROMACS software package [66]. The Python package assists the user during all steps around SBM simulations:

- generation of the files that define an SBM topology and geometry for GROMACS simulations
- optional: manipulation of the SBM depending on the study
- generation of required simulation parameter input files for GROMACS
- execution of post-processing protocols according to study
- visualization of results

The SBM is based on a native conformation defined by a PDB structure file. An XML-based topology definition that is provided by the package itself introduces bonded interactions, such as bonds, angles and planar and proper dihedral angles. The provided topology definitions describe DNA/RNA (by nucleotide building blocks) or protein (amino acid building blocks) systems in all-atom or C_α formulation. The flexible XML format enables the user to realize custom topology definitions in addition to the existing entries. This way, ligands, modified sequence constituents, or experimental markers, such as Förster resonance energy transfer (FRET)-fluorophores [110, 111], can be introduced to a model. The strategy how to include FRET-fluorophores into SBM has recently been introduced in our work group [unpublished data by Ines Reinartz]. Non-bonded interactions as atom-atom contacts are implemented by a simple cut-off formulation: A contact between two atoms is considered to be formed if the distance between the two involved atoms is below a certain threshold, by default 0.4 nm. A contact map represents every contact between two atoms by an entry, as in Fig. 5.1.

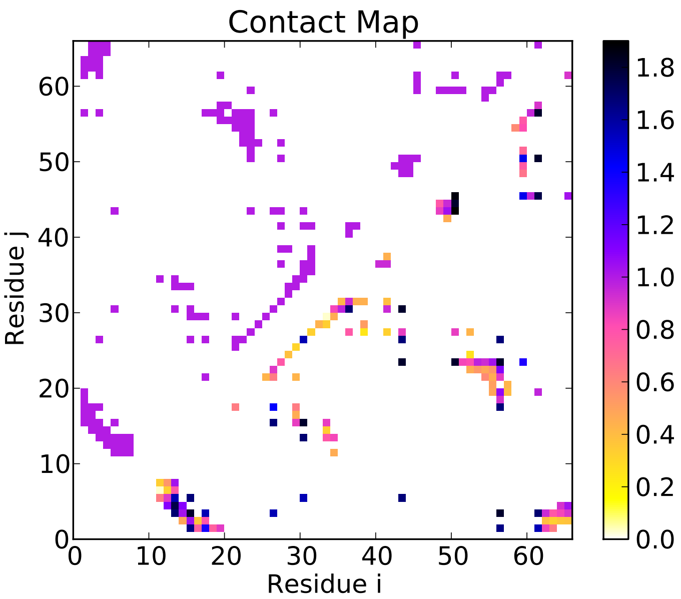


Figure 5.1.: Exemplary contact map of protein ubiquitin (PDB ID 1UBQ) with possible modifications. The top left half of the in principle symmetric binary matrix shows the plain residue-residue contact map of the protein (with weight 1.0). The bottom right half of the matrix shows the same contacts but with reweighted contact strengths. The weighting factors depend on the residue-residue pair involved in the contact and has been published by Miyazawa and Jernigan [112] based on empiric statistical values.

The generated SBM can be modified according to the requirements of the user's system of interest. The contact information can be manipulated to include, e. g., predicted contacts [25, 97] or the energetics of all contacts can be reweighted by, e. g., by amino acid interaction matrices [112], shown in Fig. 5.1. Two SBMs can be merged in order to setup combined systems for, e. g., complex formation simulations [105].

Once the files that describe the system of interest – coordinates and force field – are prepared, eSBMTools generates the required configuration files for GROMACS simulations. This generation can be customized within Python to adjust simulation parameters, e.g., temperature, duration or time resolution to steer simulations within a workflow. Examples of this capability have been presented for automatized protein folding studies by Claude Sinner in our group [109, 113].

In general, the range of functions of eSBMTools can be divided in two categories: preprocessing and postprocessing. In the following, more detailed descriptions of the functionalities within these two categories are presented.

5.2.1. Preprocessing

The preprocessing range of functions features the generation of the following set of files:

- `.gro` file: The coordinate file contains a complete list of atom positions in the system. These positions serve as starting conformation for the MD simulation.

  The coordinates are taken from a PDB (Protein Data Bank [42]) file that provides the native conformation for the SBM. Parsing of the PDB file is performed by the functionality provided by Biopython [114]. The organization by chains that the PDB file format features is lost due to the required GROMACS file format.

- `.top` file: The topology file consists of an atom type list, a full reference atom list, a list of non-bonded contacts and list of the bonded interactions bonds, angles and dihedral angles. A detailed description of the `.top` file composition can be found in Sec. A.1.

  The information in this file about the present bonded interactions is generated based on an XML based definition of bonded interactions within and between the building blocks of a biomolecule. A library of standard biomolecular building blocks, such as nucleotides (RNA, DNA) and amino acids (proteins), are included in the package. Each building block is set up by a list of atoms, bonds, angles and dihedral angles in human readable form that can be exchanged or extended at will.

  The non-bonded contact information is based on the geometry of the biomolecule and generated via a simple cut-off criterion as contact definition. Chains that have been present in the PDB file are treated independently in terms of bonded and non-bonded interactions.

- `.mdp` file: The configuration file defines all simulation parameters, such as temperatures, coupling settings, time steps, random seeds or the output frequency.

  All options available in the GROMACS configuration file format are accessible by key words within the Python implementation. Therefore, the generation is completely customizable and automatized workflows can be steered by a Python project.

During setup of the required simulation files the standard formulation of the model can be customized. The contact map can be reweighted, e. g., by statistically or biophysically motivated relative residue-residue strengths. Additional contacts can be added to the native ones to create interactions between monomers that form complexes or to provoke conformational changes in a structure. The implementation also allows merging of two existing models which would be a first step to create complexes from two independent structures or to model biomolecular assembly, crowding or the presence of ligands. Alternatively, the implementation allows to generate subsets of contacts to investigate the characteristic influence of specific contact classes on the structural stability of biomolecular systems. The Python-based access to configuration options facilitates automatized temperature scans in workflows and the modifications of random seeds guarantees statistical independence of dynamic data in different simulations.

5.2.2. Postprocessing

The postprocessing of MD simulations is based on the temporal sequence of conformational states of the system, the “trajectory”. Trajectories are saved by default in binary `.xtc` files that can be processed by GROMACS’s various extensions. From the trajectories, characteristic values can be generated:

- root mean-square deviation (RMSD) values [115]

  $$\mathrm{RMSD}(C_i, C_j) = \min\left(\frac{1}{N}\sum_{k=1}^{N} |\mathbf{r}_{ik} - \mathbf{r}_{jk}|^2\right)^{\frac{1}{2}}_{\mathrm{rot,trans}} , \tag{5.1}$$

  where C_i, C_j are the two conformations that exhibit deviations between each other, N is the number of atoms in the system and $\min\,(.)_{\mathrm{rot,trans}}$ denotes the minimum over all relative orientations between the two conformations. eSBMTools provides a wrapper for calculations that are performed by the existing GROMACS tool `g_rms`.

- Q values [116]

  $$Q_k = \sum_{\{ij\}}^{N_c} q_{ij}^k\,, \tag{5.2}$$

  where N_c is the number of contacts and

  $$q_{ij}^k = \begin{cases} 1, & \text{if} \quad 0.8 \le r_{ij}^k/\sigma_{ij}^0 \le 1.2 \\ 0, & \text{otherwise}\,, \end{cases} \tag{5.3}$$

  where r_{ij}^k is the actual distance of atom pair ij in a given conformation k and σ_{ij}^0 is the respective native distance. Therefore, a Q value is the ratio of formed native contacts in a given conformation. These values are a natural reaction coordinate for SBM simulations [82]. eSBMTools provides a wrapper for calculations that are performed by the existing custom GROMACS extension `g_kuh` [117]. Alternatively, the software package features its own implementation for Q value calculations that is about ten times slower than `g_kuh` but does not depend on a custom GROMACS extension.

The described characteristic values can be used to create evaluation scenarios, such as contact maps over time or Q value filtering for certain substructural elements. Q values of simulations at different temperatures, preferable around the folding temperature, are used for weighted histogram analysis method (WHAM) [99, 118] calculations that generate a free energy landscape $F(Q,T)$ over temperature T and Q value and the heat capacity $C_V(T)$. Visualization is realized by Python’s internal functionality provided by matplotlib as part of Scipy [119]

5.3. Discussion

I have presented the software implementation that was created in the course of my studies and contributed to all my research projects and several ongoing projects in the research group. eSBMTools has been released to the public as an open source repository. Giving the community a local implementation, in contrast to a web server, has several benefits: The user has control over the analyses without the suspicion that operators of a web server are logging submissions and scan them for interesting projects. Secondly, the Python based implementation facilitates the setup of complex workflows using the provided functionality as is. Since Python is installed on most of today's multipurpose high performance computing (HPC) facilities these workflows can also operate directly within submitted jobs. The XML-based topology definitions allows the addition of structural information without expert programming knowledge. Eventually, the more experienced users can exploit the fact that the implementation is open source and therefore extendable by own functions. I chose GROMACS as a state-of-the-art MD implementation to interface with eSBMTools since it provides the standard range of functionalities and is commonly available on HPC systems. Additionally, the force field interface of GROMACS is flexible enough to accept the simplistic parametrization of an SBM.

The software package has been successfully presented and applied in lecture tutorials. Apart from that, its application in the course of my research that is presented later on in my thesis is manifold. It plays a major role in the setup and evaluation of simulations of cotranscriptional riboswitch folding, as discussed in Chap. 6. Coordinates of a helically parameterized tube are generated by adapting the existing XML implementation and subsequent merging of the tube with existing SBM topologies and coordinates is realized with eSBMTools' standard formulation. Simulations are evaluated with the help of Q value tools that allow the setup of Q value filtering for regional Q values in helical substructures. My investigations on the assessment of coevolutionary tertiary structure contact prediction in RNA, as presented in Chap. 7, has greatly benefited from the flexible implementation of XML based topologies. The architecture gives control over every part of the forcefield definitions individually and facilitates the formulation of "generalized" SBMs, later referred to as "knowledge-based models" (KBMs).

Beyond specific individual applications of the tool, a workflow realized by eSBMTools has been deployed and published very recently as part of a computer grid implementation that gives access to SBM simulations on the Molecular Simulation Grid (MoSGrid, `https://mosgrid.de/portal`) and is started and monitored within a web browser [120].

6 Chapter 6. Cotranscriptional Riboswitch Folding

In this chapter I present my results from a study on cotranscriptional riboswitch folding. My results are then compared with an according study by Dipl. Phys. Michael Faber at the Max Planck Institute of Colloids and Interfaces in Golm, Germany. The chapter is based on our publication [24].

Riboswitches are sequences in the noncoding region at the start of messenger RNA (mRNA) that are able to sense environmental conditions by binding small molecules, often referred to as metabolites or ligands. Riboswitches can react structurally to a certain level of concentration of these metabolites, if they bind successfully. This structural reaction resembles a two-state switch that turns genetic expression of the downstream gene "on" or "off". The decision between the formation of either state, has to occur while the transcript leaves its factory, the RNA polymerase (RNAP), i. e., during transcription. The extrusion out of RNAP, folding and ligand binding all occur simultaneously and interdependently on the seconds time scale.

I employ a coarse-grained in-silico technique to investigate the influence of the crowded environment of RNAP on the folding characteristics of the SAM-I and add adenine riboswitches at varying extrusion velocities. Native structure-based model (SBM) simulations yield an atomically resolved dynamic model of riboswitch folding. The homogeneous energetics of the model introduce solvents and the presence of ligands implicitly. Depending on extrusion at various transcription rates, I observe and quantify different pathways in the formation of substructural elements. In the investigated scenarios, free-folding riboswitches can exhibit different folding characteristics compared to transcription-rate limited folding. Since the critical transcription rate distinguishing these cases is higher than physiologically relevant transcription rates, my findings suggest that cotranscriptional folding is reliably transcription rate limited in case of the SAM-I and add adenine riboswitch.

Subsequently, my results are compared to results from the energetically more detailed kinetic Monte Carlo simulations by Michael Faber. The kinetic Monte Carlo method gives access to longer time scales by describing folding on the native secondary structure level. Our findings are in robust agreement given the complementarity of both techniques.

6.1. Motivation and Overview

Riboswitches are a specialized subclass of structured RNA and situated in the untranslated region (UTR) at the 5' end of messenger RNA (mRNA). The nascent RNA strand is sequenced by the RNA polymerase (RNAP) according to a DNA sequence: the transcription process [35, 121, 122]. Small metabolites (or “ligands”) in the vicinity of the nascent RNA strand can bind to the riboswitch that itself responds by conformational changes. These conformational changes cause again structural responses that can prohibit expression of the downstream gene by transcription termination or translational repression [5–8, 123]. In order to perform this function, a riboswitch consists of two structural subdomains: aptamer region and expression platform. The aptamer region detects and binds the ligand which results in a specifically folded conformation that is stabilized by the bound ligand. As a consequence, the expression platform reacts structurally to the respective conformation of the aptamer. This reaction is then tailored to decide between two conformational states: a state that permits transcription and translation and a state that can attenuate transcription or translation. Thereby, a two-state switch depending on ligand binding is realized. Ligand binding requires a certain concentration of ligands surrounding the RNA strand during transcription. However, it has been observed that even at high ligand concentrations ligand binding can be highly suppressed [124]. These findings suggest that conformational changes in the aptamer can prohibit ligand binding which gives the metabolite not enough time to reach thermodynamic equilibrium in the bound state until the termination decision. The apparent dissociation constant for an “on” riboswitch (bound ligand allows transcription) can, therefore, be defined as

$$K_d = \frac{k_{off}}{k_{on}}\,, \tag{6.1}$$

where k_{off} is the time rate to terminate transcription and k_{on} is the time rate × concentration of ligands to continue transcription. One recent goal of riboswitch investigation is to determine whether a riboswitch is thermodynamically or kinetically controlled [125–127], i. e., whether thermodynamic equilibrium with the ligand can be reached within the time window given by transcription (thermodynamic control). In a recent experimental study, ligand binding of the *pbuE* adenine riboswitch has been directly observed *in-vivo* [39]. As a result, this riboswitch is found to be kinetically controlled.

Cotranscriptional folding of RNA differs from free folding mainly in two aspects: First, the folding chain is still growing and, thus, the subset of potential interaction partners for any atom in the nascent chain evolves with time. As a consequence, the range of possible conformations is also time-dependent and different substructures may have different accessible time windows for folding. Secondly, spatial restriction (or reduction of chain entropy) arises from steric interaction with the RNAP, the cellular machinery that reads out the genetic information stored in DNA and synthesizes a complementary RNA strand. The nascent RNA strand leaves the RNAP through an exit channel which introduces spatial constraints for the emerging RNA allowing each nucleic acid to leave the RNAP individually. Only outside RNAP the RNA is able to form secondary structural elements and, thus, the nascent RNA strand experiences drastic spatial restrictions. Considering these two factors, transcription may have an influence on the folding behavior of riboswitches and, eventually, ligand binding.

Experimental studies of RNA folding, such as single-molecule FRET measurements [128] or optical trap extension experiments [129], are complex and costly processes that motivate research for computational methods to complement such analyses. There have been several approaches to combine computational and analytical methods with experimental findings to systematically enhance our understanding of RNA folding [76, 130–132]. Atomically resolved simulations could add nanoscopic insight to the existing picture of riboswitch folding characteristics. Straightforward all-atom molecular dynamics simulations with explicit solvent, as introduced in Chap. 2, however, are typically still limited to simulated times in the order of hundreds of nanoseconds in the context of RNA [133]. Because riboswitch folding occurs at time scales in the order of seconds [134], such simulations would exceed present day's computational resources by several orders of magnitudes.

In order to overcome these computational limitations, my study focuses on a structure-based method to investigate cotranslational folding of two exemplary riboswitches. Native structure-based models (SBMs) employ a potential that is based on the native fold of a biomolecule, as described in Chap. 3. Motivated by energy landscape theory, this model exhibits a smooth, funneled energy landscape dominated by native interactions. The energetic coarse-graining of this model allows, while being still atomically resolved and dynamic, to reach the biologically relevant time scales of RNA folding with comparably moderate computational effort [22, 23, 108]. Free folding of the SAM-I riboswitch has been investigated previously by means of the native structure-based model (SBM). The focus of that study has been on ligand binding and nonlocal helix formation in the aptamer region during free folding [22]. A similar study focused on a preQ1 riboswitch [23]. By contrast, the present study focuses on cotranscriptional riboswitch folding and, thereby, on the influence of a crowded environment on the overall folding characteristics of nascent riboswitches. In my model, the stretched RNA strand is pushed out of a flexible tube with a funnel-like exit region emulating the extrusion of a nascent RNA strand out of RNAP. The acting force is distributed over a number of residues while they are inside the tube. Every segment of the strand that leaves the tube is released of the acting force and, therefore, free to fold. The physiologically relevant transcription rates of RNAP vary over about one order of magnitude (about 15 to 80 nucleotides per second, nt/s) [35]. This flexibility is accomplished by various pausing mechanisms that influence transcription speed [135–137]. According to my results, the folding characteristics of riboswitches are transcription-rate limited within the range of physiological transcription rates. Therefore, the folding characteristics of cotranscriptional riboswitch folding differs robustly from the free folding scenario. My results are then compared to respective results by a kinetic Monte Carlo study performed by Dipl. Phys. Michael Faber. The comparison of both techniques in the course of our collaboration yields good agreement that backs the significance of our computational results.

6.2. Method

I present the methodological elements that were used in the course of this study. First, the structures of interest, the SAM-I and *add* adenine riboswitches, are introduced and their characteristic structural features highlighted. Secondly, the very simplistic coarse-grained model of the RNAP is described and motivated in the context of SBM simulations. After that, the concrete formulation and application of SBMs within this investigation is discussed and the necessary steps towards an accurate model depicted. In the end, I define the evaluation protocol that allows a compact presentation of the key results and makes them at the same time comparable to results obtained by the kinetic Monte Carlo approach.

6.2.1. Structures of Interest

The investigated structures are the aptamer regions of two riboswitches with converse switching behavior and regulatory functions:

1. The SAM-I riboswitch from bacterium *Thermoanaerobacter tengcongensis* is an “off” switch that regulates transcription. The investigated sequence consists of 94 nucleotides and its structure has been published with PDB ID 2GIS [20] (shown in Fig. 6.1A). The 94 nucleotides are arranged in a four-way helical junction with two pairs of coaxially stacked helices P1 to P4. Their stems consist of 8, 7, 6 and 5 base pairs, respectively. A detailed depiction of the nucleotides involved in secondary structure is given in Fig. 6.2. The metabolite S-adenosylmethionine (SAM) binds to a binding pocket between P1 and P3.

2. The *add* adenine riboswitch from the procaryotic organism *Vibrio vulnificus* is an “on” switch that modulates translation initiation and consists of 71 nucleotides arranged in a three-way helical junction as published in PDB ID 1Y26 [21] (shown in Fig. 6.1B). The helices P1 to P3 exhibit stems consisting of 9, 6 and 6 base pairs, respectively, as seen in detail in Fig. 6.2. The ligand binds between the two coaxially stacked helices P1 and P2.

The two structures are experimentally resolved with high resolution (less than 0.3 nm) and exhibit two different switching mechanisms. In the following I want to investigate how folding is affected by transcription for these two structures, specifically if and how the folding order of the substructural elements depends on the transcription rate. Two folding scenarios are studied here: free folding and cotranscriptional folding of the whole aptamer structure.

6.2.2. Extrusion Scenarios

The presence of the RNAP during transcription needs to be modeled in simulations by means that are physiological accurate and technically realizable. Possible models would be:

- an explicitly growing strand
- extended and fixated strand that is released residue by residue
- extended strand in a completely enclosing tube with an exit region

The first model would mean restarting simulations for each nucleotide of the sequence at a group of fixated atoms that is also shifted for every new simulation. This procedure is technically quite laborious without any physiologically relevant advantages. The second

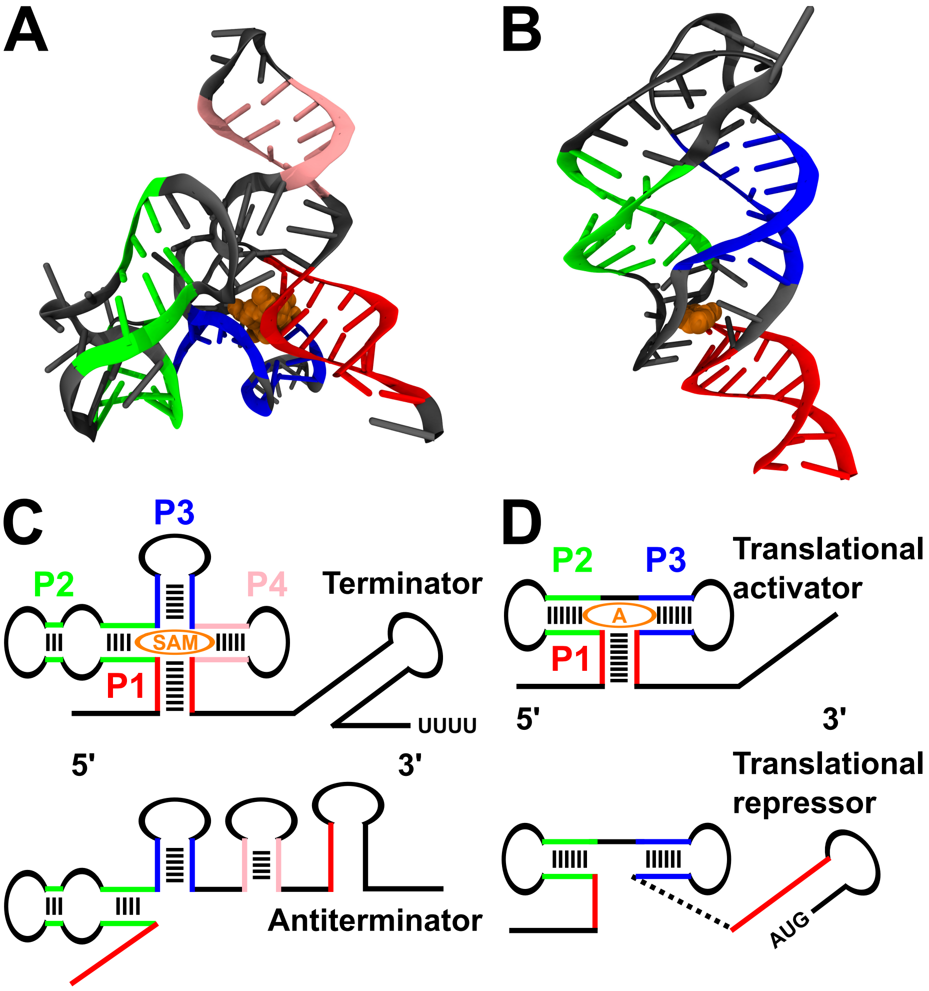


Figure 6.1.: Tertiary and secondary structures of the SAM-I and *add* adenine riboswitches in the ligand bound state. (*A*) Aptamer region of the SAM-I riboswitch (PDB ID 2GIS): The colored strands indicate elements of secondary structure. Helix P1 is highlighted in red, P2 in green, P3 in blue, and P4 in pink. The ligand is shown in orange. (*B*) Aptamer region of the *add* adenine riboswitch (PDB ID 1Y26). The same colors as in (*A*) for helices P1 to P3 and ligand are used. Helices P1 and P3 are coaxially stacked. (*C*) The SAM-I riboswitch consists of two pairs of coaxially stacked helices P1 to P4 connected by a four-way helical junction in its ligand bound state. Helix P1 forms in the presence of the ligand and acts as an anti-anti-terminator allowing the terminator (long stem loop with downstream sequence of uridines) to fold. In this case, transcription is terminated. (*D*) The *add* adenine riboswitch exhibits three helices P1 to P3 in its ligand bound state. Helix P1 forms in the presence of the ligand and prevents a translational repressor (initiation codon AUG paired in long stem loop) from forming.

The figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

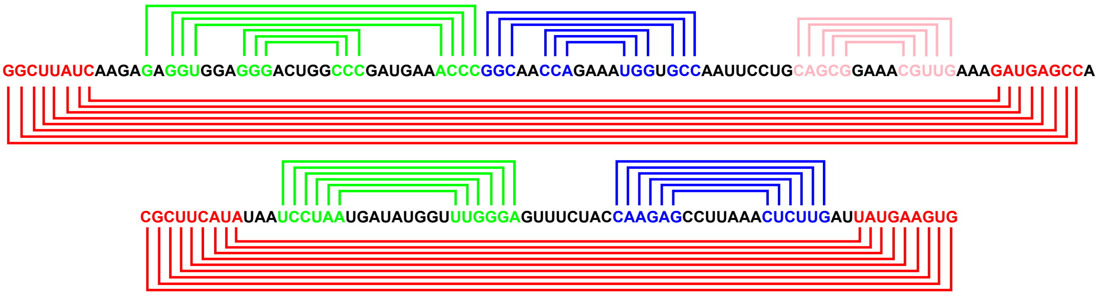

GGCUUAUCAAGAGAGGUGGAGGGACUGGCCCGAUGAAACCCGGCAACCAGAAAUGGUGCCAAUUCCUGCAGCGGAAACGUUGAAAGAUGAGCCA
CGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUG

Figure 6.2.: Secondary structure base pairs in the SAM-I (top) riboswitch and the adenine *add* (bottom) in their respective ligand bound conformations. Each connecting line between two bases in the sequence depicts a Watson-Crick (G-C, A-U) or Wobble (G-U) base pair. Different hair pin loops are color coded as in Fig. 6.1: P1 in red, P2 in green, P3 in blue and P4 in pink.

The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

model is the technical easiest implementation that lacks the effect of mechanical drag and the sterical influence of a bulky exit region. The third model, as seen in Fig. 6.3, is the favored scenario since it allows a mechanically realistic implementation at a technically feasible effort. It comes at the price of an in principle unnecessary large system due to the oversized tube that has to accommodate the whole extended RNA strand.

The beads that are forming the tube are positioned based on a helical parametrization, see Fig. 6.3. The helical parametrization positions beads with a distance Δ of 0.4 nm and a pitch h of 0.4 nm. The tube needs to accommodate a whole strechted strand of 94 nucleotides (length of the SAM-I riboswitch). Therefore, the tube has a diameter d_1 of 2 nm and a length L_1 of 110 nm. A gradual transition at the opening is realized by an exit funnel that has a length L_2 of 4 nm and an outer diameter d_2 of 3 nm. These are the steric boundary conditions that prevent secondary structure formation of the strand before it has left the RNAP. Transcription as a dynamic process is then modeled by forces that act between the rear end of the tube and every tenth nucleotide along the sequence. In steered molecular dynamic simulations, as seen in Sec. 2.1.6, the aptamer of the riboswitch is extruded out of the tube at a constant rate. Whenever a nucleotide reaches open space outside the tube, the simulation is terminated and subsequently continued without the acting force that belonged to the nucleotide that is no longer inside the tube. From that moment on this part of the strand is free to form secondary structure contacts as it would be the case for cotranscriptional folding.

6.2.3. Cotranscriptional Riboswitch Folding in SBM Simulations

The nascent RNA strand is generated by pulling apart both ends of the native structure to a linear chain of maximal length in a steered molecular dynamics simulation for an SBM

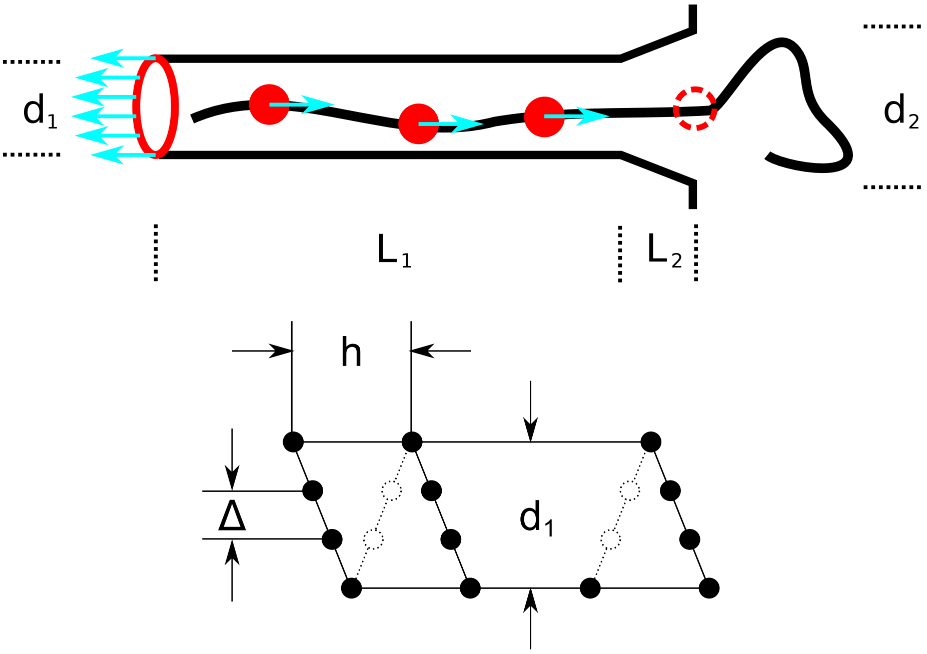


Figure 6.3.: Model of RNA polymerase (RNAP) in SBM simulations. Top: The tube with a funnel-like exit region composed of a helix of SBM atoms surrounds the stretched RNA. The tube atoms are positioned on a helix with a diameter d_1 of 2 nm and a length L_1 of 110 nm to contain the whole stretched RNA strand. The exit funnel has a length L_2 of 4 nm and an outer diameter d_2 of 3 nm. These are the spatial constraints that prevent folding before the nucleotides have left the RNAP. Forces acting between the rear end of the tube (red ring) and every tenth nucleotide along the sequence (filled red circles) extrude the RNA strand out of the tube with a constant rate. Whenever a nucleotide leaves the tube, it is released of its acting force and therefore free to fold mimicking the natural sequential transcription process. Bottom: The structure consists of helically arranged beads. The beads are positioned with a distance Δ of 0.4 nm and a pitch h of 0.4 nm.

The top figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

of the RNA. The RNAP is modeled by bonded beads positioned at helically parametrized coordinates of a hollow cylinder that is long enough to contain the respective stretched RNA strand. The beads are connected via bonds between next and next but one neighbors along the helical parametrization. Each bead is represented by the uniform exclusion radius of 0.25 nm for an SBM atom which creates spatial constraints built up by hard spheres. The stretched RNA strand is then placed inside the tube and an SBM is generated for the combined structure of tube and RNA. The combined SBM contains the geometry of both individual structures without topological interactions. The extrusion process is reproduced by forces that act between the rear end of the tube and every tenth nucleic acid of the nascent strand, which is thereby driven out of the tube.

The system of interest is represented by its potential energy from which the position dependent forces on each particle in the system can be calculated. Molecular dynamics simulations solve the corresponding Newtonian equations of motion by numeric time integration, as described in Sec. 2.1.3. Key to this procedure is the formulation of a potential energy scheme, which I choose to be the SBM, as introduced in Sec. 3.3. SBM is originally motivated by the observation that evolution favored the formation of funneled energy landscapes for proteins [12–17]. Funneled energy landscapes represent folding processes that are guided towards the native state by the cooperativity of their interactions with minimal frustration, which has also been shown for structured RNA [77, 78, 138]. SBMs realize the ideal case of a perfectly smooth, minimally frustrated and funneled energy landscape where interactions of the native conformation dominate the interaction network. The all-atom formulation of the structure-based potential [83] I use reads as

$$\begin{aligned} V = & \sum_{\text{bonds}} K_{\text{b}}(r - r_0)^2 + \sum_{\text{angles}} K_{\text{a}}(\theta - \theta_0)^2 \\ & + \sum_{\text{improper}} K_{\text{i}}(\chi - \chi_0)^2 + \sum_{\text{dihedrals}} K_{\text{d}} f_{\text{d}}(\phi) \\ & + \sum_{\text{contacts}} K_{\text{c}} \left[\left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{6} \right] \\ & + \sum_{\text{non-contacts}} K_{\text{nc}} \left(\frac{\tilde{\sigma}}{r_{ij}} \right)^{12} , \end{aligned} \tag{6.2}$$

where the dihedral (or torsional) angle potential is given by

$$f_{\text{d}}(\phi) = \left[1 - \cos(\phi - \phi_0)\right] + \frac{1}{2}\left[1 - \cos(3 \cdot (\phi - \phi_0))\right] \tag{6.3}$$

and K_{b}, K_{a}, K_{i}, K_{d}, K_{c} and K_{nc} are the corresponding force constants that are presented in more detail in Sec. A.1. The parameters r_0, θ_0, χ_0, ϕ_0 and σ_{ij}^0 are taken from the native structure and $\tilde{\sigma}$ is a global exclusion radius. Accordingly, the potential has its minimum at the native conformation. The information of bonded interactions (bonds, angles, planar dihedral and proper dihedral angles) is complemented by contact information that is introduced via Lennard-Jones terms in Eq. (6.2). Contact information can be depicted in a diagonal symmetric, binary matrix, often referred to as the “contact map” of a biomolecular structure (see Sec. 3.3). We use a “shadow map” [81] as contact map that regards atoms in contact within 0.6 nm radius as long as they are not shadowed by atoms within the connecting line. Nucleotides are allowed to form contacts between neighboring residues in order to be able to model stacking interactions (for a detailed description see Sec. 1.2.2). The contacts are introduced via Lennard-Jones terms with minima at the native atom-atom distances. Purely repulsive $1/r_{ij}^{12}$ terms that are characterized by the uniform exclusion radius $\tilde{\sigma}$ represent all other possible, non-contact pairings of atoms. The two SBMs of the riboswitch aptamer regions for my simulations in the described formulation are generated by the SMOG webserver [18].

6.2.4. Molecular Dynamics

I run the SBM simulations with the GROMACS software package [66]. The simulations are performed at temperatures in reduced units of 62. A time step of 0.001 is used for the extrusion simulations and 0.002 for the free folding simulations. The temperature is introduced and kept constant via Langevin dynamics with a coupling constant of 1 and a friction number of 1. In the extrusion scenario, we apply a constant velocity pull option with different constant rates, ranging from 0.0025 to 0.1.

6.2.5. Introducing Physical Temperature and Time Scale

The introduced force constants are homogeneous and normalized with respect to the system size and number of contacts (as shown in Sec. A.1) but they have otherwise no physical meaning. Thus, the model lacks physical time and temperature scales. In the following, the procedures employed to introduce a frame of reference for the used simulation temperature and a time scale are discussed.

I perform an all-atom molecular dynamics simulation based on the AMBER99 force field with TIP3P water and counter ions [139], a time step of 2 femtoseconds and Berendsen temperature coupling. The reference simulation of 1 μs at a physical temperature of 300 K can be compared to spatial root mean-square fluctuations (RMSF) [115] of each nucleotide n given by

$$\mathrm{RMSF}(n) = \sqrt{\frac{1}{N}\sum_{i}^{N} |\mathbf{r}_n^i - \langle \mathbf{r}_n \rangle|^2}\,, \tag{6.4}$$

where N is the number of conformations within a time-dependent trajectory and $\langle . \rangle$ denotes the time-average. RMSF values of the standard MD simulation at 300 K and several reference SBM simulations are shown in Fig. 6.4. The mean-square deviation (MSD) between the RMSF characteristics can be calculated as a global comparative value. The minimal MSD is found for a simulation at a GROMACS temperature of 90, which is shown in Fig. 6.5. As a result, I simulate at a temperature in the kinetic regime (62 in reduced units), to accelerate folding by about one order of magnitude to allow extensive sampling (see Sec. A.2 for a more detailed discussion).

I run 180 free folding simulations at the simulation temperature of 62 in reduced units. From the free folding simulations a simulated folding time can be determined that can be compared with experimental results [134]. In order to gain an estimate for the folding time in my simulations, the root mean-square deviations (RMSDs) with respect to the native fold of each simulation frame are extracted from the trajectory. When the RMSD undercuts a threshold of 0.3 nm the corresponding frame in the simulated trajectory is considered a folding event. A histogram of these 180 folding events that is presented in Sec. A.2 yields an asymmetric distribution whose maximum is regarded as an estimate for the folding time. Comparison of this simulated folding time with an experimental value for the folding time of an adenine riboswitch [134] yields 20 nt/s for the smallest extrusion rate of 0.0025.

6.2.6. Evaluation

I choose the number of formed base pairs as the reaction coordinate for my analyses. This reaction coordinate permits sampling over randomized trajectories while parametrizing the folding progress. This choice is also suitable for direct comparisons with the kinetic Monte Carlo method, see Sec. 6.3.1, where the number of base pairs is the natural reaction coordinate. The observables for my study are the numbers of formed base pairs within substructural elements, i. e., the stems of native local and non-local helical hairpin loops in the riboswitches. This observable will be referred to as the “regional Q value”. A base pair in SBM simulations is considered as formed if more than 50% of the native interatomic contacts between the two involved bases are formed. Further investigations show that the folding characteristics are stable over a wide range of choices for this threshold (40 - 80%, see Sec. A.3).

The introduced reaction coordinate and observables allow the evaluation of folding characteristics in both riboswitch systems. Normalization of the regional Q value allows a more clear arrangement of all substructural elements in a single plot for comparison, see Fig. 6.6 as part of Sec. 6.3. In a next step of information condensation a whole plot depicting the folding characteristics, e. g., for a given extrusion rate, is reduced to a single value for each substructural element: the “mid Q value”. The mid Q value of a helical folding characteristic is determined by the number of formed helical base pairs at which the normalized regional Q value of a substructural element is equal to 0.5.

6.3. Results

As a first result, 180 free folding simulations for each riboswitch are evaluated as a reference for folding characteristics. Moreover, they serve as reference simulations that yield a computational estimate of folding time that can be compared with experimental data. As discussed in Sec. 6.2.5 the SBM parametrization does not feature physical time or temperature scales. Therefore, they need to be introduced by comparison of observables with experiments or empirical force field simulations. We use the root mean-square fluctuations (RMSF) for a temperature calibration and the folding time for adjusting the time scale. My simulations are subsequently performed at a temperature of 62, well below a temperature of 90 that we identified with a physical temperature of 300 K. Simulations at a lower temperature result in kinetically driven trajectories which speeds up the folding process (see Sec. A.2). Accelerated folding reduces the computational effort of folding simulations by about one order

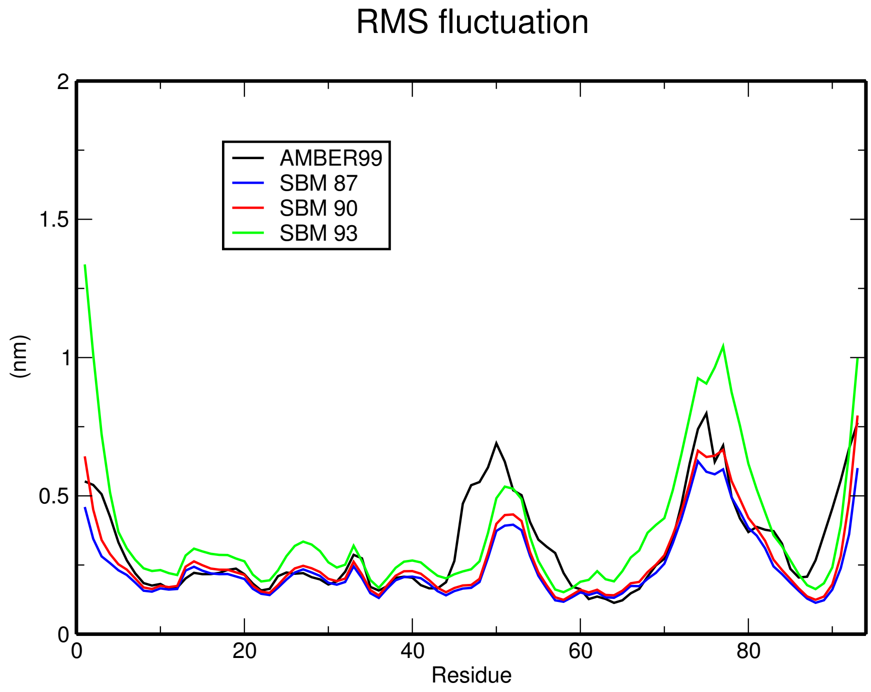


Figure 6.4.: Root mean-square fluctuations in AMBER99 and SBM simulations. A reference simulation (SAM-I riboswitch, AMBER99 force-field with TIP3P water and counter ions, 1 µs duration at 300 K in GROMACS) has been run in order to relate the reduced temperatures of the SBMs to a physical scale. The root mean-square fluctuations (RMSFs) per nucleotide for the reference simulation can be compared to SBM simulations at various temperatures in reduced units. By comparison of the respective RMSF values, the best agreement can be identified as the corresponding physical temperature, see Fig. 6.5.
The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

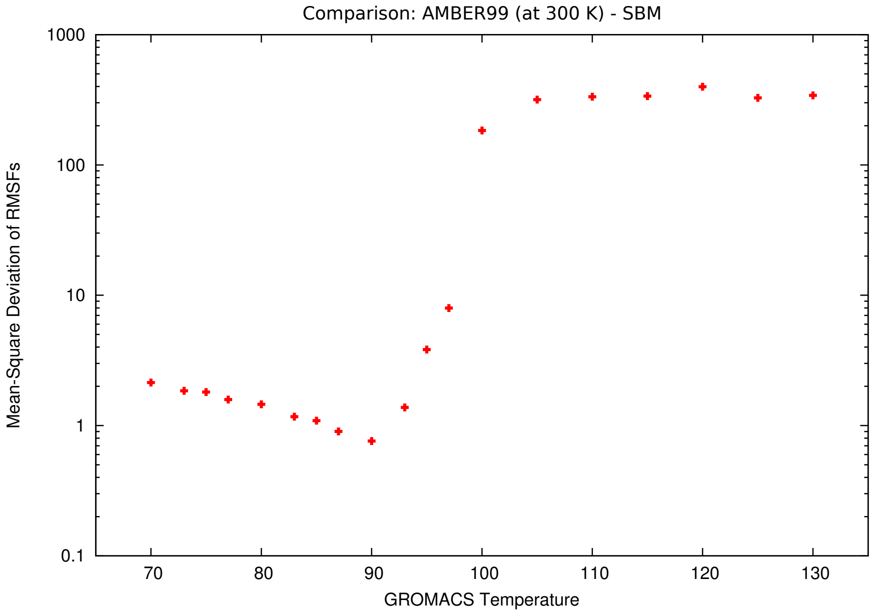


Figure 6.5.: Temperature dependence of deviations in RMSFs. The best agreement between SBM simulations and the empirical force field (AMBER99) simulations is at a temperature of 90 reduced GROMACS units.

The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

of magnitude and facilitates extensive sampling. In Fig. 6.6 the sampled results of 180 free folding trajectories are presented: The normalized regional Q value is defined as the normalized number of formed base pairs within the respective helical substructure. The mean values and standard deviations of the normalized regional Q values for each helical stem are plotted over the total number of formed contacts during free folding. Mean value and standard deviation represent an approximate Gaussian distribution of normalized regional Q values within a contact bin, as can be seen in Sec. A.4. The SAM riboswitch exhibits immediate folding of helix P4, followed by P3 and P2 with relatively small difference. The non-local helix P1 constitutes the distinct end of the folding process by tying up both ends of the RNA strand, in agreement with earlier simulations [22]. The adenine riboswitch starts folding with helix P2, followed by P3 and concluded by the non-local helix P1.

For both riboswitches, 80 simulations at 12 extrusion rates each were performed in the course of this investigations. Folding transitions can be characterized by the number of formed helical base pairs at which the normalized regional Q value equals 0.5. The characterization of the regional folding characteristics condenses each substructural folding curve into a single value – the “mid Q value”. Mid Q values for all 12 extrusion rates and substructural elements are shown in Fig. 6.7 and give a clear-cut representation of the cotranscriptional RNA folding analysis. We see an influence of the extrusion rate on the folding order of substructural elements: In the SAM riboswitch, the folding order of P2 and P4 is reversed over range of rates between 100 and 200 nt/s. The formations of P2 and P3 in the SAM riboswitch and in the adenine riboswitch are simultaneous for a wide range of rates. Folding of non-local helix P1 is independent of the extrusion rate in both riboswitches. In both riboswitches, transitions in the folding order occur at extrusion rates (more than 100 nt/sec) that are beyond the range of physiologically relevant transcription rates. Based on that observation, the free folding case can be distinguished from the transcription rate limited case in the SAM-I riboswitch. In the *add* adenine riboswitch free folding and cotranscriptional folding exhibit the same characteristic folding order. As a result of this study, riboswitch folding can be described as robustly transcription rate limited in the range in physiologically relevant transcription rates between 15 and 80 nt/s [35].

6.3.1. Comparison with Kinectic Monte Carlo Results

In addition, I collaborated closely with a research group at the Max Planck Institute of Colloids and Interfaces in Golm that studied the cotranscriptional formation of secondary structure using kinetic Monte Carlo (MC) simulations. Unlike earlier MC studies [140, 141], but similar to the SBM, Michael Faber and Stefan Klumpp employ their recently proposed approach that incorporates native secondary structural information into MC simulations [32]. The free energy of a conformation is determined by empirically parameterized values for secondary structural motifs present in a given native secondary structure, as presented in Sec. 3.4. The kinetic MC simulation scheme is based on elementary steps that consist of the opening and closing of individual base pairs. Again, two different simulation setups are employed: First, the dynamics of free folding of the full-length chain is simulated. Secondly, cotranscriptional riboswitch folding is emulated by sequentially splitting up the RNA into a growing free and a diminishing confined part. RNA transcription corresponds, therefore,

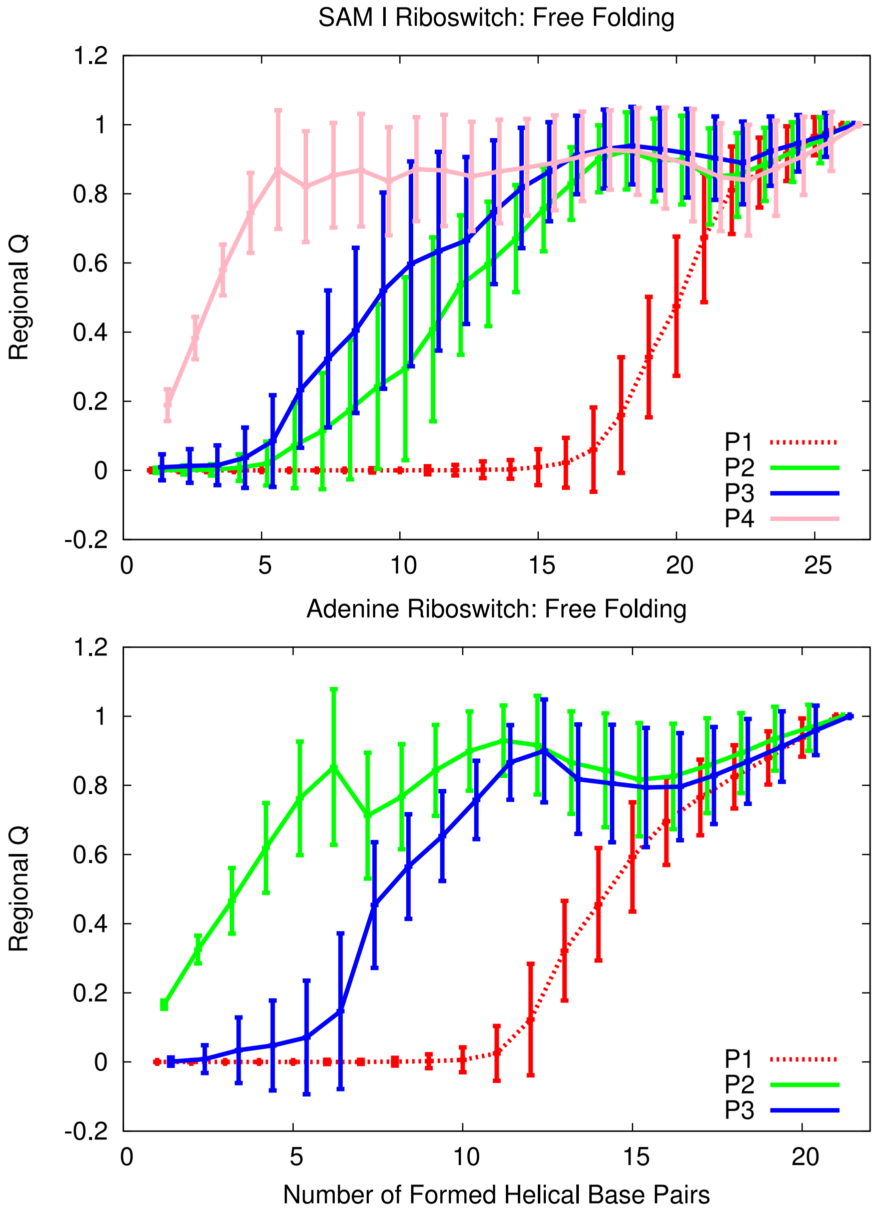


Figure 6.6.: Free folding characteristics of the SAM-I (top) and *add* adenine (bottom) riboswitches in SBM simulations. The evaluation is based on 180 free folding trajectories for each riboswitch. The mean value and standard deviation of the normalized regional Q values are plotted for each bin of number of formed helical base pairs. In the SAM ribswitch the helix formation order is: P4, P2 and P3 almost indistinguishable, P1. Accordingly for the adenine riboswitch: P2, P3, P1.

The figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

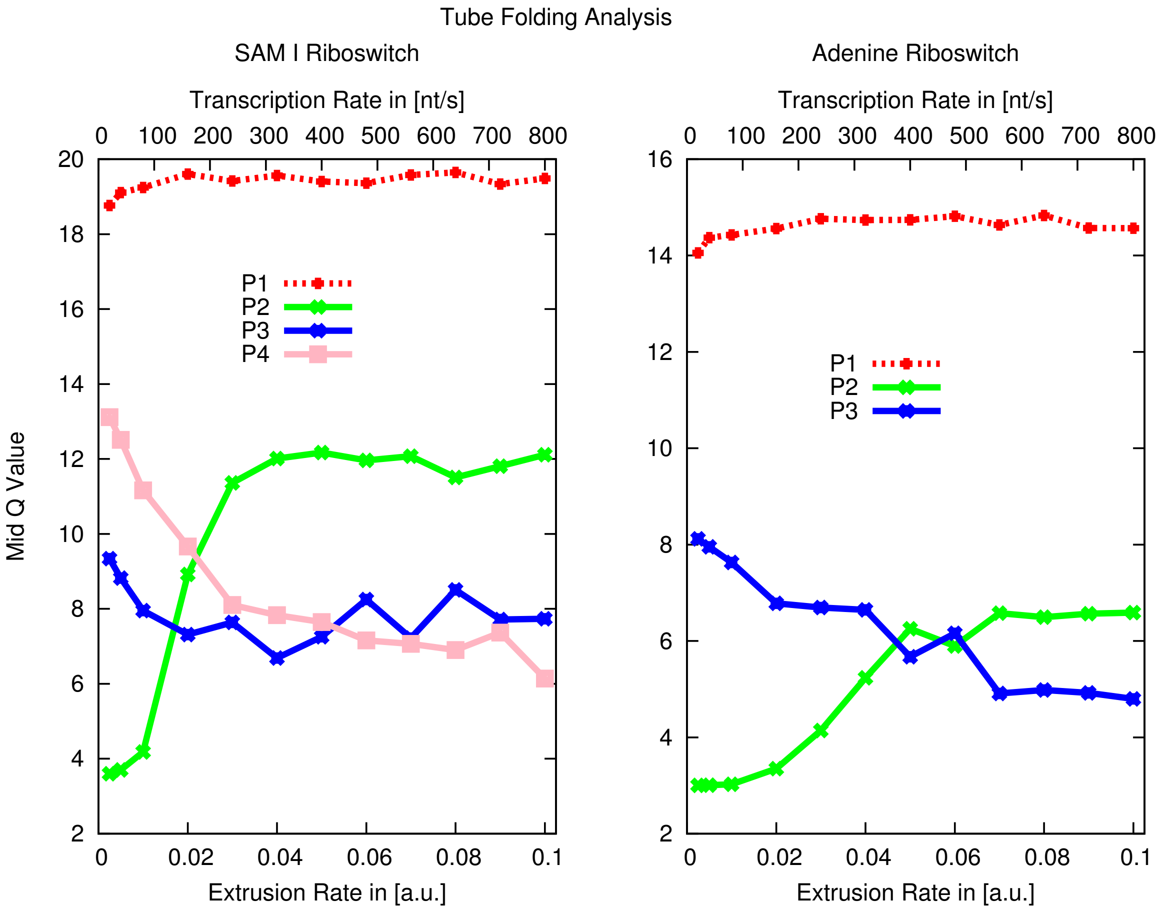


Figure 6.7.: Cotranscriptional folding characteristics of the SAM-I (left) and *add* adenine (right) riboswitches in SBM simulations. The various extrusion rates used in the simulations on the bottom axis are translated into transcription rates in nt/s on the top axis. 80 trajectories have been generated for each extrusion rate and each riboswitch. Each folding characteristic is condensed into a set of four or three mid Q values (one for each helix, depending in the riboswitch). Riboswitch folding is robustly transcription rate limited for transcription rates between 15 and 80 nt/s (physiological domain [35]).

The figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

to step-by-step shifting the boundary between these two parts. Formation of a base pair is only possible if both bases are already free, while those that are still confined are not taken into account for base pair formation.

Free folding simulations start from a conformation without formed base pairs. In both riboswitches secondary structure formation follows a distinct order. The folding order in the SAM riboswitch is the same as in SBM simulations: P4 folds first, followed by P3 and P2. The non-local P1 folds up both ends of the strand when all other secondary structure motifs are formed. Similarly, helices P2 and P3 first in the adenine riboswitch, followed by the non-local helix P1. While in kinetic MC simulations the two helices P2 and P3 fold simultaneously, P2 appears to fold slightly before P3 in SBM simulations. Tertiary interactions, which are not modeled in the kinetic MC approach, seem to cause this minor difference in folding order of free folding.

The simulations with RNA chain growth start with a single free base and add nucleotides of the RNA sequence to the free part at a given chain growth rate. The chain growth rates are varied over a wider range than in SBM simulations since the kinetic MC simulations are computationally much less expensive than the respective SBM simulations. Fig. 6.8 shows the mid Q values that characterized a helix folding event for each simulated chain growth rate. Small chain growth rates let the secondary structure elements form in the order of their appearance (P2, P3, P1 for the adenine riboswitch and P2, P3, P4, P1 for the SAM riboswitch). This order is different from the one observed in the free folding case. The folding orders in dependence of the extrusion or chain growth rate correspond qualitatively between SBM and kinetic MC simulations.

Similar to the procedure used in SBM simulations, the time scale is introduced via comparison of folding times with experimental measurements. The experimental folding times of a adenine riboswitch [134] are related to folding time histograms of kinetic MC simulations. Eventually, the used chain growth rates, given in nt per MC step, can be translated to nt/s and thereby compared to SBM results, which yields quantitative agreement. However, the transition rate that separates transcription rate limited folding from free folding is shifted to higher transcription rates compared to SBM simulations ($\simeq 0.1$ nt/MC step or 10^4 nt/s). This difference is discussed in Sec. 6.4. In both riboswitches, substructures fold limited by the transcription rate within the whole range of physiologically relevant transcription rates.

6.3.2. Competing Conformations and Ligand Model in Kinetic MC

Within the kinetic MC simualtion scheme it is directly possible to allow competing base pair formation steps by introducing them in the acceptance condition. The competing conformations of the SAM-I and the *add* adenine riboswitch are shown in detail in Fig. 6.9.

The competing base pairs are included in the decision process at each MC step, weighted by their free energy benefits. Time-resolved, normalized regional Q value trajectories alow the observation of “back tracking” of helix P1, as seen in Fig. 6.10. As long as the hairpin stem of the antiterminator (AT) helix are not available, the non-local helix P1 can start to fold. In the end, helix AT wins against P1 since it is overall energetically favorable, which is reproduced by the kinetic MC simulation scheme. This observation corresponds to the experimentally measured behavior of this riboswitch. We propose a ligand model

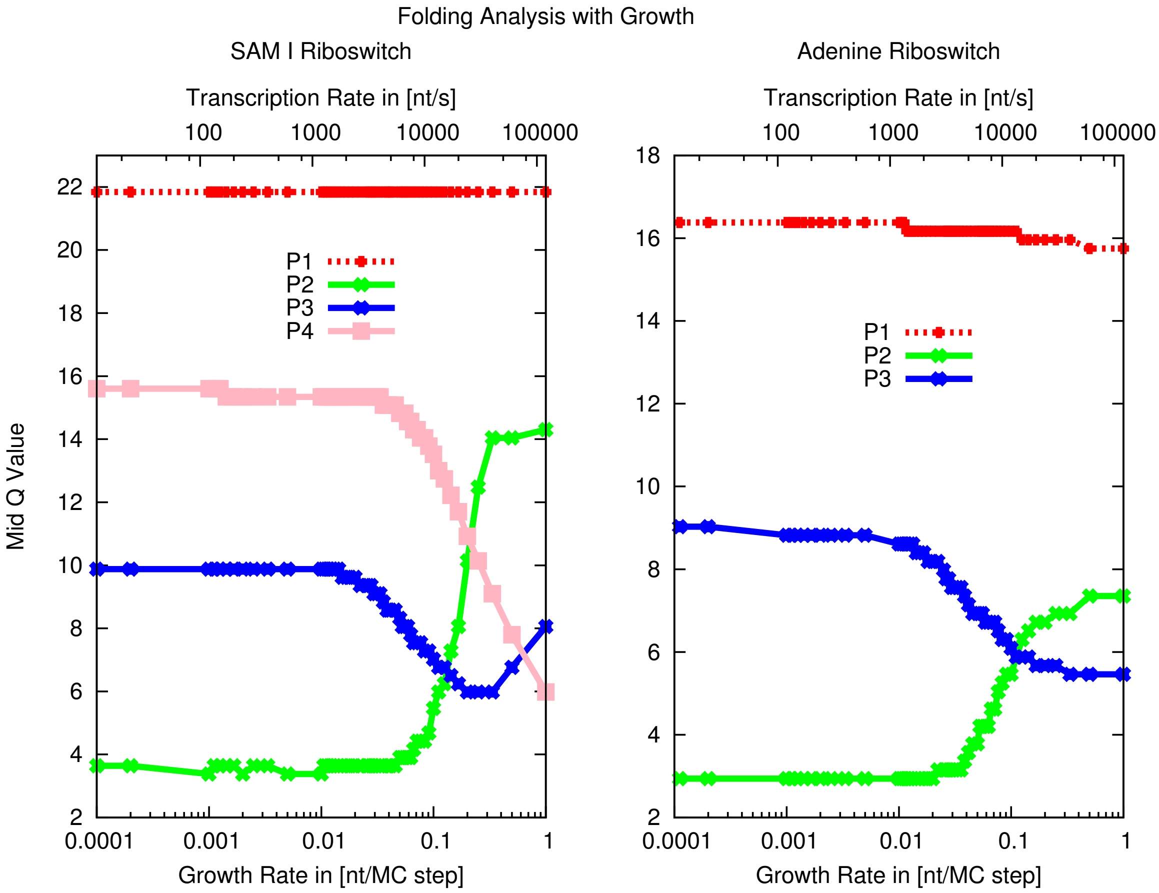


Figure 6.8.: Comparison of cotranscriptional folding in kinetic Monte Carlo simulations. The simulation data was generated by Michael Faber at the Max Planck Institute of Colloids and Interfaces, Golm. The various chain growth rates used in the simulations on the bottom axis are translated into transcription rates in nt/s on the top axis. The two plots correspond qualitatively to the characterisitcs depicted in Fig. 6.7.

The figure is taken from [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

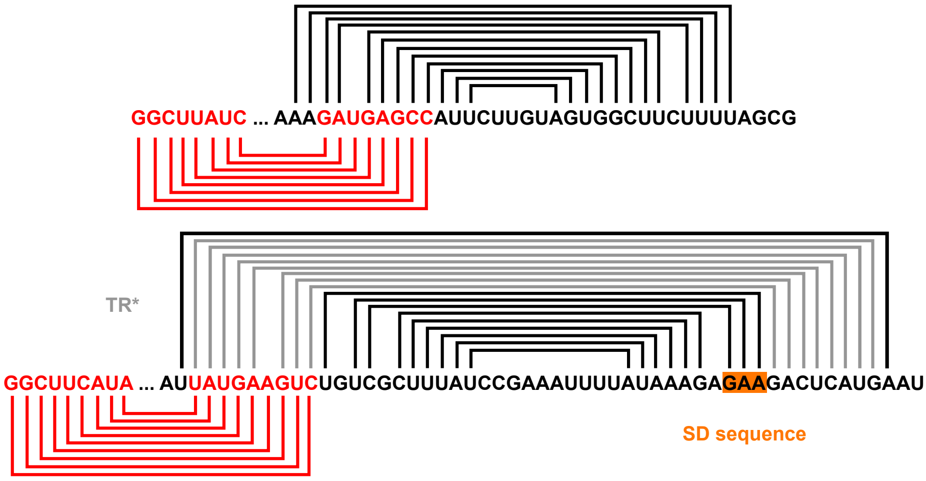

GGCUUAUC ... AAAGAUGAGCCAUUCUUGUAGUGGCUUCUUUUAGCG
TR*
GGCUUCAUA ... AUUAUGAAGUCUGUCGCUUUAUCCGAAAUUUUAUAAAGAGAAGACUCAUGAAU
SD sequence

Figure 6.9.: Secondary structure schematics of the SAM-I (top) riboswitch and the *add* adenine (bottom) in their respective ligand-free conformations. TR* denotes the part of the translational repressor helix that competes directly to helix P1.

The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

that also reproduces the experimentally determined behavior of a winning P1 helix due to the presence of SAM metabolites. The stabilizing effect of a ligand is realized by lowering the multi-loop penalty of helix P1. Varying this energetic contribution yields a minimal stabilizing free energy benefit of about 7 $k_\mathrm{B}T$ in the SAM riboswitch and 2 $k_\mathrm{B}T$ for the adenine riboswitch.

6.4. Discussion

The presented computational results originate from a novel biophysical approach to model cotranscriptional riboswitch folding. The investigated biomolecular processes of transcription, ligand binding, and folding take place interdependently during a period of time in the order of seconds. Therefore, the demands of standard MD simulations of riboswitch folding exceed today's computational capabilities by several orders of magnitude [11]. SBM implementations offer computationally acquirable simulations that yield atomically resolved dynamic trajectories of riboswitch folding. An implicit ligand model emulates the sensing conformations of the aptamer for the ligand and ligand binding. SBM simulations are performed at a temperature below the reference temperature derived from a simulation based on the AMBER99 force field at 300 K. The lower temperature accelerates folding by about one order of magnitude and reduces computational effort in favor of enhanced sampling. The extensive sampling guarantees reliable results based on averaged statistical events.

The studied ranges of transcription rates exceed the physiological transcription rates re-

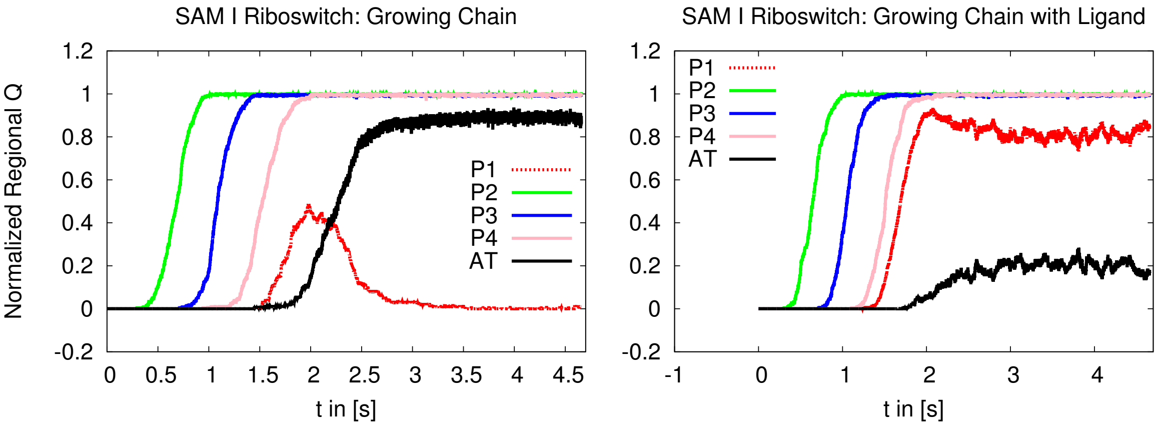


Figure 6.10.: Competing conformations and ligand model for the SAM-I riboswitch in kinetic MC simulations performed by Michael Faber. This figure displays two time-resolved trajectory averages of SAM-I riboswitch folding. Left: During transcription, helix P1 starts folding before the antiterminator (AT) sequence is available. Without the stabilizing influence of the SAM ligand, the non-local helix “back tracks” in favor of the physiologically expected AT helix. Right: In the case of cotranscriptional riboswitch folding the presence of a ligand stabilizes the formation of helix P1. The stabilizing effect of the ligand is modeled by lowering the multiloop penalty for non-local helix P1 by 12 $k_{\mathrm{B}}T$ in the presented simulation.

The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

alized by RNAP. SBM simulations yield consistent results with a complementary numerical kinetic MC method for cotranscriptional folding of riboswitch aptamer structures for both extreme cases, i. e., transcription-rate limited and free folding. However, the two models differ quantitatively regarding the value of the transcription rate at which the folding order changes between characteristic behavior of free and cotranscriptional folding. SBM simulations predict lower rates than kinetic MC simulations. This quantitative discrepancy reflects a difference in the folding times of the helical substructures in riboswitches, although the simulated time scales have been introduced in the same way in both methods: In SBM simulations, but not in kinetic MC simulations, folding of the individual hairpins is slowed down by the drag of the adjoining single-stranded regions, which is discussed in more detail in Sec. A.5. In both models, however, the transition between both scenarios occurs robustly at high transcription rates greater than 100 nt/s which are not realized in nature.

The kinetic MC method is able to model competing conformations by design. Respective simulations have been performed by Michael Faber to validate the robustness of the kinetic MC method regarding this extension. The results support the picture of a structure that has a ligand-free conformation that wins against the ligand-bound conformation if no ligand is present. The jointly developed ligand binding model in the kinetic MC framework successfully describes the influence of a ligand. The derived energetic value that causes a stable conformational change agrees with the one recently determined by a standard MD study of the *add* adenine riboswitch [142].

During transcription a ligand can bind to the aptamer region of the riboswitch and influence the transcription termination or translation attenuation decision executed by the riboswitch. This decision has to be made in a certain time range in order to be effective. Whether this time range is sufficient for the aptamer binding site to reach thermodynamic equilibrium with the metabolite defines whether a riboswitch is under kinetic or thermodynamic control. Experimental measurements typically determine dissociation constants of ligand binding *in-vitro* and compare them to metabolite concentrations that are responsible for 50% termination efficiency *in-vivo* [143, 144]. These measurements are quite rough keeping the complexity of the involved mechanisms in mind. Depending on the transcription progress, the riboswitch offers the metabolite a "growing" binding site with a time-dependent binding affinity. A binding pocket composed by substructures of the aptamer close to the 5'-end effectively increases the time window available for ligand binding compared to a binding pocket with significant substructures close to the 3'-end. Considering transcription times of typically seconds for both the aptamer and expression platform, this accounts to a significant difference compared to a fully transcribed free-folding riboswitch. Simulations complement this picture by atomistic insights regarding folding paths and dynamic behavior.

Riboswitches – structured RNA in the 5' UTR of mRNAs – regulate gene expression of genetic information contained in mRNAs on a transcriptional or translational level. A deeper understanding of possible influences on the involved mechanisms will facilitate insights in gene regulation, evolutionary processes or the RNA world hypothesis [145]. I employ a protocol based on a coarse-grained approach that emulates spatial constraints of the RNAP and sequential release of the riboswitch aptamer region: A homogenized, minimally frustrated force field based on the systems native tertiary structure. Michael Faber, on the other hand, proposes a kinetic Monte Carlo method based on base pair formation weighted by free energy benefits determined from native secondary structure [32]. A comparison between the two complementary models as part of our collaboration yields agreeing results that back the significance of the presented findings. The results gained by both numerical approaches give a more detailed insight in cotranscriptional riboswitch folding. Thereby, the folding appears to occur in a transcription-rate limited order possibly different from free folding. The emerging picture of cotranscriptional riboswitch folding is surprisingly simple and concurs with recent experimental findings in single molecule measurements [39].

An explicit ligand representation in my model and the implementation of the mutually exclusive structures in the unbound state will be future refinements to the model. Moreover, the SBM introduces all native contacts by Lennard-Jones terms and assumes stabilizing ion concentrations that were present during crystallization of the underlying PDB structure. The implicitly modeled electrostatics prohibit the investigation of different ion concentrations on cotranscriptional riboswitch folding. In order to study the influence of varying ion concentrations, the SBM would have to be expanded by electrostatics terms, as recently presented in Debye-Hückel approximation [146, 147].

7

Chapter 7.

Advances to RNA Structure Prediction

This chapter gives an overview of the results acquired in the course of my studies on a novel approach to RNA structure prediction based on direct coupling analysis (DCA) contact predictions. The studies have been conducted in a close collaboration with M.Sc. Eleonora De Leonardis in the group of Prof. Martin Weigt at the Université Pierre et Marie Curie Paris, France.

Structured RNA fulfills various catalytic or regulatory functions in cells. Since the discovery of base pairs and the double helix of DNA by Watson and Crick in the 1960s, there have been tremendous advances in experimental structure measurements both in DNA and RNA. The strong link between structure and function motivates the refinement of structure resolution methods and simplification of the procedures.

The automatization of experimental RNA sequencing methods resulted in vast databases of RNA sequences over a huge variety of organisms and families. The available amount of data facilitates research in the fields of bioinformatics and statistical physics to look for correlations or characteristic motifs. Since there is a growing gap between the number of sequences available and the number of experimentally resolved RNA structures, a method that exploits sequence information for structure prediction is a striking ansatz. DCA is a method that originated in the field of protein contact prediction and is adapted to RNA contact prediction in the course of this study.

Contact information predicted by DCA is translated into interatomic Lennard-Jones potential terms and added to a knowledge-based model (KBM) force field for bonded and non-bonded interactions. The subsequent folding simulations yield ensembles of stable conformations with deviations from the experimentally measured reference structures that are comparable to other state-of-the-art RNA structure prediction approaches.

7.1. Motivation and Overview

Similar to proteins as the more investigated biomolecular systems, structured RNA systems are subject to the dogma that there is a link between their structures and their functionalities. Therefore, the thorough investigation and resolution of RNA structures can give some indication of their physiological functions. The first experimentally investigated structured RNA systems were the transfer RNA (tRNA) [2, 3] in the early 1970s. tRNA delivers amino acids, the building blocks of proteins, to the complementary RNA codons in the ribosome. Thus, tRNA structure features a part that has selective affinity for a specific amino acid and a part that contains the corresponding codon. More recent examples are ribozymes [4] – catalytic elements involved in the biosynthesis of RNA or RNA splicing – or riboswitches that are discussed in more detail in Sec. 1.4 and Chap. 6. The functions of structured RNA that are crucial in the complex network of biomolecular processes, are of high interest but their experimental investigations are scientifically challenging and costly. On the other hand, tremendous progress has been made in DNA and RNA sequencing techniques allowing fast and automated determination of sequence information or "primary structure" [148]. Therefore, it is desired to develop analytic and numerical approaches that are able to predict RNA structures based on available sequences. The established focus so far has been on secondary structure prediction. The methods used for secondary structure prediction are based on statistical analysis methods [90, 91] or dynamic programming approaches [89]. Statistical methods in this field are not based on a single sequence but on multiple sequence alignments (MSA) of RNA families that are experimentally accessible due to the before mentioned automatized sequencing techniques. The most comprehensive collection of such sequence families are published in the RFAM database [68] where the sequences are deposited as MSA data sets. A great challenge to the present day is the prediction of tertiary structure in RNA. Very recent approaches in this field of research are template based implementations that assemble sequence motifs by homology modeling algorithms, e. g., ROSETTA [149]. Progress that has been achieved so far has been compiled recently in a publication by several groups [150].

The field of statistical analysis methods in the context of RNA structure prediction has focused in the past on mutual information analysis for secondary structure prediction. The mutual information (MI) is defined as

$$\mathrm{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \left(\frac{f_{ij}(A,B)}{f_i(A) f_j(B)} \right) , \tag{7.1}$$

where $f_i(A)$ and $f_j(B)$ are the single site frequency counts of nucleotides A,B at positions i,j and$f_{ij}(A,B)$ are the pair frequency counts, as introduced in Sec. 4.1.2. If the pair frequency factorizes into the respective single site frequencies, MI equals to zero and is positive otherwise. This method is able to predict canonical base pairs in MSAs reliably because, apparently, there is a strong correlation in mutations of nucleotides participating in canonical base pairs [85]. Therefore, secondary structure prediction can be considered as a solved problem and consensus secondary structure information is already provided in the databases together with the corresponding MSAs. Going beyond the analysis of apparent statistical couplings, the direct coupling analysis (DCA) disentangles direct couplings from the whole set of correlations. For tertiary contacts, this disentanglement is necessary to distinguish direct couplings that correspond to spatial closeness from indirect couplings that comprise any kind of mediated correlations. DCA has become an established method in protein contact prediction [25] or protein complex formation [105]. In the course of a close collaboration with M.Sc. Eleonora De Leonardis in the group of Prof. Martin Weigt at the Université Pierre et Marie Curie Paris I worked on the adaption of the DCA implementation to process RNA alignments.

A smaller alphabet for RNA sequences (4 nucleotides and one gap) compared to proteins (20 amino acids and one gap) yields more frustration and RNA conformations feature different structural motifs. After the method is adjusted to the different biomolecular system respective predictions can be generated. The predictions contain on top of secondary structure information also a substantial amount of tertiary contact predictions. Tertiary structure predictions, however, have a lower signal than the secondary structure predictions and are disturbed by underground noise in the DCA scores. Therefore, the quality of true positive predictions and the influence of false positive predictions need to be assessed in order to be able to improve the reliability of this new approach. To this end, I compile a corresponding list of RNA sequence families and experimental PDB (Protein Data Bank [42]) structures that meet certain requirements. In this defined “gold standard”, predictions can then be used as constraints for KBM simulations and the outcome of respective folding simulations can be compared to the PDB structures. The KBM force field uses cataloged values for the quantities of equilibrium instead of values taken from a native fold. I create a catalog based on studies of a training set, the SAM-I riboswitch (PDB ID 2GIS). The catalog features default values and values that depend on the consensus secondary structure elements provided together with the MSAs in RFAM. Four systems are then investigated by SBM simulations in the generalized formulation: SAM-I riboswitch (PDB ID 2GIS), *add* adenine riboswitch (PDB ID 1Y26), glycine riboswitch (PDB ID 3OWI) and fluoride riboswitch (PDB ID 3VRS). I present the results of four different contact scenarios:

a) only consensus secondary structure

b) consensus secondary structure and the 50 highest ranked true positives

c) consensus secondary structure and all available true positives

d) consensus secondary structure and the 100 highest ranked predictions (true and false positives)

I show that all scenarios that are based on tertiary structure predictions (b, c, d) decrease the average RMSD values of the folded ensembles in simulations compared to the plain secondary information scenario (a). The presented proof of principle can also quantify the influence of false positive disturbances in a set of predicted contacts.

7.2. Method

The research protocol that facilitates the investigation of contact predictions in structured RNA is introduced in the following. First, the adjustments necessary to use the original DCA formulation are discussed and the steps towards a comparison with PDB structures are described. To this end, I collect a "gold standard", i. e., a list of highly resolved PDB structures that correspond to RNA families in the RFAM database with sufficient sequence data. The emerging coevolutionary RNA contact predictions can be assessed by evaluating a true positive rate relative to the "true" PDB structure.

Instead of a criterion that is purely based on DCA predictions compared to the respective PDB contact map I present a technique that introduces contact predictions to a simplistic *de-novo* force field. The construction of this KBM force field is described in detail. Bonded and non-bonded interactions are introduced by an analysis of a learning set and then transfered to systems of interest. Subsequent SBM simulations yield folding trajectories that provide ensembles of folded conformations based on DCA predictions for non-bonded interactions. The root mean-square deviations (RMSDs) of the ensembles with respect to the native fold represent a measure to evaluate the quality of a predicted contact map.

7.2.1. Gold Standard to Evaluate Prediction Quality

In order to assess the quality of DCA predictions it is necessary to have a "gold standard" of "true" structures with sufficient respective sequence data. To this end, I compile a list of PDB structures whose sequences have a best match overlap with their respective RFAM MSAs, as shown in Tab. 7.1. PDB structures and RFAM sequence families have to meet a list of requirements:

- experimental X-ray diffraction crystal structure with less than 0.4 nm resolution

  The PDB contains several types of experimentally determined structures by techniques such as X-ray crystallography, NMR, neutron scattering and cryo-electron microscopy. X-ray structures provide the best resolved structures and are therefore a requirement in this study.

- non-trivial

  Some of the families contained in the RFAM database are single hairpin structures. These structures are considered as trivial and therefore disregarded in this study.

- monomeric

  Some of the families contained in the RFAM database represent single strands of dimeric RNA structures. The investigation of polymeric complexes has additional challenges and is outside the scope of this study.

- less than 1500 nucleotides

RFAM ID	M_{eff}	PDB ID	Chain
RF00001	57991.33	3CC2	9
RF00017	8145.33	1L9A	B
RF00059	3347.90	2HOJ	A
RF00504	1828.54	3OWI	A
RF00010	2309.67	1U9S	A
RF00023	2143.30	4ABR	Y
RF00162	1165.56	2GIS	A
RF00050	1045.85	3F2Q	X
RF02001	636.43	3BWP	A
RF00167	588.88	1Y26	X
RF00168	552.38	3DIL	A
RF01051	983.21	3IRW	R
RF00380	206.70	2QBZ	X
RF01734	532.03	3VRS	A

Table 7.1.: Gold standard of structured RNA families. The table contains a list of 14 RNA families that are suitable for a structural comparison. The listed families have published X-ray crystal structures with a resolution of less than 0.4 nm, are neither trivial nor dimeric and their alignments contain at least 1000 sequences. This list is sorted by the number of sequences in RFAM database, version 11.0 [68]. The families are indicated by their RFAM ID. M_{eff} denotes the effective number of independent sequences, as discussed in Sec. 4.1.1. The respective structure deposited in PDB is given by its structure ID and a chain ID.

The current formulation of the DCA implementation reaches memory limits with structures bigger than 1500 nucleotides.

- more than 1000 sequences or $M_{\text{eff}} > 200$, respectively

The DCA requires a minimal number of independent sequences to provide reliable statistical results.

The effective length of each sequence, as discussed in Chap. 4, is given by

$$M_{\text{eff}} = \sum_{a=1}^{M} \frac{1}{m^a}\,, \tag{7.2}$$

where we define the number of similar sequences by

$$m^a := \Big|\{b \mid 1 \leq b \leq M, \text{seqid}(A^a, A^b) \geq xL\}\Big| \tag{7.3}$$

for a given similarity ratio $x = 0.8$.

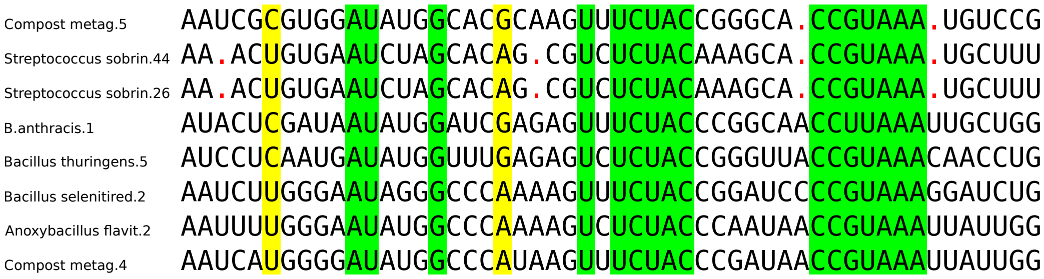

Compost metag.5 AAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCA.CCGUAAA.UGUCCG
Streptococcus sobrin.44 AA.ACUGUGAAUCUAGCACAG.CGUCUCUACAAAGCA.CCGUAAA.UGCUUU
Streptococcus sobrin.26 AA.ACUGUGAAUCUAGCACAG.CGUCUCUACAAAGCA.CCGUAAA.UGCUUU
B.anthracis.1 AUACUCGAUAAUAUGGAUCGAGAGUUUCUACCCGGCAACCUUAAAUUGCUGG
Bacillus thuringens.5 AUCCUCAAUGAUAUGGUUUGAGAGUCUCUACCGGGUUACCGUAAACAACCUG
Bacillus selenitired.2 AAUCUUGGGAAUAGGGCCCAAAAGUUUCUACCGGAUCCCCGUAAAGGAUCUG
Anoxybacillus flavit.2 AAUUUUGGGAAUAUGGCCCAAAAGUCUCUACCCAAUAACCGUAAAUUAUUGG
Compost metag.4 AAUCAUGGGGAUAUGGCCCAUAAGUUUCUACCCGAUAACCGUAAAUUAUUGG

Figure 7.1.: MSA excerpt of RFAM family RF00167 (*add* adenine riboswitch). Each row of the alignment represents an instance of the given family found in a specific organism (on the left). The red dots (.) denote gaps that are introduced in the aligned sequence with a certain penalty to match the other sequences as well as possible. The green columns denote conservation of nucleotides in the sequences that are often present in regions of high functional importance. The two yellow columns denote commutations of nucleotides that can indicate correlation and spatial closeness. To decide, whether this apparent commutations have direct or indirect causes, is the task of statistical algorithms, such as the DCA method.

7.2.2. Coevolutionary RNA Contact Prediction

The coevolutionary approach to RNA contact prediction is based on the technique described in detail in Sec. 4.2. Published formulations of this approach have been applied in the field of protein structure prediction. The established DCA approach has been successfully employed for, e. g., contact prediction in proteins [25] and protein-protein complexes [105]. RNA structures have different challenges for the statistical method. The "alphabet" or number of states in the Potts model is reduced to 5 compared to 21 for proteins. This reduction yields a drastically increase in frustration at each position of the sequence. The number of alternatives in case of a mutation is very small and thus the probability of random commutations comparably high. In the course of a close collaboration with M. Sc. Eleonora De Leonardis during my visit in the group of Prof. Martin Weigt at the Université Pierre et Marie Curie Paris I helped to adapt a respective implementation to be able to treat RNA multiple sequence alignments, as shown in Fig. 7.1.

In a first step, all columns that contain more than 50% gaps are removed from the alignment to reduce the alignment to relevant entries. Key to the statistical method, as discussed in Chap. 4, are the redefined frequency counts

$$f_i(A) = \frac{1}{\lambda + M_{\text{eff}}} \left(\frac{\lambda}{q} + \sum_{a=1}^{M} \delta_{A,A_i^a} \right) , \tag{7.4}$$

$$f_{ij}(A,B) = \frac{1}{\lambda + M_{\text{eff}}} \left(\frac{\lambda}{q^2} + \sum_{a=1}^{M} \delta_{A,A_i^a} \delta_{B,A_i^a} \right) , \tag{7.5}$$

where we introduce the pseudo-count $\lambda = M_{\text{eff}}$ and the effective number of independent sequences M_{eff}, see Eq. (7.2). The score obtained by a simple mutual information analysis needs be refined by disentangling direct and indirect correlations. To this end, we look for a statistical model for the entire RNA sequence that reproduces the empirical frequency counts as marginals. The most general model is revealed by the principle of maximum entropy. This principle has the known solution

$$P(A_1, \dots, A_L) = \frac{1}{Z} \exp\left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i) \right) , \tag{7.6}$$

where $e_{ij}(A, B)$ are pairwise couplings and $h_i(A)$ are local fields. In this study

$$e_{ij}(A, q) = e_{ij}(q, A) = h_i(q) = 0 \tag{7.7}$$

is the used gauge where couplings and fields are considered relative to the last ribonucleic acid $A = q = 5$. Based on a mean-field approximation $e_{ij}(A, B)$ and $h_i(A)$ can be determined. The score value for DCA predictions is the direct information (DI) as introduced in Sec. 4.2.5:

$$\text{DI}_{ij} = \sum_{A,B=1}^{q} P_{ij}^{\text{dir}}(A, B) \ln \left(\frac{P_{ij}^{\text{dir}}(A, B)}{f_i(A) f_j(B)} \right) , \tag{7.8}$$

where the direct statistical model

$$P_{ij}^{\text{dir}}(A, B) = \frac{1}{Z_{ij}} \exp\left(e_{ij}(A, B) + \tilde{h}_i(A) + \tilde{h}_j(B) \right) \tag{7.9}$$

and the frequency counts

$$f_i(A) = \sum_{B=1}^{q} P_{ij}^{\text{dir}}(A, B) \quad , \quad f_j(B) = \sum_{A=1}^{q} P_{ij}^{\text{dir}}(A, B) \tag{7.10}$$

are contained.

The general challenge is to distinguish the very high signal for secondary structure contacts (Watson-Crick and Wobble base pairs) from tertiary structure contacts with signals that are only slightly above the noise level in the DI signal.

In order to compare the DCA predicted contact map to a "true" contact map read from a PDB structure it is necessary to align the PDB sequence to its best match in the MSA. This alignment is performed by a Smith-Waterman algorithm [88] that was chosen over the Needleman-Wunsch algorithm [87] and is used in the MATLAB workflow of the DCA software package. DCA predictions are then generated by the MATLAB implementation that is distributed by the group of Prof. Martin Weigt at the Université Pierre et Marie Curie Paris [25].

7.2.3. Knowledge-based Model Force Field

The general idea is to translate DCA residue contact prediction for structured RNA into interatomic contacts, combine it with a knowledge-based model (KBM) force field for bonded interactions in RNA. Subsequent folding simulations are expected to yield stable conformations as close to the experimentally measured structures as possible. The magnitude of the deviations from the native structure can be used to assess the quality of the initial contact predictions by DCA.

The KBM force field follows the parametrization scheme of regular SBMs but does not take the quantities of equilibrium from a given structure. The quantities of equilibrium are instead taken from a catalog that is generated from a combination of learning sets. In the following I present the different steps necessary to generate the catalog and setup a respective force field for a given RNA sequence and consensus secondary structure.

The potential energy of SBM simulations reads as

$$\begin{aligned} V = & \sum_{\text{bonds}} K_{\mathrm{b}}(r - r_0)^2 + \sum_{\text{angles}} K_{\mathrm{a}}(\theta - \theta_0)^2 \\ & + \sum_{\text{improper}} K_{\mathrm{i}}(\chi - \chi_0)^2 + \sum_{\text{dihedrals}} K_{\mathrm{d}} f_{\mathrm{d}}(\phi) \\ & + \sum_{\text{contacts}} K_{\mathrm{c}} \left[\left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{12} - 2 \cdot \left(\frac{\sigma_{ij}^0}{r_{ij}} \right)^{6} \right] \\ & + \sum_{\text{non-contacts}} K_{\mathrm{nc}} \left(\frac{\tilde{\sigma}}{r_{ij}} \right)^{12} \end{aligned} \tag{7.11}$$

where the dihedral (or torsional) angle potential is given by

$$f_{\mathrm{d}}(\phi) = \left[1 - \cos(\phi - \phi_0)\right] + \frac{1}{2}\left[1 - \cos(3 \cdot (\phi - \phi_0))\right] \tag{7.12}$$

and K_{b}, K_{a}, K_{i}, K_{d}, K_{c} and K_{nc} are the corresponding force constants that are presented in more detail in Sec. A.1. $\tilde{\sigma}$ is a global exclusion radius and the parameters r_0, θ_0, χ_0, ϕ_0 and σ_{ij}^0 are taken from the native structure in its original formulation. The variation of this formulation that is necessary to be able to use this framework for *de-novo* folding simulations is to choose the parameters r_0, θ_0, χ_0, ϕ_0 and σ_{ij}^0 in a generalized or knowledge-based scheme.

First, bonded interactions need to be introduced in the KBM. The general idea is to analyze the learning set by histograms for all possible bonded interactions. The learning set for the bonded interactions in this study is the SAM-I riboswitch structure published in PDB ID 2GIS. The bond distances are very narrowly distributed and have standard deviations of less than 0.001 nm for bonds that range between 0.12 and 0.17 nm. Similarly, the angles are also very narrowly distributed with standard deviations of less than 6 degrees for angles that range between 101 and 131 degrees. The third class of bonded interactions that are very well defined by all appearances in the learning set are the planar dihedral angles that stabilze planarity in the rings of the bases. The angles are either 180 or 360 degrees (depending on the order of listed atoms defining the angle) with standard deviations of less than 3 degrees. Therefore, all bonds, angles and planar dihedral angles can be cataloged for an arbitrary structure, independently of their involvement in substructural elements. This can

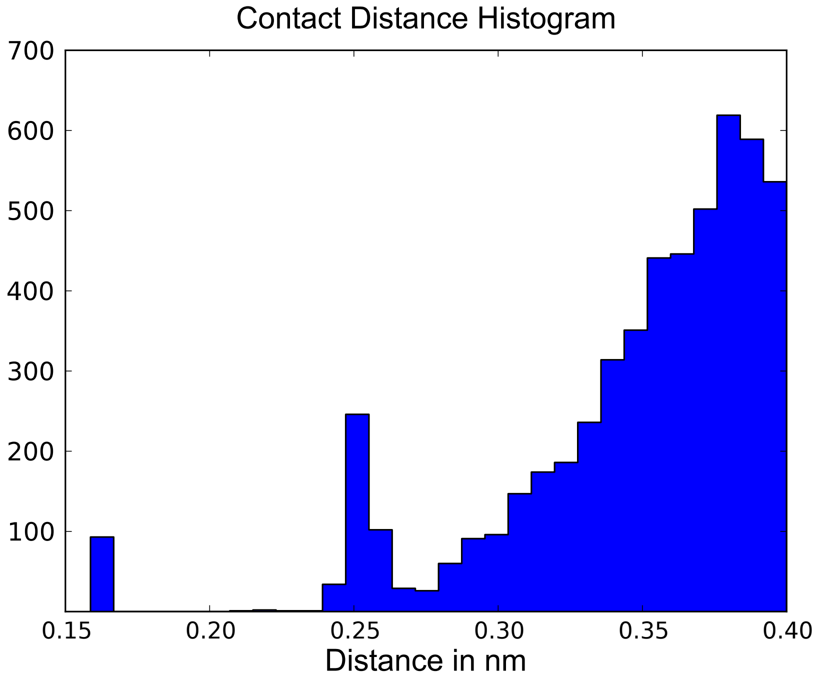


Figure 7.2.: Interatomic contact distance histogram in the SAM-I riboswitch. Two peaks in the distribution at 0.16 and 0.25 nm can be identified with contacts across the interface of two nucleotides along the backbone. The continuous distribution above these peaks needs to be resolved by more distinct analyses of specific structural motifs, such as base pairing or base stacking interactions.

be physically motivated by the fact that bond distances as well as angles are determined by the overlap between next-neighbor atomic orbitals and molecular rings are kept planar by the collective overlap of their constituents' orbitals.

The quantities of equilibrium for proper dihedral angles are dependent on their involvement in substructural elements. Dihedral angles are the defining geometrical values that describe a conformation within the range of its structural flexibility. The analysis of the learning set reveals a set of proper dihedral angles that are narrowly distributed in helical regions. I choose a cut-off of 9 degrees for the standard deviation of helical dihedral angles within a nucleotide and a cut-off of 17 degrees for the standard deviations of dihedral angles at the interface between two nucleotides. The set comprises 24 dihedral angles within nucleotides and 2 connecting dihedral angles. The required information for determining helical regions is the consensus secondary structure of the RNA that is stored in the database together with the MSA.

In a next step the non-bonded interactions need to be introduced. The analysis of the overall distribution of contact distances in the learning set is shown in Fig. 7.2. Contact distances along the back bone are the same for all nucleotides and correspond to the two sharp peeks in the histogram: C3' - P (0.16 nm) and O3' - {P, OP1, OP2, O5'} (0.25 nm).

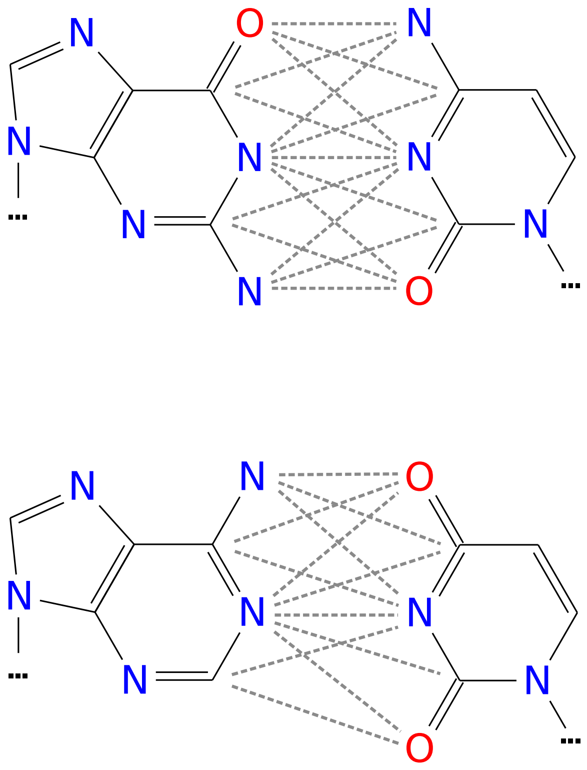


Figure 7.3.: Interatomic contacts in Watson-Crick base pairs. The contacts are determined by calculating the average atom-atom distances of 24 Watson-Crick base pairs in the SAM-I riboswitch. The average distances between atoms in the depicted base pairs are smaller than 0.4 nm with a standard deviation of 0.01 nm or less. This yields 15 contacts in the G-C (top) and 12 contacts in the A-U (bottom) base pair.

A continuum of contacts above the sharp peeks needs to be resolved by the analysis of explicit structural motifs. The two motifs that are included in the model are base pairing and base stacking contacts. Base pairing contacts, as shown in Fig. 7.3 and Fig. 7.4, are found in base pairs of the learning set with a cut-off of 0.4 nm. Base stacking contacts can be extracted from the learning set with a cut-off of 0.4 nm and symmetric constructions since not all 16 possible realizations of stacking base sequences are present in the learning set.

The bonded and non-bonded interactions of the KBM are compiled in an XML catalog and integrated in the eSBMTools implementation via the existing XML-based topology definition, as discussed in Sec. 5.2.1.

The remaining vital ingredient to the knowledge-based model is the incorporation of tertiary contact predictions from DCA. A residue-residue contact predicted by DCA needs to be mapped onto a set of corresponding atom-atom contacts. To this end, I choose a

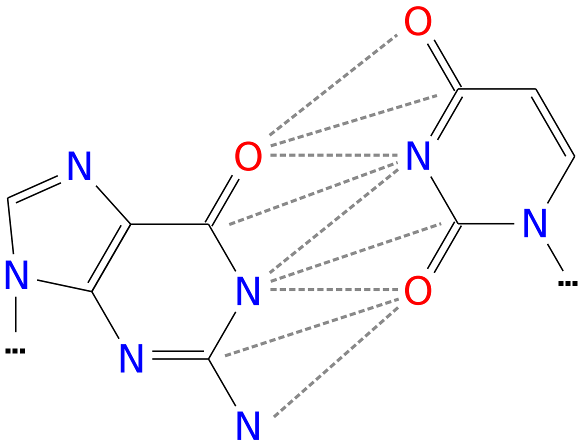


Figure 7.4.: Interatomic contacts in Wobble base pairs. The contacts are determined by calculating the average value of 2 Wobble base pairs in the SAM-I riboswitch. The average distances between atoms in the depicted Wobble base pair are smaller than 0.4 nm with a standard deviation of 0.01 nm or less. This yields 9 contacts in the G-U base pair.

set of typical RNA contact classes proposed by the research group of Eric Westhof [26]. This set classifies possible inter-nucleotide contacts by their relative base coordination and orientation and provides a collection of representative all-atom structures to determine characteristic atom-atom distances. For a predicted nucleotide-nucleotide contact announced by DCA, all respective RNA structures in the set are analyzed and the averages and standard deviations of according atom-atom distances are calculated. A list of atom-atom contacts with mean values less than 0.6 nm and standard deviations less than 0.3 nm is then included as all-atom Lennard-Jones contacts in the knowledge-based model.

The required model is now complete and features the following potential terms:

- cataloged values for all bonds, angles and planar dihedral angles
- cataloged values for proper dihedral angles in helical regions
- cataloged values for base stacking and base pairing contacts
- mean values of atom-atom distances (based on a set of characteristic structures [26]) for a given nucleotide-nucleotide contact

The generalized formulation of the potential in the SBM framework enables atomically resolved and computationally affordable simulations. Therefore, RNA folding can be observed and extensive sampling allows investigation of the folded ensembles.

RFAM ID	PDB ID	Organism
RF00001	3CC2	Haloarcula_ marismort.2
RF00017	1L9A	M.murinus.681
RF00059	2HOJ	Escherichia_ coli_E11.2
RF00504	3OWI	Vibrio_cholerae_bv._.2
RF00010	1U9S	Thermus_thermophilus.2
RF00023	4ABR	marine_metag.496
RF00162	2GIS	Thermoanaerobacter_t.3
RF00050	3F2Q	Fusobacterium.18
RF02001	3BWP	Oceanobacillus_iheye.5
RF00167	1Y26	Vibrio_vulnificus_CM.1
RF00168	3DIL	Thermotoga_maritima_.2
RF01051	3IRW	Vibrio.6
RF00380	2QBZ	Bacillus_pumilus_SAF.1
RF01734	3VRS	Thermotoga.2

Table 7.2.: Gold standard of structured RNA families with their best matches to the corresponding PDB sequences. The PDB sequences are compared to the sequences in the RNA family alignment by a standard Smith-Waterman algorithm [88]. The found sequences and their indicated organisms of origin correspond to the organisms given in their PDB entries, respectively.

7.3. Results

In the following section I compile the outcomes of my studies conducted in the context of DCA contact prediction for structured RNA. The results consist of two categories. First the general analysis of structures in the gold standard is presented. This analysis already demonstrates capability of DCA to discover secondary structural elements and tertiary contacts. Secondly, the incorporation of DCA contact predictions in SBM simulations gives a more quantitative evaluation of the prediction quality in a subset of the gold standard. The ensemble of folded conformations based on KBM simulation incorporating coevolutionary contact predictions yields RMSD values comparable to a recent publication of state-of-the-art RNA structure predictions [150].

7.3.1. Gold Standard

I present the compiled "gold standard" of RNA families (as stored in the RFAM database) that have corresponding structures available in the PDB. The alignments for all structure sequences with their best matches in the MSAs is shown in Tab. 7.2. The best matching alignments are found by a standard Smith-Waterman algorithm [88] that allows gaps in the alignment at the cost of a score penalty.

For the families contained in the gold standard I calculated contact map predictions. 4 examples – RF00167, RF00162, RF01734 and RF00504 – are shown in Fig. 7.5 and Fig. 7.6. The contact predictions of the other 10 families are shown in the appendix in Sec. A.6. The contact maps show the native contact map in grey, the consensus secondary structure in blue and the 70 highest ranked DI predictions in red. The secondary structure is present in the top ranked predictions. The secondary structure information could also be produced by mutual information. Novel information is the enrichment in the tertiary structure contact regions of the DCA predictions compared to predictions that are gained by mutual information.

7.3.2. KBM Simulations

The simulations are based on a potential that was generated from the learning set structure SAM-I riboswitch (PDB ID 2GIS) and a collection of representative nucleotide-nucleotide contacts [26]. The workflow allows investigations of various tertiary contact scenarios in systems of interest and is able to demonstrate the reliability of the cataloged bonded interactions. Four scenarios are applied for the structures of an *add* adenine riboswitch (PDB ID 1Y26), a SAM-I riboswitch (PDB ID 2GIS), a fluoride riboswitch (PDB ID 3VRS) and a glycine riboswitch (PDB ID 3OWI):

a) only secondary structure

b) secondary structure + 100 predictions (true positive + false positive)

c) secondary structure + 50 true positive predictions

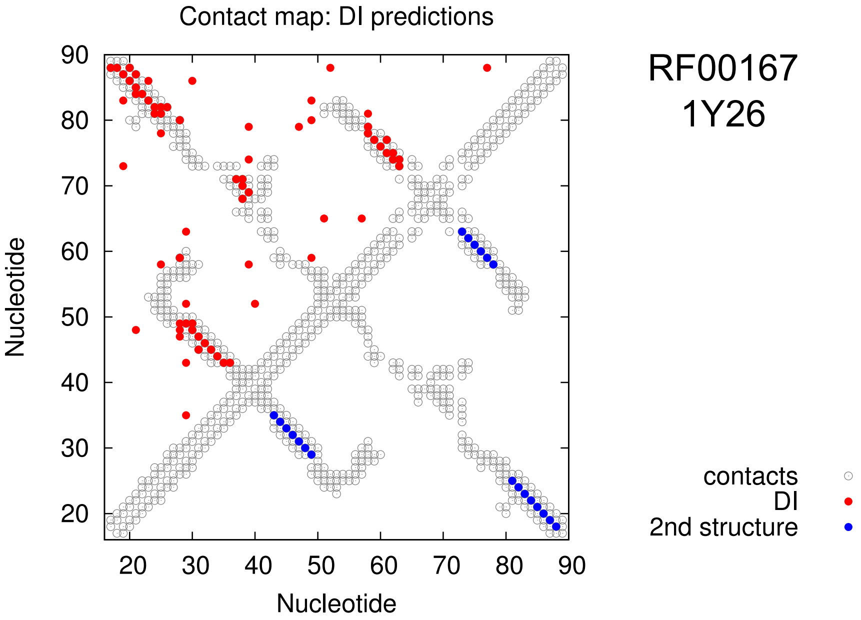
d) secondary structure + all true positive predictions

Eleonora De Leonardis provided me with predictions by the most recent version of the DCA prediction implementation. Among the 100 predictions of scenario d there are 27 true positives in RF00167, 30 true positives in RF00162, 16 true positives in RF01734 and 21 true positives in RF00504. The average and minimal RMSD values for 10 simulations in each scenario are shown in Fig. 7.7. Scenario b) and c) are the same for the fluoride riboswitch, since the riboswitch features only 41 native nucleotide-nucleotide contacts. In all cases the following general trends are evident:

- Any of the included tertiary structure information reduces the RMSD values compared to simulations with only secondary structure information.
- The addition of false positive predictions increases the RMSD values.
- Including all true positive predictions instead of just 50 has a comparably small, but slightly decreasing influence on the RMSD values.

7.4. Discussion

The presented approach allows the direct observation of a stabilizing influence of predicted tertiary contacts on an RNA strand towards its native fold. The predictions are purely based on openly accessible sequence information. The gold standard outlines the evaluation of 10 more than the 4 simulated systems of interest. I focus on a subset of the gold standard that



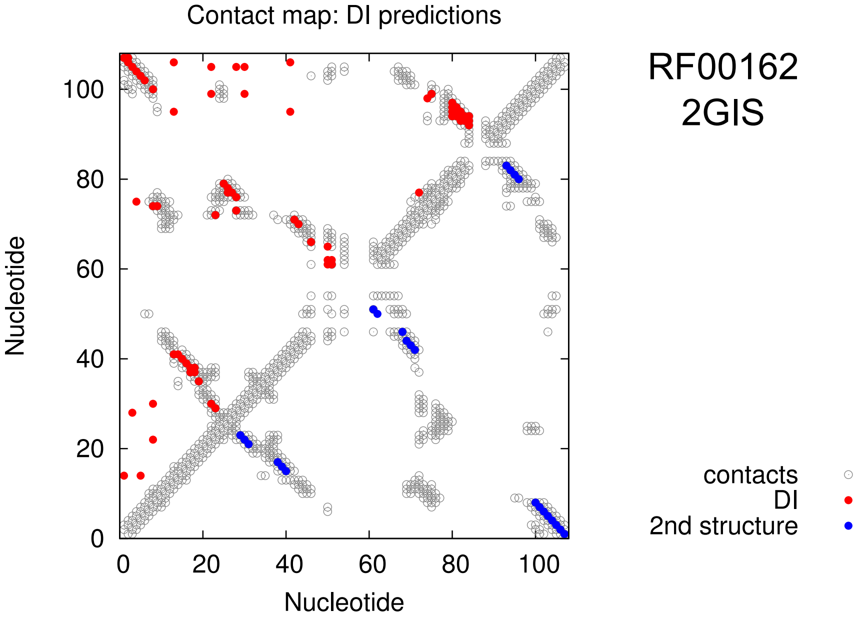


Figure 7.5.: Predicted contact maps for RNA families RF00167 (top) and RF00162 (bottom). The corresponding structures are PDB ID 1Y26 (top) and PDB ID 2GIS (bottom).

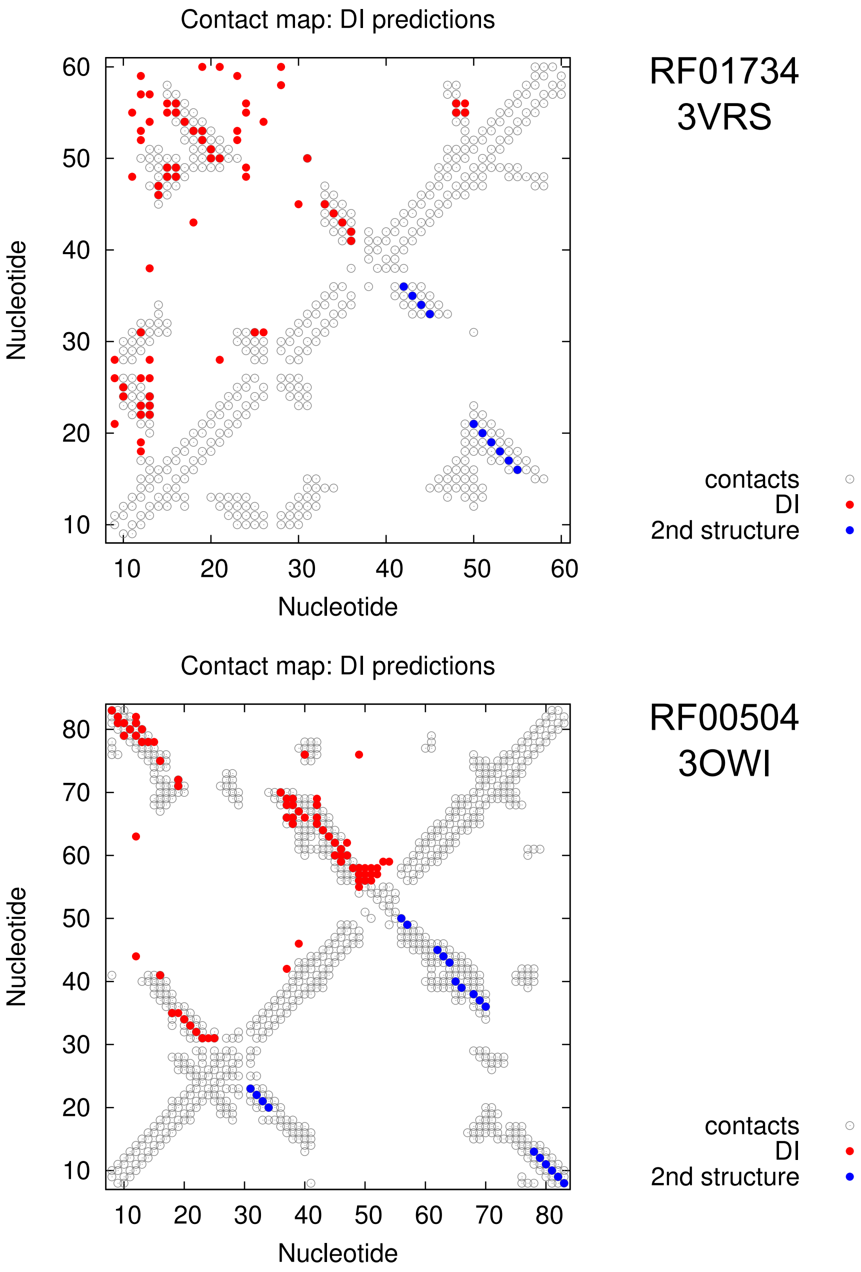


Figure 7.6.: Predicted contact maps for RNA families RF01734 (top) and RF00504 (bottom). The corresponding structures are PDB ID 3VRS (top) and PDB ID 3OWI (bottom).

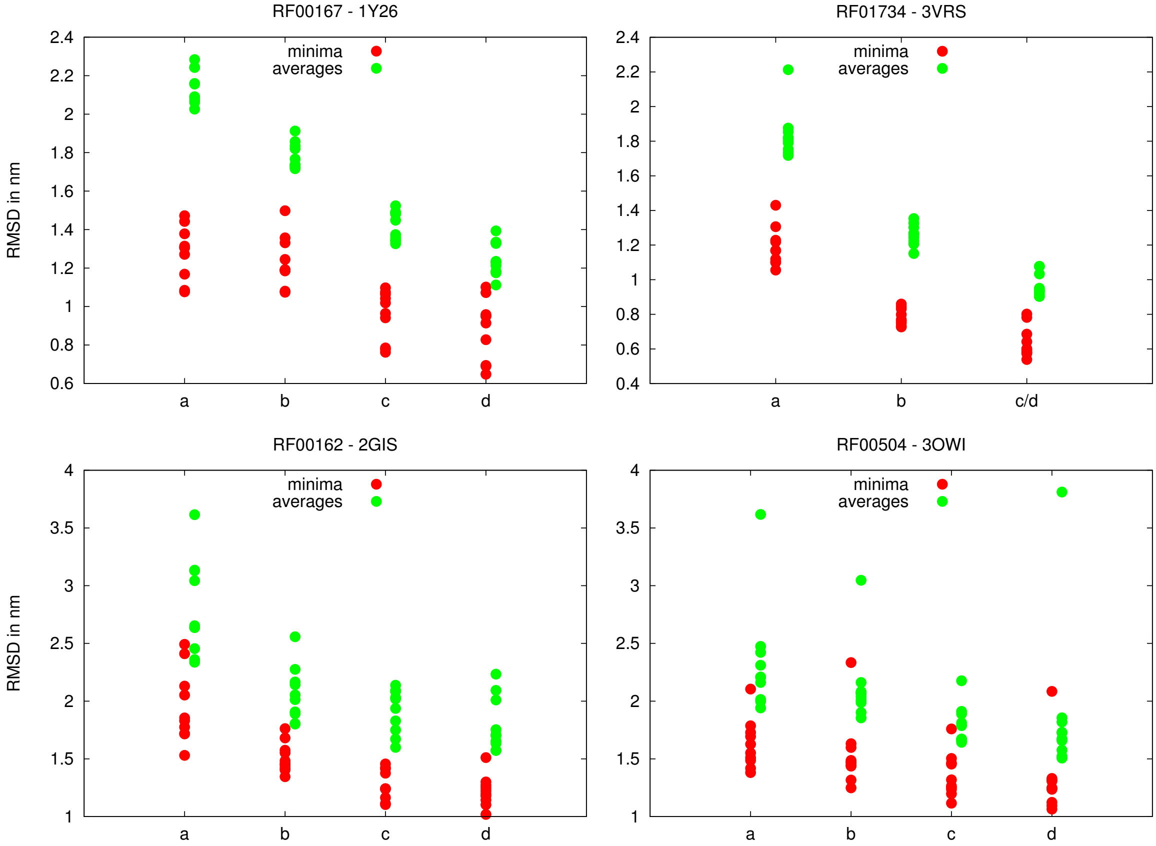


Figure 7.7.: Quantitative analysis of prediction quality. The averages of the last 2000 frames and the the minima are plotted. In each diagramm four scenarios are presented: a) only secondary structure, b) secondary structure and 100 predictions, c) secondary structure and 50 true positives, d) secondary structure and all true positives. Since the native structure of RF01734 only contains 41 contacts, scenario c and d are the same. The overall tendency is that the inclusion of tertiary contact information provided by DCA predictions in addition to consensus secondary structure information improves the prediction quality compared to simulations based only on secondary structure information. The reduction of false positive predictions seems to have a stronger influence than to increase the absolute number of positive predictions.

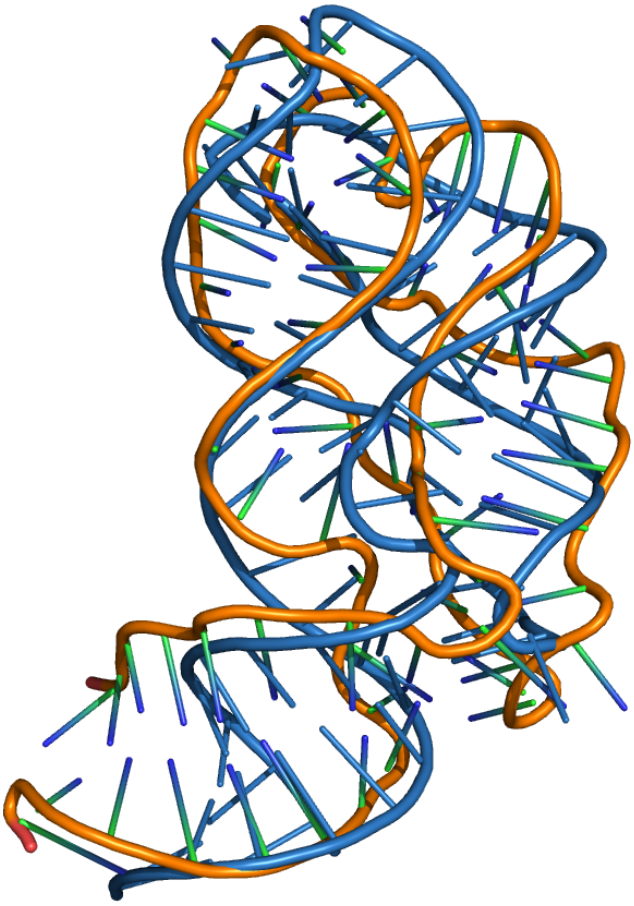


Figure 7.8.: Structure overlay of the native (PDB ID 1Y26) and a simulated conformation of the *add* adenine riboswitch. The simulated conformation has an RMSD value of 0.65 nm and illustrates the general features of simulation outcome in an exemplary, hand-picked simulation frame: The global arrangement of the structure is correct and base pairs are formed correctly. The shapes of helical substructures is quite wobbly and undefined compared to, e. g., template-based prediction techniques. This simulation was performed under the ideal conditions of all true positive predictions without false positive predictions (scenario b) and therefore indicates the limits the present knowledge-based model parametrization.

consists of the learning set itself (SAM-I riboswitch), the other riboswitch that I analyzed in the cotranscriptional folding study (*add* adenine riboswitch), the smallest riboswitch in the gold standard (fluoride riboswitch) and a reference structure (glycine riboswitch) that has been discussed in an earlier competitive study [150]. Although, our collaborative project was originally designed to evaluate the quality of DCA predictions for structural RNA formation, the calculated average RMSD values reach the range of RMSD values in this competitive study that predicted the structure of exemplary RNA structures, such as the glycine riboswitch (PDB ID 3OWI). Besides the riboswitch, only structures that were artificially designed, such as single hairpins and a constructed fourway function, are presented in the competition with much lower RMSD values. Therefore naturally occurring structures are considered the more challenging systems.

These surprisingly good results regarding prediction quality motivate further improvements to the presented novel approach of combining coevolutionary, statistical methods with KBM simulations. Several improvements both on the side of DCA and on the side of KBM simulations are indicated by the experience gained so far. On the side of DCA the major challenge is the reduction of the false positive ratio among the considered predictions. The simulations show that it is possible to reduce the RMSD by including only true positive predictions and also by increasing the absolute number of true positive predictions. An exemplary simulation frame, based on a contact map with all true positive predictions (scenario d), as an overlay with the native fold is shown in Fig. 7.8. Therefore, any technique that is able to filter or reorder the given DCA predictions for true positive predictions has potential to reduce the RMSD values to the given base line.

At the interface between DCA and KBM it may be possible determine the atom-atom contacts more precisely if there can be discovered a statistical inference of Westhof classifications [26] on the base-base contact predictions. A given Westhof classification for a DCA prediction would allow more definite contact definitions similar to the current realization of canonical base pairs in the KBM.

The KBM parametrization could be improved by a larger learning set, i. e., a combination of several different RNA structures. The larger data base might reduce the statistical errors in the dihedral angle distributions. Further inclusions of dihedral angles that can be justified by a reduced threshold for standard deviations that might stabilize the helical regions. Compared to template-based prediction techniques the helices in my *de-novo* force field are less prominent and distinct which could be improved by additional dihedral angles. Furthermore, the variation of cut-offs for the atom-atom distance averaging procedure based on Westhof classifications has not been investigated and could have an influence on the outcome of the simulations.

On the technical side there are possible future investigations: Annealing simulations with low RMSD frames as starting conformation at lower temperatures can lower the RMSD by “freezing” the helical base pairs and keeping the helical twists more rigid. The computationally more demanding approach would be to use a low RMSD conformation from the KBM simulations as starting conformation for a standard molecular dynamics simulation with stabilizing ions and explicit water. The local relaxations in this simulation would happen on time scales in the nanosecond to microsecond regime and could drive the system towards the native state.

In order to utilize this approach in the course of blind RNA structure prediction, strategies have to be developed to identify trajectory frames of minimal RMSD. This identification could be achieved by evaluating independently accessible quantities, such as (regional) Q values, energies or intra-trajectory RMSD values, directly or correlations among them. This would again reduce the given RMSD value (to the minimal instead of the average) and would provide even better starting conformations for any annealing techniques.

8 Chapter 8.

Conclusion

The field of riboswitch research is comparably young since the first discovery of a riboswitch in 2002 [9]. The investigations are focusing on the identification of new representatives of riboswitches and the characterization of regulatory mechanisms by recording the dynamics of folding and ligand binding under various experimental conditions. The inherent connection between structure and function motivates the research for tertiary structure prediction methods that give access to experimentally still unresolved riboswitches and their regulatory mechanisms. If the structure is resolved and an impression of its function is available, more detailed studies that probe the dynamics of such a riboswitch are required to improve the understanding of the interplay of involved processes.

8.1. Summary

I have presented the results of three projects in the course of my scientific work in the context of native structure-based models (SBMs) for regulatory RNA. They comprise a new, published and openly accessible software implementation of native structure-based model generation and evaluation, a published study that employs a multiscale model to investigate cotranscriptional riboswitch folding and advances to a novel approach in the field of RNA tertiary structure prediction.

eSBMTools is a Python-based software implementation of SBM generation and modification [19] that enables flexible definition of workflows of corresponding simulations. The software package assists the setup process and the evaluation of SBM simulations for protein and RNA systems. Thereby, it renders the generation process flexible enough to implement the knowledge-based model formulation utilized in the third project. In addition, eSBMTools enables several current projects in the research group, ranging from an automated workflow to analyze protein folding pathways [109] to incorporating FRET dyes in SBM simulations. The project is under ongoing development and released online under GNU General Public License version 3.0 and has recently been integrated in a UNICORE grid portlet on the MoSGrid portal [120].

The computational analysis of cotranscriptional riboswitch folding [24] represents a timely and extensive study of riboswitch folding dynamics. The general challenges of this analysis are the treatment of transcription, the timescale of RNA folding, competing conformations (the switching states), respective ligand binding and the influence of stabilizing ions. Transcription is introduced by a simplistic, coarse-grained model and the seconds timescale is reached by employing a native structure-based model that allows the investigation of folding pathways. The position restraints of the RNA polymerase during transcription are realized by an enclosing tube out of which the nascent RNA strand is extruded by acting forces. Cotranscriptional folding is evaluated via projecting the regional folding progress on the global folding progress. Thereby, the folding order of substructural elements can be identified. My findings show robust transcription-rate limited folding of substructural helical elements in the range of physiologically relevant transcription rates. The results are in agreement with the outcome of a complementary computational technique by Michael Faber at the Max Planck Institute of Colloids and Interfaces in Golm who presents a secondary structure-based model of cotranscriptional RNA folding in a kinetic Monte Carlo simulation scheme. The remaining challenges, i. e., competing conformations, explicit ligand binding and the treatment of ion concentrations are subject to future studies, as discussed in the next section.

The third project in collaboration with Eleonora De Leonardis in the group of Prof. Martin Weigt at the Université Pierre et Marie Curie Paris explores the possibilities that are offered by coevolutionary contact predictions from a direct coupling analysis [25] (DCA) in the field of RNA structure prediction. The basic idea is to exploit available and easily accessible sequence information of RNA families in order to get access to missing and comparably hard-to-reach tertiary structure information. Therefore, one needs to analyze the available aligned sequence data that is organized in tables with one line for each recorded representative and find directly coupled columns in the alignments. Direct coupling between sequence positions in the analyzed motif indicates spatial closeness due to the possibility of commutating residues that try to maintain structural integrity. The empiric single-site and pair frequency counts in a given sequence database can be described by the least constrained statistical model which is described in the maximum entropy theory. The statistical physics technique of Lagrangian multipliers provides an *ansatz* for solving this statistical model. Independent-site and mean-field approximations reduce the computational effort by avoiding the computationally demanding execution of involved partition functions. From the thereby gained stochastic model for single-site and pair frequency counts a direct coupling score can be calculated by which the predicted contacts can be ranked. The ranked predictions contain, besides secondary structure information that has already been identified by means of other statistical analyses, a sufficiently high signal of tertiary structure contacts. The relevance of the predicted contacts in combination with the respective amount of false positive predictions needs to be evaluated in order to be able to assess the novel technique. Therefore, I incorporate the predicted contact map into a knowledge-based variation of SBM simulations. The knowledge-based concept determines the geometrical values of quantities of equilibrium from a representative native conformation. Thresholds for the standard deviation of geometric values in the reference system allow the inclusion of bonded and non-bonded interactions in helical substructures. The nucleotide-nucleotide contacts predicted by DCA can be mapped onto representative atom-atom contacts by averaging over a published list of typical contact classes [26]. The knowledge-based force field can be included in the molecular dynamics simulation scheme in the same way as the native structure-based force field. Simulations of four riboswitch systems yield stable folds that show a systematic improvement of structure quality, measured by means of root mean-square deviation (RMSD) values, by the inclusion of predicted tertiary contacts in addition to secondary structure contacts.

8.2. Outlook

The presented approach to tertiary RNA structure prediction is promising since it serves as a proof of principle for the systematic improvement of RNA structure prediction quality by the inclusion of tertiary contact information from a predicted contact map. The predictions are purely based on accessible sequence data and can be utilized in combination with a simplistic energetic model to generate all-atom conformations. The evaluation of this novel approach indicates the following refinements in the course of developing a fully predictive model.

On the side of DCA, an increase of the signal-to-noise ratio or the "true positive rate" of DCA contact prediction holds out the prospect of systematic improvement to the prediction quality according to my simulations. Basic parameters such as the pseudo-count or the similarity threshold have been transferred from former protein contact prediction analyses and might need to be reevaluated in the new context of RNA contact prediction. Another possibility to reduce the number of false positive predictions is to evaluate the final predicted contact map by means of clustering techniques in order to identify accumulations of contacts that tend to be characteristic for appearances of true positive predictions. It also needs to be investigated how inferring contact classes, such as the classes proposed by the group of Eric Westhof [26], could enhance the outcome of the direct information score by the specification of the expected contact class. Such inference could determine the kind of contact according to the respective classification scheme and therefore define the mapping of nucleotide-nucleotide contacts to atom-atom contacts more precisely.

The knowledge-based force field offers a basic setup of parameters that can be optimized regarding the structure prediction quality measured by RMSD values compared to the native fold. In particular, it is necessary to find optimal cut-offs for mean value and standard deviation as part of the distance averaging while mapping predicted nucleotide-nucleotide contact onto atom-atom contacts. Characteristic values for bonded interactions and base pairing interactions can be improved by determining respective values based on multiple native RNA structures. Also for this procedure it remains to be evaluated how to choose optimal thresholds for the required maximal standard deviation that determines geometric values that are included. In a last step the exploration of appropriate scoring values to identify the lowest RMSD conformations within a folding simulation is an important task. Blind predictions rely on a robust identification of conformations close to the desired native one. In this context clustering algorithms for projections of suitable energetic score values need to be evaluated in the pursuit of low RMSD conformations.

The computational analysis of cotranscriptional riboswitch folding consists of, as discussed before, five major challenges: The treatment of transcription in atomistic simulations has to be realized, a timescale in the seconds regime needs to be reached, the two competing conformations of the switch should be included in the model, ligand binding should have an explicit effect on the folding and the involved ions should be introduced by a corresponding electrostatic description. The treatment of transcription and the desired timescale are successfully implemented in the presented study. The representation of competing conformations in the native structure-based model can be realized by introducing double minimum potential terms for the proper dihedral angles. The challenge in this case is to get estimations for the geometric values in the ligand-free conformation since their three-dimensional conformations are usually, while the mechanism is understood on the secondary structure level, experimentally not resolved. Assuming the knowledge of a native ligand-free conformation and the solution to technical difficulties of implementing the double minimum potential, the findings can be compared to results obtained by native secondary structure-based kinetic Monte Carlo simulations [24]. Introducing an explicit model of ligand binding in SBM simulation would require the parametrization of the ligand itself in combination with additional contacts that allow ligand binding. In principle this can be achieved by following the methodology presented in [22]. The treatment of electrostatics in native structure-based models can be realized by Debye-Hückel theory [146, 147].

Additions to the existing model in these regards would mean further refinement of the reduced model as a physically motivated theory of biopolymer folding. A more precise description of riboswitch dynamics aims ultimately at answering the question whether a riboswitch is kinetically or thermodynamically driven which is a crucial step in order to identify the respective regulation mechanism. Control over genetic expression by the influence of ligands could promote or suppress bacterial activity and therefore cure diseases.

Appendix A.

A Appendix

This appendix covers additional detailed elaborations and complements.

First, it gives a detailed overview over the structure of a SBM topology file that follows the GROMACS file syntax. The differences between an all-atom formulation and a C_α formulation are pointed out.

In the second section, the folding times of an add adenine riboswitch at different temperatures in SBM simulations are presented. The folding event histograms over time show a speed up in folding for the lower temperature (kinetic regime) which motivates simulations at the lower temperature in order to reduce the computational effort.

The next section investigates the influence of the base pair contact threshold on the folding characteristics and comes to the conclusion that a choice of 0.5 agrees with a wide range of choices.

Normalized regional Q value distributions within bins of the total number of formed helical base pairs are discussed in the fourth section. The histograms exhibit single maximum distributions which justifies the rough approximations by normal distributions characterized by mean value and standard deviation.

Subsequently, the mechanical drag on hairpin stem loops in SBM simulations is discussed in comparison to the behavior in kinetic Monte Carlo simulations. The observed difference in the model causes the quantitative differences in the predicted transitions of folding orders, as seen in Sec. 6.3.

Additional contact map predictions from the gold standard of structured RNA families as presented in Sec. 7.3.1 are shown in the sixth section of this chapter. The contact maps exhibit high signals for secondary structure elements while in contrast to established contact prediction techniques some enrichment of tertiary contact predictions are present.

The last section states the hardware and software used for realizing the scientific work presented in this thesis.

A.1. SBM Topology File Composition for GROMACS

This section describes the various sections in a GROMACS topology file for the use with SBM force field definitions. It is divided in an all-atom formulation, as it is used for all RNA simulations in this thesis, and a C$_\alpha$ formulation that is included in the presented implementation eSBMTools, as discussed in Sec. 5.2. The general structure is described in the GROMACS manual [49] and the default parameters are based on [18, 22, 83].

All-atom

- defaults
  This section defines the non-bonded function type as 1 (Lennard-Jones), the Lennard-Jones combination rule as 1 (parametrization choice as presented in the "pairs" section below) and prohibits the automatic generation of pair parameters.

non-bonded function	combination rule	generate pairs
1	1	no

- atomtypes
  This section represents a list of all atom types appearing in the structure.

name	mass	charge	ptype	$c6'$	$c12'$
	1.0	0.0	A	0.0	

  $c12'$ stands for the (only) repulsive term of non-bonded, non-contact interactions and is given by

$$c12' = \epsilon_2 \cdot \sigma_r^{12} = 0.01 \cdot 0.25^{12} \,.$$

- moleculetype
  This sections gives an identifier to the contained molecule and sets the number of the excluded bond distance for non-bonded interactions to 3.

name	number exclusions
	3

- atoms
  This section represents a list of all atoms appearing in the structure. Each atom has a number, a type, is associated with a residue (number and name), has a name, a charge group number and is characterized by its charge and mass.

number	type	res nr	res name	atom name	charge group nr	charge	mass
						0.0	1.0

- pairs
  This section introduces all non-bonded interactions.

i	j	function	$c6$	$c12$
		1		

  The potential terms in the Lennard-Jones formulation are calculated by

$$c6 = K_c \cdot 2 \cdot \sigma_{ij}^{6} \,, \qquad c12 = K_c \cdot \sigma_{ij}^{12} \,,$$

where σ_{ij} are the native distances. The force constant K_c is determined by

$$K_\mathrm{c} = \frac{N_\mathrm{A}}{N_\mathrm{C}} \cdot \frac{R_\mathrm{C/D}}{1 + R_\mathrm{C/D}} \, ,$$

where N_A is the total number of atoms, N_C is the total number of contacts and $R_\mathrm{C/D}$ is the ratio between total contact and total dihedral energy. The default value of $R_\mathrm{C/D}$ is 2. The total number of contacts N_C takes into account a relative weight of contacts, e. g., a factor of 1/3 for stacking contacts in nucleic acids.

- bonds

  This section contains all bonds in the structure.

i	j	function	r_0	K_b
		1		

  The default value of K_b is 20000.

- exclusions

  This sections defines all non-native contacts in a structure by a negative list that contains as a logical consequence the indices of all non-bonded interactions (pairs).

i	j

- angles

  This section contains all angles in the structure.

i	j	k	function	θ_0	K_a
			1		

  The default value of K_a is 40.

- dihedrals

  This section contains all the improper and proper dihedral angles. The improper (planar) dihedral angles are characterized by:

i	j	k	l	function	χ_0	K_i
				2		

  The default value of K_i is 40.

  The proper dihedral angles are characterized by:

i	j	k	l	function	ϕ'_0	K'_d	multiplicity
				1			1 or 3

  In case of multiplicity 1 the force constant K'_d is given by:

$$K'_\mathrm{d} = \frac{N_\mathrm{A}}{(1 + R_\mathrm{C/D}) \cdot N_\mathrm{D}} \, .$$

  In case of multiplicity 3 the force constant K'_d is given by:

$$K'_\mathrm{d} = 0.5 \cdot \frac{N_\mathrm{A}}{(1 + R_\mathrm{C/D}) \cdot N_\mathrm{D}} \, .$$

ϕ_0' needs to cause a sign change in front of the cosine since GROMACS defines its potential with a $+$ sign. In addition, ϕ_0' has to absorb the factor 3 in case of multiplicity 3 due to the GROMACS definition.

- system
  This section simply gives the described system an arbitrary name.

- molecules
  This section compiles all present molecule types (by name) present in the system (usually only one).

name	number

C$_\alpha$

The C$_\alpha$ formulation is slightly different and I state the sections that exhibit differences in the following.

- atomtypes
  This section represents a list of all atom types appearing in the structure.

name	mass	charge	ptype	$c6'$	$c12'$
	1.0	0.0	A	0.0	

$c12'$ stands for the (only) repulsive term of non-bonded, non-contact interactions and is given by

$$c12' = \epsilon_2 \cdot \sigma_r^{12} = 1 \cdot 0.4^{12} \,.$$

- pairs
  This section introduces all non-bonded interactions.

i	j	function	$c6$	$c12$
		1		

The potential terms in the modified Lennard-Jones 10-12 formulation (see Eq. (3.16)) are calculated by

$$c6 = K_\mathrm{c} \cdot 6 \cdot \sigma_{ij}^{10} \,, \qquad c12 = K_\mathrm{c} \cdot 5 \cdot \sigma_{ij}^{12} \,,$$

where σ_{ij} are the native distances and the force constant K_c has the homogeneous default value 1.

- dihedrals
  This section contains all the proper dihedral angles.

i	j	k	l	function	ϕ_0	K_d'	multiplicity
				1			1 or 3

In case of multiplicity 1 the force constant K_d' is given by

$$K_d' = 1 \,.$$

In case of multiplicity 3 the force constant K'_d is given by

$$K'_d = 0.5\,.$$

ϕ'_0 needs to cause a sign change in front of the cosine since GROMACS defines its potential with a + sign. In addition, ϕ'_0 has to absorb the factor 3 in case of multiplicity 3 due to GROMACS definition.

A.2. Folding Time of Adenine Riboswitch

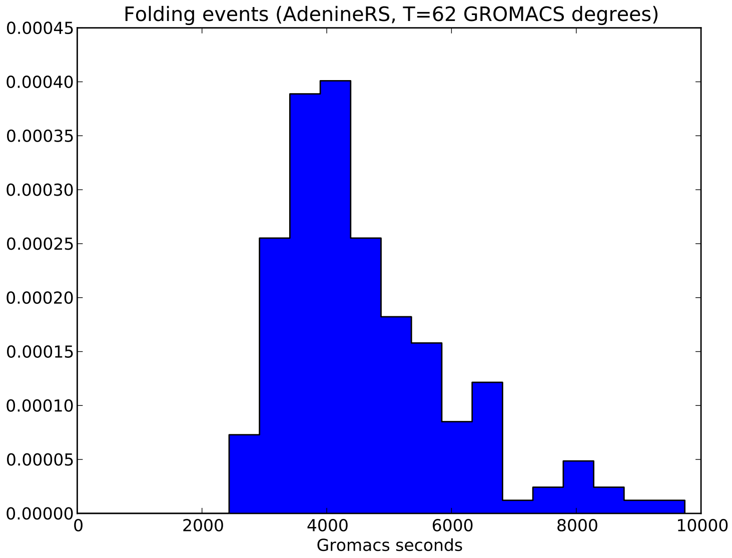
Lowering the temperature in SBM simulations speeds up folding times without changing the overall characteristics of the folding process. In order to quantify this effect folding simulations are performed for the *add* adenine riboswitch (PDB ID 1Y26 [21]). The simulations are performed at two different temperatures: 90 and 62 in reduced GROMACS units. The results are two normalized folding time distribution histograms, as shown in Fig. A.1. The folding time is characterized by the first moment in time, when the trajectory undercuts an RMSD value of 0.3 nm compared to the native conformation. The comparison yields a speed-up factor of about 6.

A.3. Choice of Base Pair Contact Threshold

For the evaluation of SBM simulations in the context of RNA folding it is appropriate to choose the number of formed base pairs as the reaction coordinate. The data contained in all-atom trajectories needs to be projected on this reaction coordinate. Therefore, it is necessary to define a measure for a formed base pair based on all-atom information. A sensible choice is the introduction of a relative threshold for the number of formed atom-atom contacts between two bases. The determination of 0.5 for this threshold is motivated by a robust behavior of the substructural folding characteristics over a wide range (0.4 - 0.8) of thresholds, as presented in Fig. A.2.

A.4. Regional Q Value Distribution in Total Q Bin

The characterization of hairpin stem loop folding is based on a distribution of normalized regional Q values over the number of formed helical base pairs. In order to be able to asign a representative number of normalized regional Q value to the corresponding number of formed helical base pairs the shape of this distributions has to be taken into account. Fig. A.3 (free folding) and Fig. A.4 (cotranscriptional folding) show the distributions for the SAM-I riboswitch (PDB ID 2GIS [20]). The observed distributions are single-maximum distributions that are crudely approximated by a Gaussian distribution to depict the magnitude of deviations in the presented folding characteristics throughout this thesis.



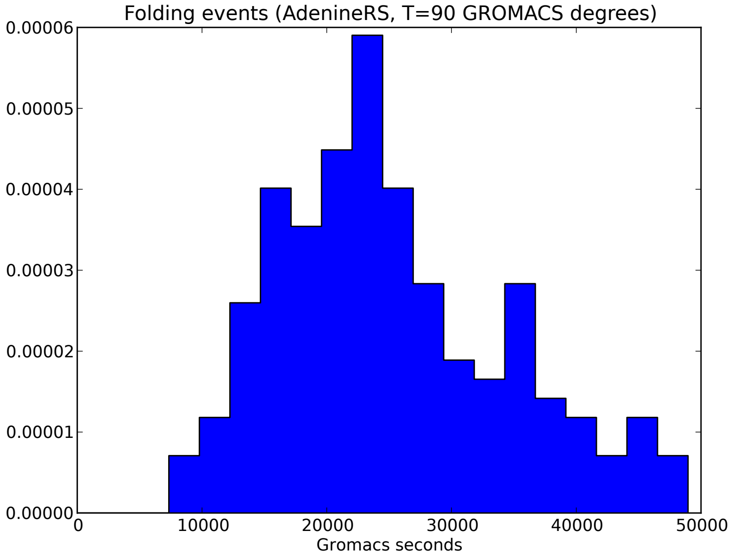


Figure A.1.: Folding times of *add* adenine riboswitch at two different temperatures. Top: The histogram of folding times at a temperature of 62 in reduced GROMACS units has its peak at about 4000 in reduced GROMACS time units. Bottom: The histogram of folding times at a temperature of 90 in reduced GROMACS units has its peak at about 24000 in reduced GROMACS time units. There is a factor of 6 in folding time due to the higher temperature. Without a change in the folding characteristics it is possible to reduce the computational effort for the folding simulations by one order of magnitude.

The histogram at the top is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

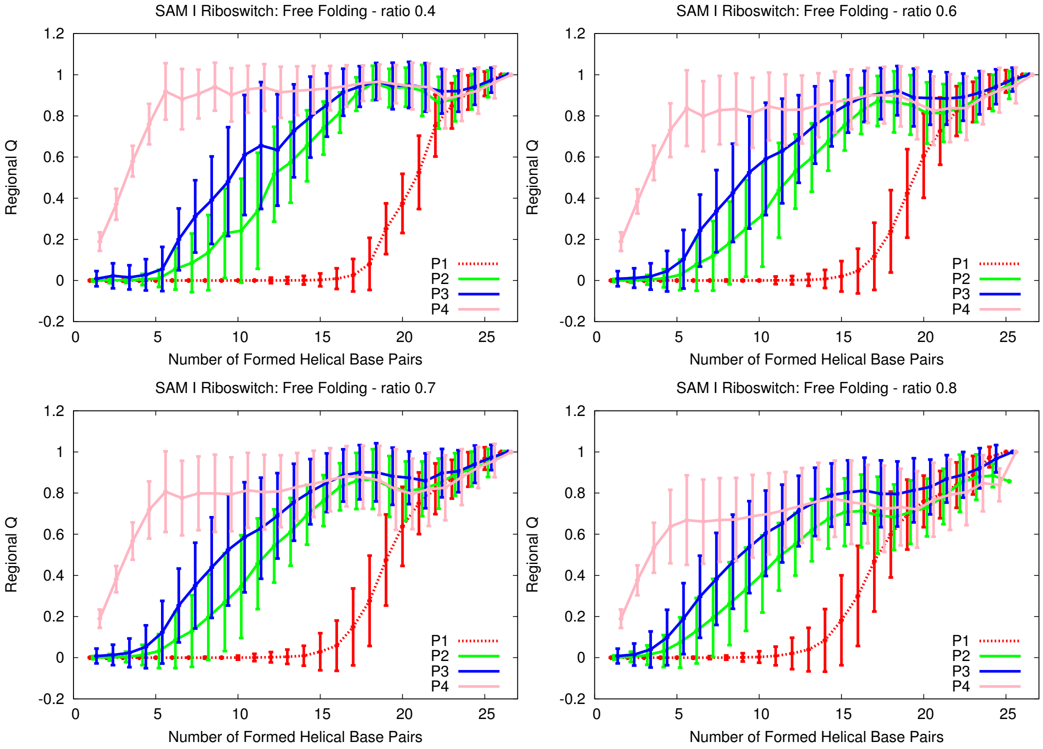


Figure A.2.: Influence of base pair contact threshold on folding characteristics of the SAM-I riboswitch. The ratio of the number of necessary formed contacts to consider a base pair as formed can be varied. Over the range of investigated base pair contact thresholds (0.4 - 0.8) there is no change in the folding characteristics. The conducted evaluations in Chap. 6 are performed at a contact threshold of 0.5.
The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

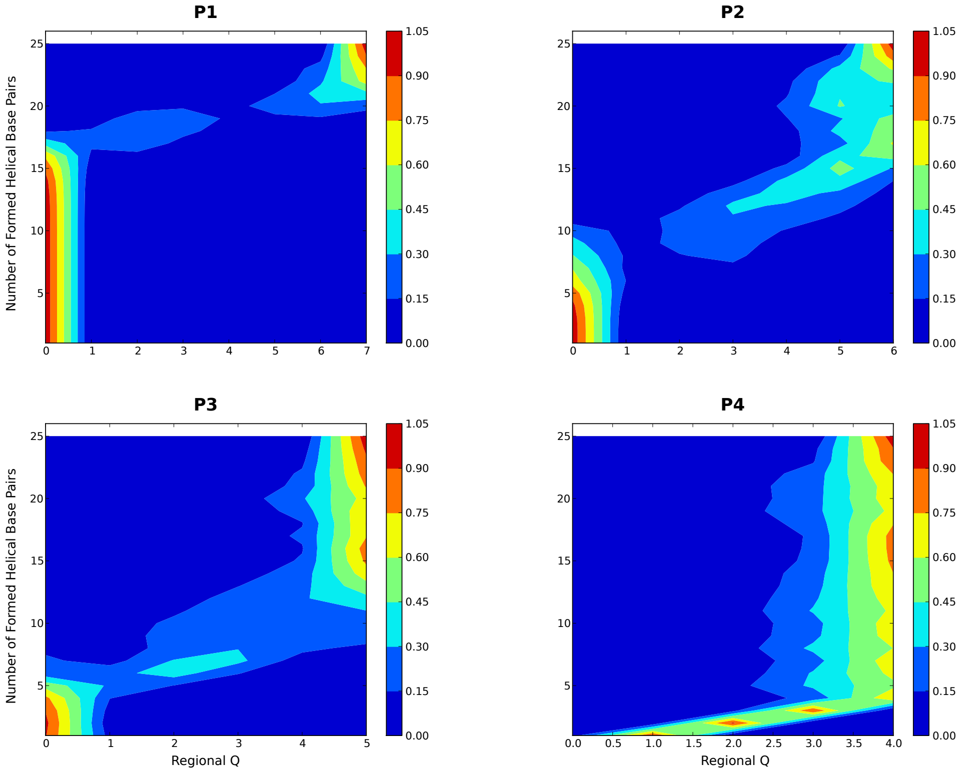


Figure A.3.: Regional Q value distribution for free folding of a SAM-I riboswitch. Depicted are the folding characteristics of helices P1 - P4. The data is extracted from the 180 simulated trajectories of free folding.

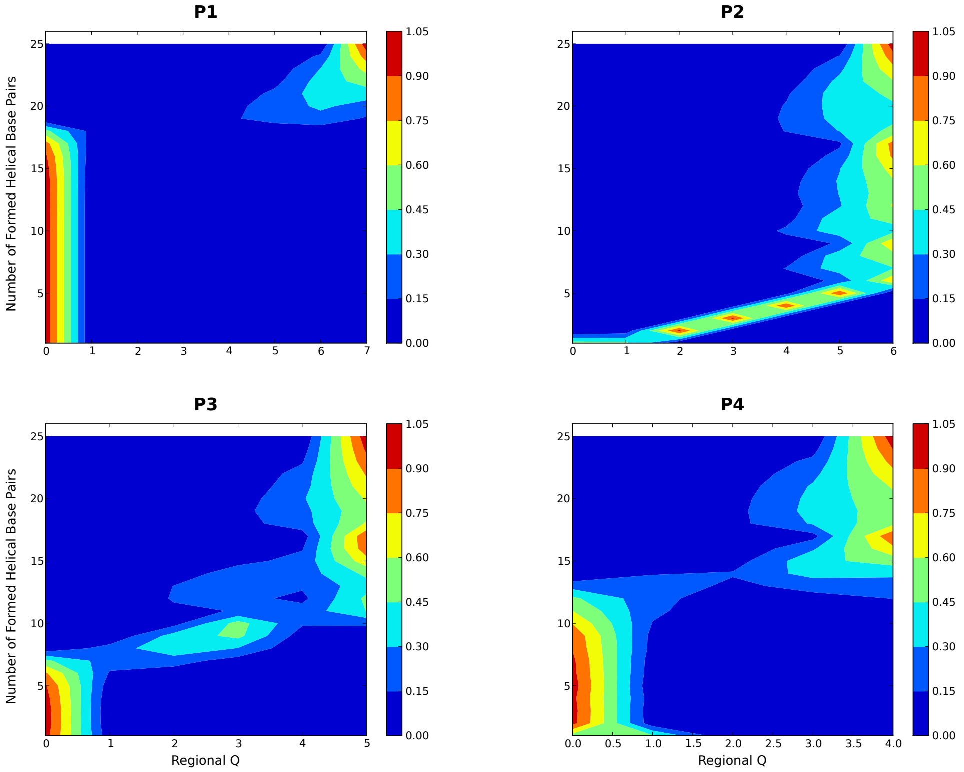


Figure A.4.: Regional Q value distribution for cotranscriptional folding of a SAM-I riboswitch. Depicted are the folding characteristics of helices P1 - P4. The data is extracted from the 80 simulated trajectories of cotranscriptional folding at a rate of 0.0025 in reduced GROMACS units (simulation parameter).

A.5. Mechanical Drag in SBM Simulations

The comparison between SBM and kinetic Monte Carlo (MC) simulations reveals a quantitative difference in the transcription rates at which the transitions of folding orders occur. The discovery of this discrepancy motivates an additional consideration of differences in the modeling of single hairpin formation. The folding time of an isolated single hairpin in SBM simulations differs from the same hairpin embedded in the whole structure, whereas in the kinetic MC simulation this is not the case. A histogram of the folding times in SBM simulations of hairpin stem loop P2 in the *add* adenine riboswitch is shown in Fig. A.5. A comparison with experimental values, as discussed in Sec. 6.2.5, yields an estimate of the folding time of 0.0175 seconds in SBM simulations. The same hairpin stem loop inside the surrounding chain of the riboswitch aptamer region exhibits a folding time of 0.375 seconds based on a distinct evaluation of the simulations in Sec. 6.2.5. The complete chain acts with a drag on both ends of the hairpin resulting in an higher estimate of the folding time. The kinetic MC approach yields for both scenarios the same estimate of the folding time, i. e., 0.00635 seconds, as shown in Fig. A.6. This folding time corresponds well to the folding time of the isolated hairpin stem loop in SBM simulations. The observed faster single hairpin folding embedded in the complete aptamer structure explains why the kinetic MC approach predicts higher transcription rates at which cotranscriptional folding starts to leave the transcription rate limited regime.

A.6. Contact Predictions for Gold Standard

This section compiles the remaining predicted DCA contact maps in the course of my studies that are introduced and discussed in Sec. 7.3.1. Families RF00001 and RF00017 are shown in Fig. A.7, families RF00059 and RF00010 in Fig. A.8, families RF00023 and RF00050 in Fig. A.9, families RF02001 and RF00168 in Fig. A.10, and families RF01051 and RF00380 in Fig. A.11. The contact maps show the native contact map in gray, the consensus secondary structure in blue and the 70 highest ranked direct information (DI) predictions in red. The secondary structure is present in the top ranked predictions. The secondary structure information could also be produced by mutual information. Novel information is the enrichment in the tertiary structure contact regions of the DCA predictions.

A.7. Used Hard- and Software

The desktop computer I used was equipped with an Intel (R) Core (TM) i5 CPU 650, running at 3.2 GHz on 4 cores. The memory was 6 GB RAM and as an operating system I chose Ubuntu Linux 11.04, 11.10, and 12.04 (LTS), consecutively. I used this machine for

- prototyping and software development (eclipse 3.7, Python 2.7, ipython, MATLAB 2012b (MathWorks))
- test runs (GROMACS [66], NAMD [64])

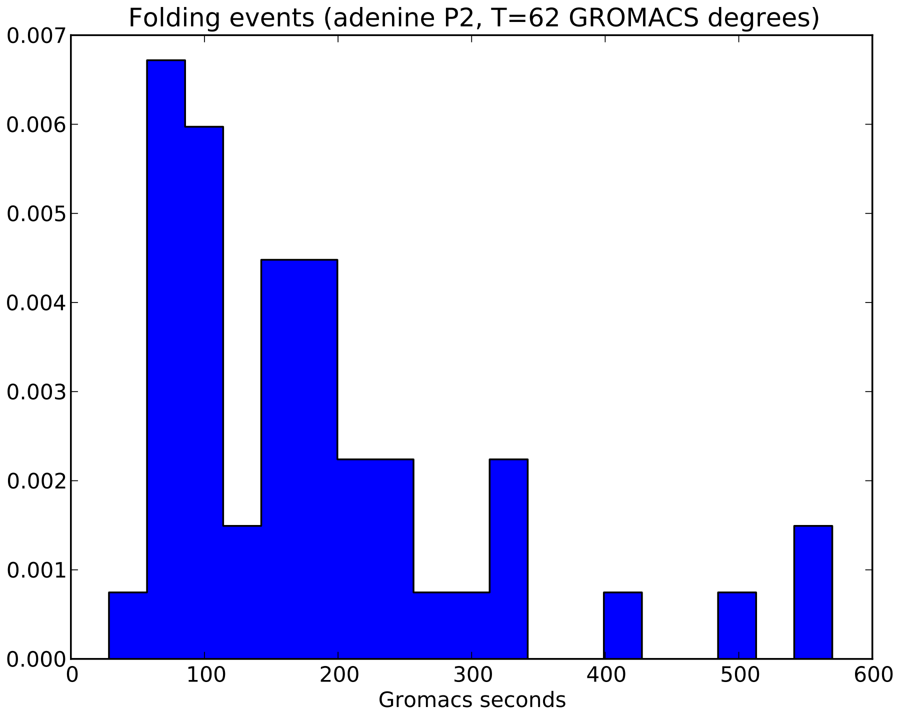


Figure A.5.: Normalized histogram of folding times in SBM simulations for free folding of isolated helix P2 (*add* adenine riboswitch). The isolated helix folds in about 90 GROMACS reduced time units instead of about 1300 for the hairpin stem loop embedded in the whole aptamer region.

The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

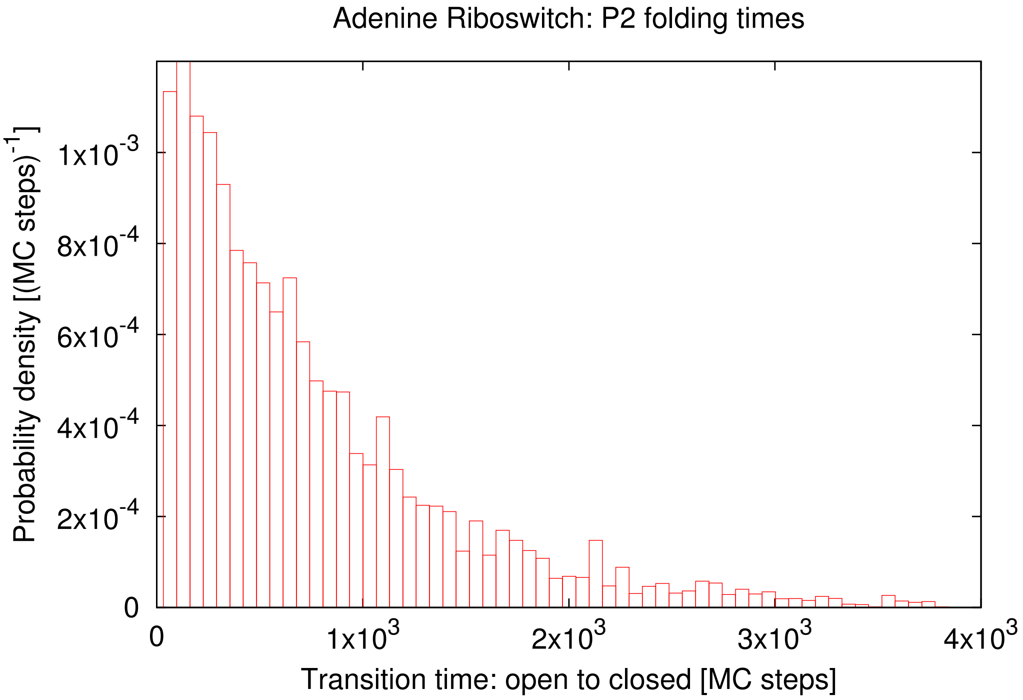


Figure A.6.: Histogram of folding times in kinetic MC simulations for free folding of helix P2 (*add* adenine riboswitch). The simulation data was generated by Michael Faber at the Max Planck Institute of Colloids and Interfaces, Golm. This characteristics is independent of the surrounding structure of the considered substructural element.
The figure is taken from the supplementary information of [24] and used under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

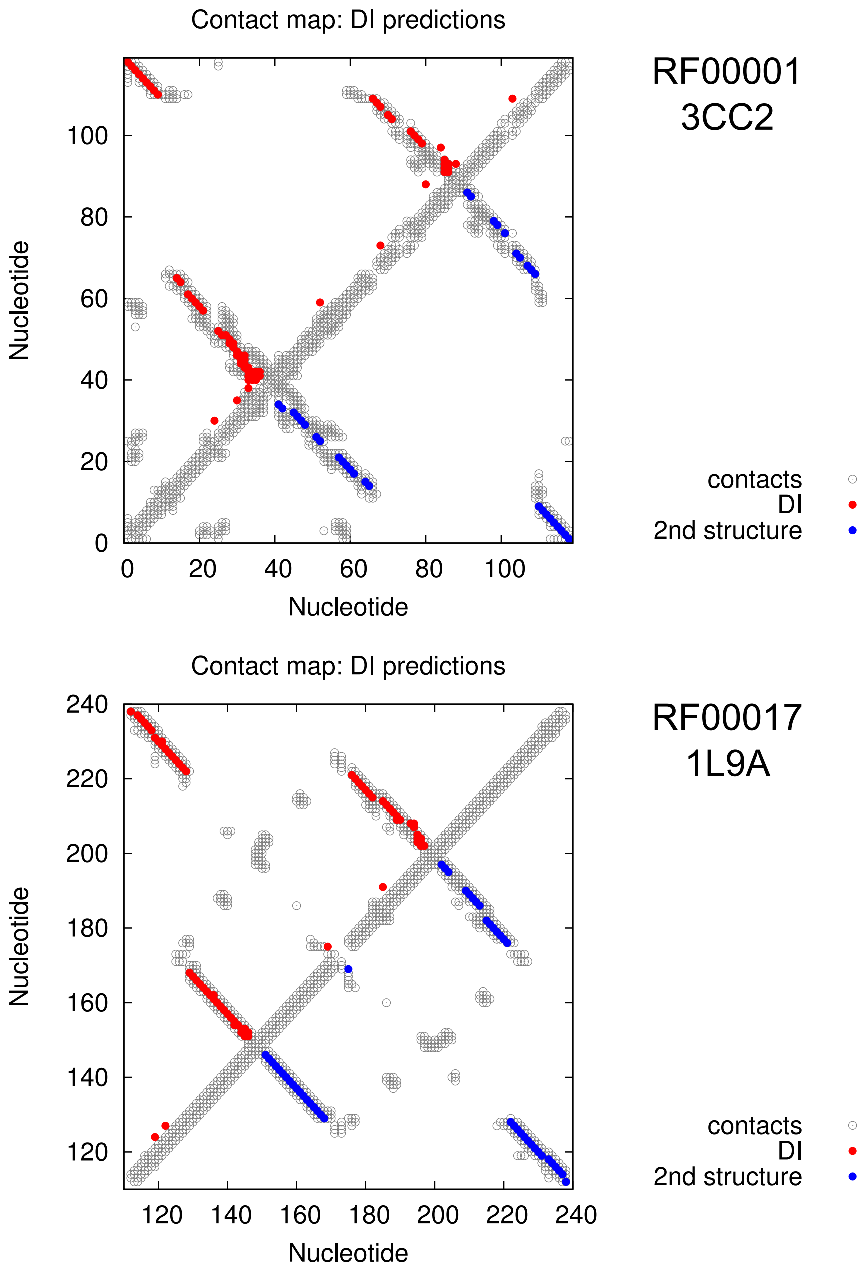


Figure A.7.: Predicted contact maps for RNA families RF00001 (top) and RF00017 (bottom). The corresponding structures are PDB ID 3CC2 (top) and PDB ID 1L9A (bottom).

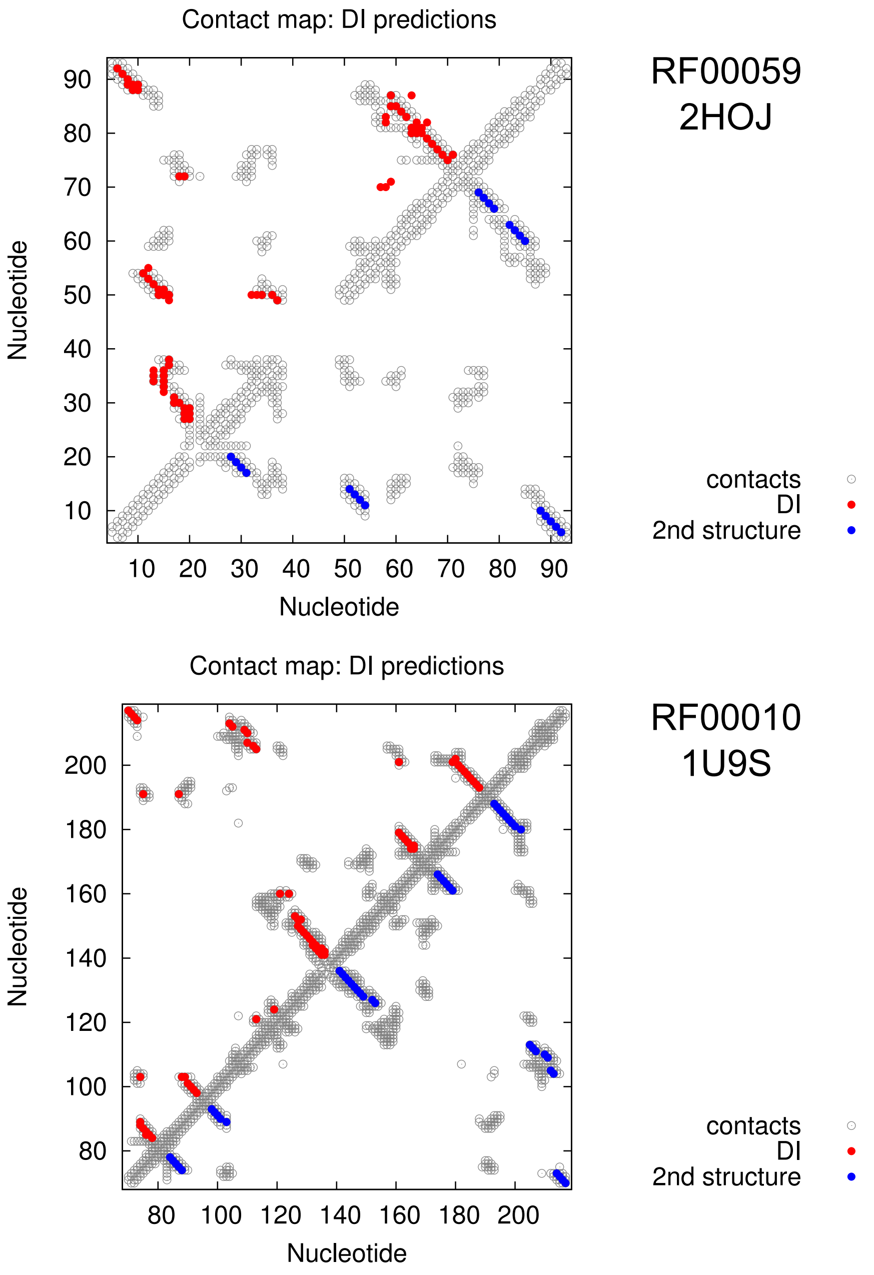


Figure A.8.: Predicted contact maps for RNA families RF00059 (top) and RF00010 (bottom). The corresponding structures are PDB ID 2HOJ (top) and PDB ID 1U9S (bottom).

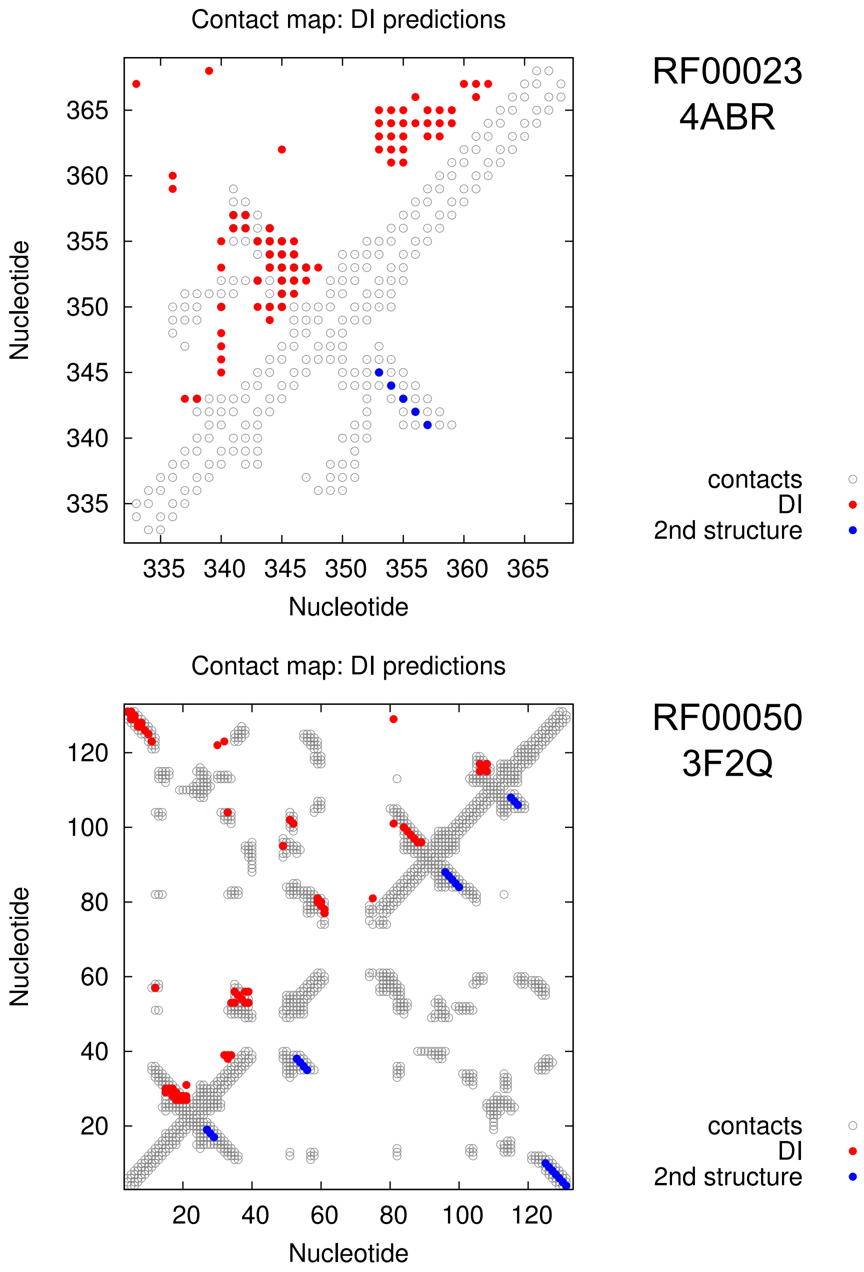


Figure A.9.: Predicted contact maps for RNA families RF00023 (top) and RF00050 (bottom). The corresponding structures are PDB ID 4ABR (top) and PDB ID 3F2Q (bottom).

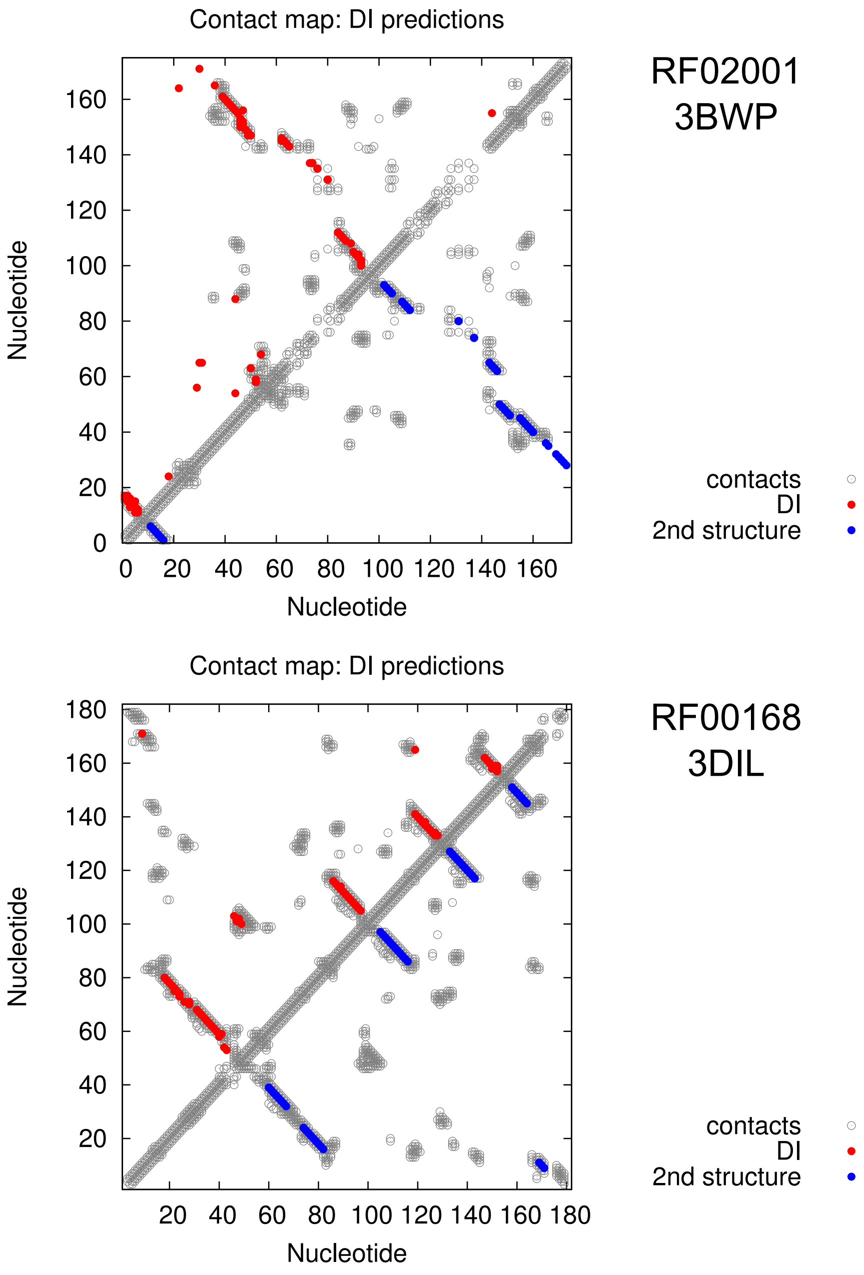


Figure A.10.: Predicted contact maps for RNA families RF02001 (top) and RF00168 (bottom). The corresponding structures are PDB ID 3BWP (top) and PDB ID 3DIL (bottom).

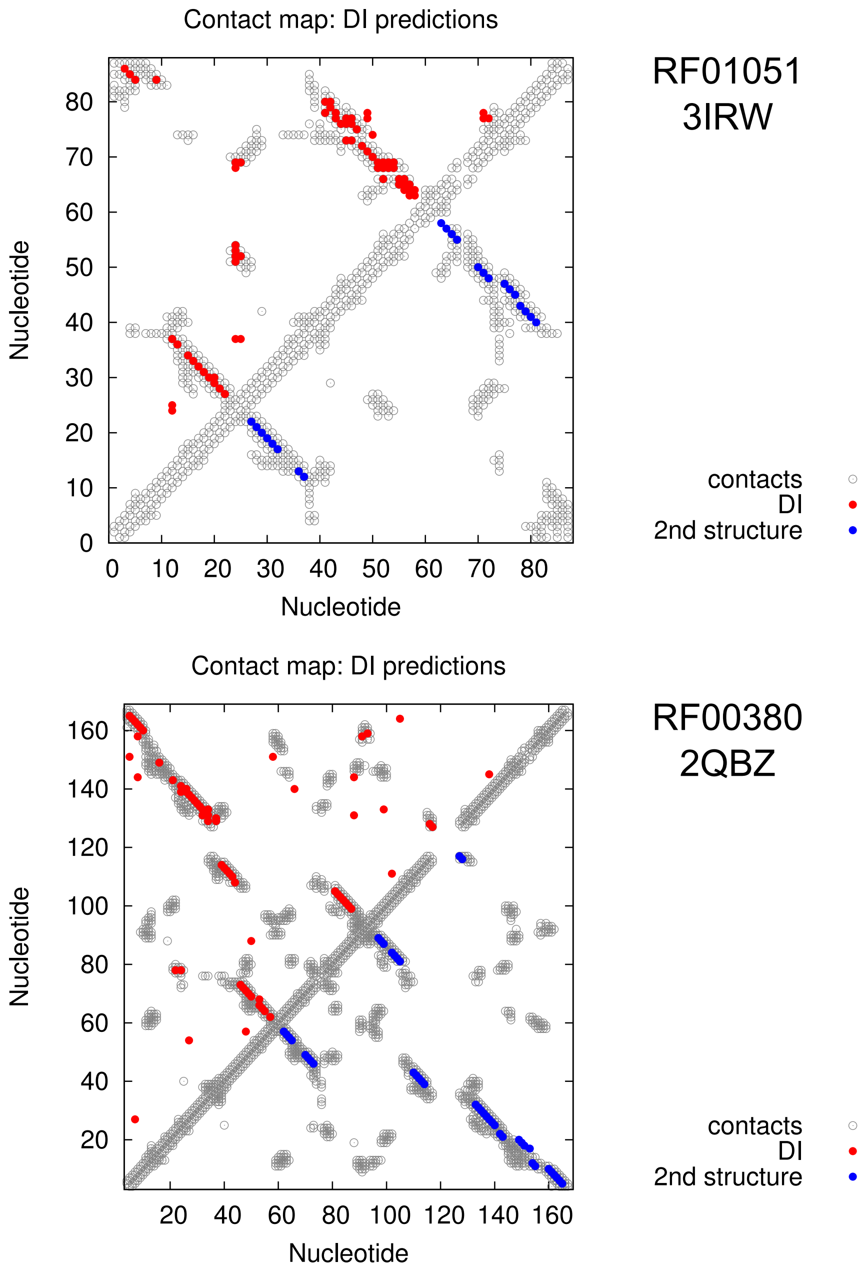


Figure A.11.: Predicted contact maps for RNA families RF01051 (top) and RF00380 (bottom). The corresponding structures are PDB ID 3IRW (top) and PDB ID 2QBZ (bottom).

- evaluation and visualization (eSBMTools [19], pymol, UCSF Chimera [151], VMD [65], xmgrace, gnuplot)
- documentation (Kile, LibreOffice, Mendeley, inkscape, gimp)

Productive runs of GROMACS simulations in the context of cotranscriptional riboswitch folding were performed on the bwGRiD cluster (`http://www.bw-grid.de`) at the sites in Karlsruhe, Esslingen, Mannheim, Heidelberg and Stuttgart. Productive runs of GROMACS simulations in the context of advances to RNA structure prediction were performed on the bwUniCluster (`http://www.bwhpc-c5.de/wiki/index.php/BwUniCluster_User_Guide`) in Karlsruhe.

Bibliography

[1] Eric H Lee, Jen Hsin, Marcos Sotomayor, Gemma Comellas, and Klaus Schulten. Discovery through the computational microscope. *Structure (London, England)*, 17(10):1295–306, October 2009.

[2] J D Robertus, Jane E Ladner, J T Finch, Daniela Rhodes, R S Brown, B F Clark, and A Klug. Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature*, 250(467):546–51, August 1974.

[3] S H Kim, F L Suddath, G J Quigley, A McPherson, J L Sussman, A H Wang, N C Seeman, and A Rich. Three-dimensional tertiary structure of yeast phenylalanine transfer RNA. *Science (New York, N.Y.)*, 185(4149):435–40, August 1974.

[4] Wendy K Johnston, Peter J Unrau, Michael S Lawrence, Margaret E Glasner, and David P Bartel. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science (New York, N.Y.)*, 292(5520):1319–25, May 2001.

[5] Jeffrey E Barrick and Ronald R Breaker. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology*, 8(11):R239, January 2007.

[6] Maumita Mandal and Ronald R Breaker. Gene regulation by riboswitches. *Nature reviews. Molecular cell biology*, 5(6):451–63, June 2004.

[7] Alexander S Mironov, Ivan Gusarov, Ruslan Rafikov, Lubov Errais Lopez, Konstantin Shatalin, Rimma A Kreneva, Daniel A Perumov, and Evgeny Nudler. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111(5):747–56, November 2002.

[8] Wade C Winkler, Ali Nahvi, and Ronald R Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910):952–6, October 2002.

[9] Ali Nahvi, Narasimhan Sudarsan, Margaret S Ebert, Xiang Zou, Kenneth L Brown, and Ronald R Breaker. Genetic control by a metabolite binding mRNA. *Chemistry & biology*, 9(9):1043, September 2002.

[10] Stewart A Adcock and J Andrew McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chemical reviews*, 106(5):1589–615, May 2006.

[11] Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–20, October 2011.

[12] Joseph D Bryngelson, José N Onuchic, Nicholas D Socci, and Peter G Wolynes. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 21(3):167–95, March 1995.

[13] José N Onuchic, Zaida A Luthey-Schulten, and Peter G Wolynes. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600, January 1997.

[14] Cecilia Clementi, Hugh Nymeyer, and José N Onuchic. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *Journal of molecular biology*, 298(5):937–53, May 2000.

[15] Aaron R Dinner, Andrej Sali, Lorna J Smith, Christopher M Dobson, and Martin Karplus. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends in biochemical sciences*, 25(7):331–9, July 2000.

[16] Steven S Plotkin and José N Onuchic. Understanding protein folding with energy landscape theory Part I: Basic concepts. *Quarterly Reviews of Biophysics*, 35(2):111–67, August 2002.

[17] Steven S Plotkin and José N Onuchic. Understanding protein folding with energy landscape theory Part II: Quantitative aspects. *Quarterly Reviews of Biophysics*, 35(3):205–86, January 2003.

[18] Jeffrey K Noel, Paul C Whitford, Karissa Y Sanbonmatsu, and José N Onuchic. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic acids research*, 38(Web Server issue):W657–61, July 2010.

[19] Benjamin Lutz, Claude Sinner, Geertje Heuermann, Abhinav Verma, and Alexander Schug. eSBMTools 1.0: enhanced native structure-based modeling tools. *Bioinformatics (Oxford, England)*, 29(21):2795–6, November 2013.

[20] Rebecca K Montange and Robert T Batey. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*, 441(7097):1172–5, June 2006.

[21] Alexander Serganov, Yu-Ren Yuan, Olga Pikovskaya, Anna Polonskaia, Lucy Malinina, Anh Tuân Phan, Claudia Hobartner, Ronald Micura, Ronald R Breaker, and Dinshaw J Patel. Structural basis for discriminative regulation of gene expression by adenine- and guanine-sensing mRNAs. *Chemistry & biology*, 11(12):1729–41, December 2004.

[22] Paul C Whitford, Alexander Schug, John Saunders, Scott P Hennelly, José N Onuchic, and Kevin Y Sanbonmatsu. Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function. *Biophysical journal*, 96(2):L7–9, January 2009.

[23] Jun Feng, Nils G Walter, and Charles L Brooks. Cooperative and directional folding of the preQ1 riboswitch aptamer domain. *Journal of the American Chemical Society*, 133(12):4196–9, March 2011.

[24] Benjamin Lutz, Michael Faber, Abhinav Verma, Stefan Klumpp, and Alexander Schug. Differences between cotranscriptional and free riboswitch folding. *Nucleic acids research*, 42(4):2687–96, February 2014.

[25] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–301, December 2011.

[26] Neocles B Leontis, Jesse Stombaugh, and Eric Westhof. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic acids research*, 30(16):3497–531, August 2002.

[27] Robert Zwanzig, Attila Szabo, and Biman Bagchi. Levinthal's paradox. *Proceedings of the National Academy of Sciences of the United States of America*, 89(January):20–22, 1992.

[28] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, 254(5038):1598–603, December 1991.

[29] Peter G Wolynes, Zaida A Luthey-Schulten, and José N Onuchic. Fast-folding experiments and the topography of protein folding energy landscapes. *Chemistry & biology*, 3(6):425–32, June 1996.

[30] Peter G Wolynes. Symmetry and the energy landscapes of biomolecules. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14249–55, December 1996.

[31] Hue Sun Chan and Ken A Dill. Protein folding in the landscape perspective: chevron plots and non-Arrhenius kinetics. *Proteins*, 30(1):2–33, January 1998.

[32] Michael Faber and Stefan Klumpp. Kinetic Monte Carlo approach to RNA folding dynamics using structure-based models. *Physical Review E*, 88(5):052701, November 2013.

[33] Marat M Yusupov, Gulnara Z Yusupova, Albion Baucom, Kate Lieberman, Thomas N Earnest, J H Cate, and Harry F Noller. Crystal structure of the ribosome at 5.5 A resolution. *Science (New York, N.Y.)*, 292(5518):883–96, May 2001.

[34] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1-2):17–24, January 2013.

[35] Evgeny Nudler. RNA polymerase active center: the molecular engine of transcription. *Annual review of biochemistry*, 78:335–61, January 2009.

[36] Innokenti Toulokhonov and Robert Landick. The flap domain is required for pause RNA hairpin inhibition of catalysis by RNA polymerase and can modulate intrinsic termination. *Molecular cell*, 12(5):1125–36, November 2003.

[37] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)*, 7(4):499–512, April 2001.

[38] Jie Chen, Seth A Darst, and D Thirumalai. Promoter melting triggered by bacterial RNA polymerase occurs in three steps. *Proceedings of the National Academy of Sciences of the United States of America*, 107(28):12523–8, July 2010.

[39] Kirsten L Frieda and Steven M Block. Direct Observation of Cotranscriptional Folding in an Adenine Riboswitch. *Science*, 338(6105):397–400, October 2012.

[40] Jenny L Baker, Narasimhan Sudarsan, Zasha Weinberg, Adam Roth, Randy B Stockbridge, and Ronald R Breaker. Widespread genetic switches and toxicity resistance proteins for fluoride. *Science (New York, N.Y.)*, 335(6065):233–5, January 2012.

[41] Adam Roth, Wade C Winkler, Elizabeth E Regulski, Bobby W K Lee, Jinsoo Lim, Inbal Jona, Jeffrey E Barrick, Ankita Ritwik, Jane N Kim, Rüdiger Welz, Dirk Iwata-Reuyl, and Ronald R Breaker. A riboswitch selective for the queuosine precursor preQ1 contains an unusually small aptamer domain. *Nature structural & molecular biology*, 14(4):308–17, April 2007.

[42] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, T N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The Protein Data Bank. *Nucleic acids research*, 28(1):235–42, January 2000.

[43] Charles E Dann, Catherine A Wakeman, Cecelia L Sieling, Stephanie C Baker, Irnov Irnov, and Wade C Winkler. Structure and mechanism of a metal-sensing regulatory RNA. *Cell*, 130(5):878–92, September 2007.

[44] Lili Huang, Alexander Serganov, and Dinshaw J DJ Patel. Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Molecular cell*, 40(5):774–86, December 2010.

[45] Alexander Serganov, Lili Huang, and Dinshaw J Patel. Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature*, 455(7217):1263–7, October 2008.

[46] Alexander Serganov, Lili Huang, and Dinshaw J Patel. Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *Nature*, 458(7235):233–7, March 2009.

[47] Kathryn D Smith, Sarah V Lipchock, Tyler D Ames, Jimin Wang, Ronald R Breaker, and Scott A Strobel. Structural basis of ligand binding by a c-di-GMP riboswitch. *Nature structural & molecular biology*, 16(12):1218–23, December 2009.

[48] Aiming Ren, Kanagalaghatta R Rajashankar, and Dinshaw J Patel. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature*, 486(7401):85–9, June 2012.

[49] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C Berendsen. Gromacs user manual version 4.5.6. Technical report, 2010.

[50] Loup Verlet. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, July 1967.

[51] R W Hockney, S P Goel, and J W Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, February 1974.

[52] Herman J C Berendsen, J P M Postma, W F van Gunsteren, A DiNola, and J R Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684, 1984.

[53] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, January 2007.

[54] William Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695–1697, March 1985.

[55] W F van Gunsteren and Herman J C Berendsen. A Leap-frog Algorithm for Stochastic Dynamics. *Molecular Simulation*, 1(3):173–185, March 1988.

[56] Matthias Rief, Mathias Gautel, Filipp Oesterhelt, Julio M Fernandez, and Hermann E Gaub. Reversible Unfolding of Individual Titin Immunoglobulin Domains by AFM. *Science*, 276(5315):1109–1112, May 1997.

[57] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[58] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, Thomas Fox, James W Caldwell, and Peter A Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, May 1995.

[59] A. D. MacKerell, D. Bashford, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux,

M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, April 1998.

[60] Chris Oostenbrink, Alessandra Villa, Alan E Mark, and Wilfred F van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of computational chemistry*, 25(13):1656–76, October 2004.

[61] Djurre H. de Jong, Gurpreet Singh, W F Drew Bennett, Clement Arnarez, Tsjerk A Wassenaar, Lars V. Schäfer, Xavier Periole, D Peter Tieleman, and Siewert J Marrink. Improved Parameters for the Martini Coarse-Grained Protein Force Field. *Journal of Chemical Theory and Computation*, 9(1):687–697, January 2013.

[62] Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, March 2013.

[63] Bernard R Brooks, Charles L Brooks, A D Mackerell, Lennart Nilsson, R J Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, A Caflisch, L Caves, Q Cui, A R Dinner, M Feig, S Fischer, J Gao, M Hodoscek, W Im, K Kuczera, T Lazaridis, J Ma, V Ovchinnikov, E Paci, R W Pastor, C B Post, J Z Pu, M Schaefer, B Tidor, R M Venable, H L Woodcock, X Wu, W Yang, D M York, and Martin Karplus. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, 30(10):1545–614, July 2009.

[64] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26(16):1781–802, December 2005.

[65] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–8, 27–8, February 1996.

[66] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, Berk Hess, and Erik Lindahl. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics (Oxford, England)*, 29(7):845–54, May 2013.

[67] Konstantinos Liolios, I-Min A Chen, Konstantinos Mavromatis, Nektarios Tavernarakis, Philip Hugenholtz, Victor M Markowitz, and Nikos C Kyrpides. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 38(Database issue):D346–54, January 2010.

[68] Sarah W Burge, Jennifer Daub, Ruth Y Eberhardt, John Tate, Lars Barquist, Eric P Nawrocki, Sean R Eddy, Paul P Gardner, and Alex Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, 41(Database issue):D226–32, January 2013.

[69] Christian B Anfinsen, E Haber, M Sela, and F H White. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 47:1309–14, September 1961.

[70] Ken A Dill and Hue Sun Chan. From Levinthal to pathways to funnels. *Nature Structural Biology*, 4(1):10–19, January 1997.

[71] Joseph D Bryngelson and Peter G Wolynes. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America*, 84(November):7524–7528, 1987.

[72] Richard A Goldstein, Zaida A Luthey-Schulten, and Peter G Wolynes. Optimal protein-folding codes from spin-glass theory. *Proceedings of the National Academy of Sciences of the United States of America*, 89(11):4918–22, June 1992.

[73] J W Moore and R G Pearson. *Kinetics and Mechanism*. A Wiley-Interscience publication. Wiley, 1961.

[74] John D. Ferry, Lester D. Grandine, and Edwin R. Fitzgerald. The Relaxation Distribution Function of Polyisobutylene in the Transition from Rubber-Like to Glass-Like Behavior. *Journal of Applied Physics*, 24(7):911, 1953.

[75] Leslie L Chavez, José N Onuchic, and Cecilia Clementi. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *Journal of the American Chemical Society*, 126(27):8426–32, July 2004.

[76] D Thirumalai and Sarah A Woodson. Maximizing RNA folding rates: a balancing act. *RNA*, 6(6):790–4, June 2000.

[77] D Thirumalai and Changbong Hyeon. RNA and protein folding: common themes and variations. *Biochemistry*, 44(13):4957–70, April 2005.

[78] Reza Behrouzi, Joon Ho Roh, Duncan Kilburn, R M Briber, and Sarah A Woodson. Cooperative tertiary interaction network guides RNA folding. *Cell*, 149(2):348–57, April 2012.

[79] Benjamin Lutz, Michael Faber, Abhinav Verma, Stefan Klumpp, and Alexander Schug. Computational Analysis of Co-Transcriptional Riboswitch Folding. *Biophysical Journal*, 106(2):284a, 2014.

[80] Claude Sinner, Benjamin Lutz, Shalini John, Ines Reinartz, Abhinav Verma, and Alexander Schug. Simulating Biomolecular Folding and Function by Native-Structure-Based/Go-Type Models. *Israel Journal of Chemistry*, (accepted), June 2014.

[81] Jeffrey K Noel, Paul C Whitford, and José N Onuchic. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *The journal of physical chemistry. B*, 116(29):8692–702, July 2012.

[82] Robert B Best, Gerhard Hummer, and William A Eaton. Native contacts determine protein folding mechanisms in atomistic simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44):17874–9, October 2013.

[83] Paul C Whitford, Jeffrey K Noel, Shachi Gosavi, Alexander Schug, Kevin Y Sanbonmatsu, and José N Onuchic. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins*, 75(2):430–41, May 2009.

[84] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, 38(Database issue):D280–2, January 2010.

[85] Erwin Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences of the United States of America*, 91(1):98–102, January 1994.

[86] David de Juan, Florencio Pazos, and Alfonso Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, March 2013.

[87] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–53, March 1970.

[88] T F Smith and M S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–7, March 1981.

[89] Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National ...*, 77(11):6309–6313, November 1980.

[90] Eva Freyhult, Vincent Moulton, and Paul Gardner. Predicting RNA structure using mutual information. *Applied bioinformatics*, 4(1):53–9, January 2005.

[91] Eckart Bindewald and Bruce A Shapiro. RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *Rna*, 12(3):342–52, March 2006.

[92] Martin Weigt, Robert A White, Hendrik Szurmant, James A Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1):67–72, January 2009.

[93] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, July 1998.

[94] David J C MacKay. *Information Theory, Inference & Learning Algorithms.* Cambridge University Press, New York, NY, USA, 2002.

[95] T Plefka. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General*, 15(6):1971–1978, June 1982.

[96] Antoine Georges and Jonathan S Yedidia. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General*, 24(9):2173–2192, May 1991.

[97] Angel E Dago, Alexander Schug, Andrea Procaccini, James A Hoch, Martin Weigt, and Hendrik Szurmant. Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26):E1733–42, June 2012.

[98] Guido Rossum. Python reference manual. Technical report, Amsterdam, The Netherlands, 1995.

[99] Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry*, 13(8):1011–1021, October 1992.

[100] Paul C Whitford, Peter Geggier, Roger B Altman, Scott C Blanchard, José N Onuchic, and Karissa Y Sanbonmatsu. Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways. *RNA (New York, N.Y.)*, 16(6):1196–204, June 2010.

[101] Cecilia Clementi and Steven S Plotkin. The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Science*, 13(7):1750–66, July 2004.

[102] Yaakov Levy, Samuel S Cho, and Tongye Shen. Symmetry and frustration in protein energy landscapes: A near degeneracy resolves the Rop dimer-folding mystery. *Proceedings of the National Academy of Sciences of the United States of America*, 102(7):2373–8, February 2005.

[103] Alexander Schug, Paul C Whitford, Yaakov Levy, and José N Onuchic. Mutations as trapdoors to two competing native conformations of the Rop-dimer. *Proceedings of the National Academy of Sciences of the United States of America*, 104(45):17674–9, November 2007.

[104] Paul C Whitford, Osamu Miyashita, Yaakov Levy, and José N Onuchic. Conformational transitions of adenylate kinase: Switching by cracking. *Journal of molecular biology*, 366(5):1661–71, March 2007.

[105] Alexander Schug, Martin Weigt, José N Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52):22124–9, December 2009.

[106] Joanna I Sulkowska and Faruck Morcos. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26):10340–5, June 2012.

[107] Torsten H Walther, Christina Gottselig, Stephan L Grage, Moritz Wolf, Attilio V Vargiu, Marco J Klein, Stefanie Vollmer, Sebastian Prock, Mareike Hartmann, Sergiy Afonin, Eva Stockwald, Hartmut Heinzmann, Olga V Nolandt, Wolfgang Wenzel, Paolo Ruggerone, and Anne S Ulrich. Folding and self-assembly of the TatA translocation pore based on a charge zipper mechanism. *Cell*, 152(1-2):316–26, January 2013.

[108] Paul C Whitford, Aqeel Ahmed, Yanan Yu, Scott P Hennelly, Florence Tama, Christian M T Spahn, José N Onuchic, and Karissa Y Sanbonmatsu. Excited states of ribosome translocation revealed through integrative molecular modeling. *Proceedings of the National Academy of Sciences of the United States of America*, 108(47):18943–8, November 2011.

[109] Claude Sinner, Benjamin Lutz, Abhinav Verma, and Alexander Schug. Effects of Energetic Heterogeneity on Protein Folding Dynamics Across Many Non-Homologous Proteins. *Biophysical Journal*, 106(2):673a, 2014.

[110] Theodor Förster. Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der physik*, 437(1-2):55–75, 1948.

[111] Volkard Helms. *Principles of Computational Cell Biology.* Wiley, 2008.

[112] Sanzo Miyazawa and Robert L Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of molecular biology*, 256(3):623–44, March 1996.

[113] Claude Sinner, Benjamin Lutz, Abhinav Verma, and Alexander Schug. Analyzing Protein Folding by High-throughput Simulations. *Biophysical Journal*, 104(2):398a, January 2013.

[114] Peter J. A. Cock, Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, and Michiel J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

[115] Antonija Kuzmanic and Bojan Zagrovic. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophysical journal*, 98(5):861–71, March 2010.

[116] Samuel S Cho, Yaakov Levy, and Peter G Wolynes. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 103(3):586–91, January 2006.

[117] g_kuh - part of the Gaussian contact extension for GROMACS available at http://smog-server.org/sbmextension.html, 2011.

[118] In-Chul Yeh, Michael S Lee, and Mark A Olson. Calculation of protein heat capacity from replica-exchange molecular dynamics simulations with different implicit solvent models. *The Journal of Physical Chemistry B*, 112(47):15064–73, November 2008.

[119] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001.

[120] Stefan Bozic, Jens Krüger, Claude Sinner, Benjamin Lutz, Alexander Schug, and Ivan Kondov. Integration of eSBMTools into the MoSGrid portal using the gUSE technology. *IWSG 2014 Proceedings*, (accepted).

[121] Dmitry G Vassylyev. Elongation by RNA polymerase: a race through roadblocks. *Current opinion in structural biology*, 19(6):691–700, December 2009.

[122] Sergei Borukhov and Evgeny Nudler. RNA polymerase: the vehicle of transcription. *Trends in microbiology*, 16(3):126–34, March 2008.

[123] Joseph A Liberman and Joseph E Wedekind. Riboswitch structure in the ligand-free state. *Wiley interdisciplinary reviews. RNA*, 3(3):369–84, 2012.

[124] J Kenneth Wickiser, Wade C Winkler, Ronald R Breaker, and Donald M Crothers. The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Molecular cell*, 18(1):49–60, April 2005.

[125] Andrea Haller, Marie F Soulière, and Ronald Micura. The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Accounts of chemical research*, 44(12):1339–48, December 2011.

[126] Jinwei Zhang, Matthew W Lau, and Adrian R Ferré-D'Amaré. Ribozymes and riboswitches: modulation of RNA function by small molecules. *Biochemistry*, 49(43):9123–31, November 2010.

[127] J Kenneth Wickiser, Ming T Cheah, Ronald R Breaker, and Donald M Crothers. The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry*, 44(40):13404–14, October 2005.

[128] Andrei Yu Kobitski, Alexander Nierth, Mark Helm, Andres Jäschke, and G Ulrich Nienhaus. Mg2+-dependent folding of a Diels-Alderase ribozyme probed by single-molecule FRET analysis. *Nucleic acids research*, 35(6):2047–59, January 2007.

[129] William J Greenleaf, Kirsten L Frieda, Daniel a N Foster, Michael T Woodside, and Steven M Block. Direct observation of hierarchical folding in single riboswitch aptamers. *Science (New York, N.Y.)*, 319(5863):630–3, February 2008.

[130] Changbong Hyeon and D Thirumalai. Chain length determines the folding rates of RNA. *Biophysical journal*, 102(3):L11–3, February 2012.

[131] Changbong Hyeon, Ruxandra I Dima, and D Thirumalai. Size, shape, and flexibility of RNA structures. *The Journal of chemical physics*, 125(19):194905, November 2006.

[132] Giulio Quarta, Ken Sin, and Tamar Schlick. Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. *PLoS computational biology*, 8(2):e1002368, January 2012.

[133] Ryan L Hayes, Jeffrey K Noel, Udayan Mohanty, Paul C Whitford, Scott P Hennelly, José N Onuchic, and Karissa Y Sanbonmatsu. Magnesium fluctuations modulate RNA dynamics in the SAM-I riboswitch. *Journal of the American Chemical Society*, May 2012.

[134] Jean-François Lemay, J Carlos Penedo, Renaud Tremblay, David M J Lilley, and Daniel A Lafontaine. Folding of the adenine riboswitch. *Chemistry & biology*, 13(8):857–68, August 2006.

[135] Rachel Anne Mooney, Irina Artsimovitch, and Robert Landick. Information processing by RNA polymerase: recognition of regulatory signals during RNA chain elongation. *Journal of bacteriology*, 180(13):3265–75, July 1998.

[136] Robert Landick. RNA polymerase slides home: pause and termination site recognition. *Cell*, 88(6):741–4, March 1997.

[137] Sandra J Greive and Peter H von Hippel. Thinking quantitatively about transcriptional regulation. *Nature reviews. Molecular cell biology*, 6(3):221–32, March 2005.

[138] Paul C Whitford, Karissa Y Sanbonmatsu, and José N Onuchic. Biomolecular dynamics: order-disorder transitions and energy landscapes. *Reports on progress in physics. Physical Society (Great Britain)*, 75(7):076601, July 2012.

[139] Alessandra Villa, Jens Wöhnert, and Gerhard Stock. Molecular dynamics simulation study of the binding of purine bases to the aptamer domain of the guanine sensing riboswitch. *Nucleic acids research*, 37(14):4774–86, August 2009.

[140] Herve Isambert and Eric D Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12):6515–20, June 2000.

[141] Christoph Flamm, Walter Fontana, IL Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *Rna*, 6(3):325–38, March 2000.

[142] Francesco Di Palma, Francesco Colizzi, and Giovanni Bussi. Ligand-induced stabilization of the aptamer terminal helix in the add adenine riboswitch. *RNA (New York, N.Y.)*, 19(11):1517–24, November 2013.

[143] Maumita Mandal and Ronald R Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature structural & molecular biology*, 11(1):29–35, January 2004.

[144] Maumita Mandal, Mark Lee, Jeffrey E Barrick, Zasha Weinberg, Gail Mitchell Emilsson, Walter L Ruzzo, and Ronald R Breaker. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science (New York, N.Y.)*, 306(5694):275–9, October 2004.

[145] Walter Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618–618, February 1986.

[146] Dana Vuzman, Ariel Azia, and Yaakov Levy. Searching DNA via a "Monkey Bar" mechanism: the significance of disordered tails. *Journal of molecular biology*, 396(3):674–84, March 2010.

[147] Levani Zandarashvili, Dana Vuzman, Alexandre Esadze, Yuki Takayama, Debashish Sahu, Yaakov Levy, and Junji Iwahara. Asymmetrical roles of zinc fingers in dynamic DNA-scanning process by the inducible transcription factor Egr-1. *Proceedings of the National Academy of Sciences of the United States of America*, 109(26):E1724–32, June 2012.

[148] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[149] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian Kaufman, P Douglas Renfrew, Colin a Smith, Will Sheffler, Ian W Davis, Seth Cooper, Adrien Treuille, Daniel J Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J Fleishman, Jacob E Corn, David E Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J Gray, Brian Kuhlman, David Baker, and Philip Bradley. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in enzymology*, 487(11):545–74, January 2011.

[150] José Almeida Cruz, Marc-Frédérick Blanchet, Michal Boniecki, Janusz M Bujnicki, Shi-Jie Chen, Song Cao, Rhiju Das, Feng Ding, Nikolay V Dokholyan, Samuel Coulbourn Flores, Lili Huang, Christopher a Lavender, Véronique Lisi, François Major, Katarzyna Mikolajczak, Dinshaw J Patel, Anna Philips, Tomasz Puton, John Santalucia, Fredrick Sijenyi, Thomas Hermann, Kristian Rother, Magdalena Rother, Alexander Serganov, Marcin Skorupski, Tomasz Soltysinski, Parin Sripakdeevong, Irina Tuszynska, Kevin M Weeks, Christina Waldsich, Michael Wildauer, Neocles B Leontis, and Eric Westhof. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. *RNA (New York, N.Y.)*, 18(4):610–25, April 2012.

[151] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–12, October 2004.

Acknowledgements

Finally, I would like to thank the people who supported me in the course of my 'Doktorarbeit'. In particular, I thank the following people:

- Prof. Dr. Wolfgang Wenzel, for supervising my work and examining my thesis;
- Prof. Dr. Gerd Ulrich Nienhaus, for being the co-examiner of my thesis;
- Dr. Alexander Schug, for supervising my work and giving me the great opportunity to perform my scientific studies in his research group at the Steinbuch Centre for Computing;
- Dr. Stefan Klumpp and Dipl. Phys. Michael Faber at the MPI of Colloids and Interfaces in Golm for the fruitful collaboration on cotranscriptional riboswitch folding;
- Prof. Dr. Martin Weigt and M. Sc. Eleonora De Leonardis at the Université Pierre et Marie Curie Paris for letting me join their interesting work on RNA contact prediction and giving me the outstanding opportunity to spend two months in their lab;
- Dr. Abhinav Verma, for being a constant help and person of contact for all both technical and biophysical questions and a great co-worker;
- Dipl. Phys. Claude Sinner, Dipl. Phys. Angela Poschlad, Dr. Shalini John Lovis and Ines Reinartz for their productive assistance, the creative discussions, thorough proof reading and all the fun we had;

Furthermore, I give many thanks to the research group of Wolfgang Wenzel, especially Timo, Moritz, Julia, Priya, Simon and Martin. I enjoyed the cooperative, nice and open atmosphere, the diverting conversations and various common activities.

I am extraordinarily grateful for my wife's multifaceted support, for her heartly indorsement and for setting me a bright and inspiring example of ambition and joy.

Last but not least, I thank my parents for their steady support and their encouragement.