Felix Kühnl

Folding Kinetics of Riboswitches

Master's Thesis

Master's Thesis

Folding Kinetics of Riboswitches

A tractable concentration-dependent approach

Felix Kühnl
ID 2516047

March 2016

Supervisors:
Prof. Dr. Peter F. Stadler
Dr. Sebastian Will

Chair for Bioinformatics
Institute of Computer Science
Faculty of Mathematics and Computer Science
University of Leipzig

Abstract

The interaction of RNAs and their ligands strongly depends on folding kinetics and thus requires explanations that go beyond thermodynamic effects. Whereas the computational prediction of minimum energy secondary structures, and even RNA–RNA and RNA–ligand interactions, are well established, the analysis of their kinetics is still in its infancy. Due to enormous conformation spaces, the exact analysis of the combined processes of ligand binding and structure formation requires either the explicit modeling of an intractably large conformation space or—often debatable—simplifications. Moreover, concentration effects play a crucial role. This increases the complexity of modeling the interaction kinetics fundamentally over single molecule kinetics.

This work presents a novel tractable method for computing RNA–ligand interaction kinetics under the widely-applicable assumption of ligand excess, which allows the pseudo-first order approximation of the process. It rigorously outlines the approach and discusses model parametrization from empirical measurements. Furthermore, the kinetics of the designed theophylline riboswitch RS3 are studied at different ligand concentrations and with respect to co-transcriptional effects. Additionally, the concept of *canonical landscapes* is put on a solid theoretical foundation, defining a symmetrical move set yielding a connected landscape as well as direct paths in these. Furthermore, a heuristic approach for partially exploring energy landscapes around a given structure of interest is described. All results are implemented as usable software tools.

Contents

Chapter 1.

Introduction

Riboswitches are regulatory RNA elements usually located in the 5′-UTR of genes. They enable the specific response to the presence of *ligands*, i. e. small molecules that can bind to the RNA, by transcriptional or translational control of gene expression. Their ability to switch genes on or off depending on small molecules such as theophylline or tetracycline makes them valuable biotechnological tools. The design of tailored riboswitches for specific applications and advanced control logic is therefore an attractive endeavor in synthetic biology (Wachsmuth et al. 2013). A riboswitch can be understood as the composition of its aptamer and its actuator domain. It senses the ligand by binding it to a binding pocket of the aptamer domain; this influences the conformations of the actuator domain and thereby leads to a measurable response to ligand binding, e. g. by terminating transcription (*off* switch) or suppressing the terminator hairpin (*on* switch).

The computational design of artificial riboswitches requires a sufficiently accurate model of the ligand binding process and the structural response of the RNA to ligand binding. The equilibrium thermodynamics of RNA–ligand binding have been studied for RNA–RNA interactions e. g. in Bernhart et al. (2006) and Dimitrov and Zuker (2004), and for small molecule binding in RNA–ligand (Espah Borujeni et al. 2015). As in the case of single molecule RNA folding, purely thermodynamic models are sometimes insufficient because they disregard the dynamics of the process. This can cause dramatic mis-predictions. Various approaches have analyzed the kinetics of single molecule RNA folding (Flamm et al. 2000; Hofacker et al. 2010; Mann et al. 2014; Wolfinger et al. 2004). For tractability the continuous process is decomposed into elementary steps, simplified based on heuristic assumptions, and approximated by a coarse-grained process.

One especially important simplification is the restriction of a RNA's conformation space to *canonical* structures, which are going to be defined as the structures that do not contain any isolated base pairs. This greatly reduces the conformation space of secondary structures and makes

computations feasible for RNAs of sizes that could not have been handled before. Though the notion of canonical structures is well-known (Bompfünewerer et al. 2007) and basic support for this approximation is implemented in some computer programs (e. g. `barriers`, Flamm et al. 2002), these implementations are incomplete and lack a solid theoretical foundation. Therefore, a symmetric move set for canonical RNA landscapes is introduced formally. Furthermore, the well-known notion of *direct paths* (Flamm et al. 2000) is extended to canonical landscapes, allowing for a highly efficient approximation of path searches and barrier height estimations between arbitrary canonical structures.

Another important simplification is the *pruning* of RNA landscapes to structures that lie within an energy band of defined width above the minimum free energy of the given RNA. This radical heuristic is essential because of the enormous number of secondary structures which grows exponential with the RNA's length. However, as a consequence, structures of interest might be removed from the landscape, too. Therefore a heuristic has been developed that tries to remedy this dilemma by exploring only parts of an energy landscape and connecting them to the other structures, yielding a connected landscape again.

This work is structured as follows. In Chapter 2, the biochemical background of structure and function of RNA and especially riboswitches is explained. Next, in Chapter 3, mathematical formalizations of these concepts are given, which will be used throughout this work. This chapter also states important natural laws that are used to derive the folding model later on. Chapter 4 defines canonical RNA landscapes and shows important properties of these. Further, the notion of direct canonical paths is introduced and their existence is proved. The partial, heuristic exploration of energy landscapes is considered in Chapter 5. Finally, a tractable model of riboswitch folding kinetics is developed in Chapter 6. The results of this thesis are concluded in Chapter 7.

Publication of the results. Parts of this work, especially Chapter 6, have been used in the preparation of the manuscript (Kühnl et al. 2016) which was accepted for publication in the *Proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA)* (2016).

Chapter 2.

Biochemical Background

This chapter outlines the basic biochemical background of this thesis. The structure and function of RNA will be explained. Special emphasis is put on *riboswitches*, a class of regulatory RNA elements that is highly interesting for synthetic biology. That is because a riboswitch allows an external regulation of the expression of genes and functional RNAs depending on the presence of a certain ligand. Therefore, riboswitches may proof to be a useful tool to analyze the function of specific genes.

2.1. The structure of DNA and RNA

Both DNA and RNA are important biomolecules present in any known life form. They are chains of *nucleobases*, namely adenine, guanine, cytosine, thymine and uracil, which are connected by a *sugar–phosphate backbone* (Vollhardt and Schore 2003, p. 1179). While thymine is only present in DNA, uracil can only be found in RNA. The nucleobases are often abbreviated using their initial letters A, G, C, T and U, respectively. The structural difference between DNA and RNA can be inferred from their names: DNA stands for deoxyribonucleic acid and RNA for ribonucleic acid, indicating that DNA is missing a hydroxyl (OH) group that is present in the sugar of RNAs, cf. Fig. 2.1 on the following page. Because of that missing reactive group, DNA is more stable than RNA. The nucleobases in both RNA and DNA have a strong tendency to form pairs connected by hydrogen bonds. This pairing, however, is not arbitrary: the very stable *canonical* or Watson-Crick *base pairs* are A−T, A−U and G−C, and the less stable *wobble pair* is G−U. Other pairings are energetically unfavorable and seldom observed. Each nucleobase can pair with at most one other base. The canonical base pairing schema induces a notion of complementarity: for a given sequence of nucleobases, its *complementary sequence* is defined as the sequence consisting of the canonical pairing partner of each nucleobase in the original sequence.

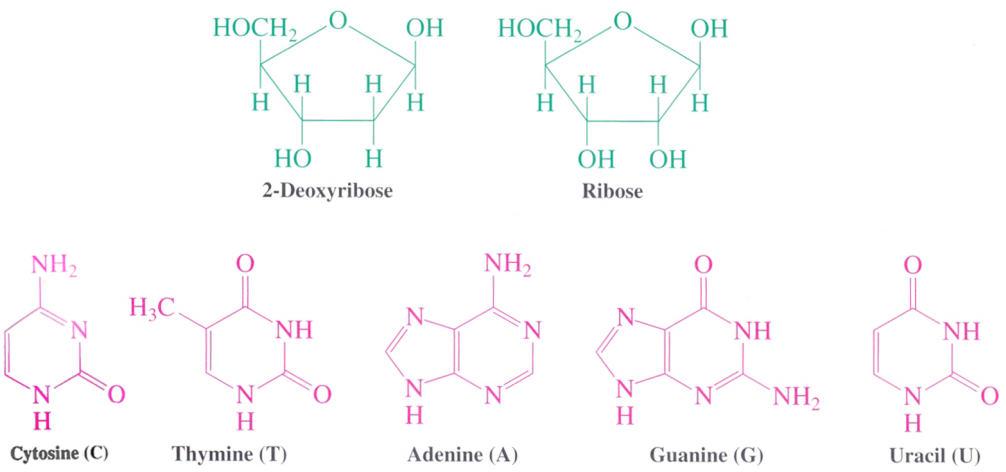


Figure 2.1.: *Top:* Structure of 2-deoxyribose and ribose as found in DNA and RNA, respectively. Note the missing hydroxyl group at the second C atom. *Bottom:* Structure of the five nucleobases. In RNAs, the nucleobase thymine is not present but replaced with uracil and *vice versa*. Source: McQuarrie and Simon (1997).

The sugar–phosphate backbone of RNA consists of ribose molecules with an attached phosphate group (cf. Fig. 2.2 on the next page). In DNA, as mentioned before, a hydroxyl group is missing, so instead of ribose, the sugar is called deoxyribose. The sugar molecule consists of a ring of four carbon atoms and an oxygen atom, additional groups attached to it, and a fifth dangling carbon atom. The carbon atoms in the ring are numbered in clock-wise order, beginning after the oxygen atom. The fifth carbon atom is attached to the fourth one. The phosphate group is attached to the third carbon atom of its associated ribose as well as the fifth carbon atom of the next ribose. Therefore, the sugar-phosphate backbone gives the sequence of nucleobases attached to it a *direction*. By convention, and because this is the direction of DNA transcription (cf. Section 2.2 on page 15), DNA and RNA sequences are written from $5'$ to $3'$, where n' refers to the n-th carbon atom. The first and the last nucleobase of a sequence represent its $5'$ and $3'$ end, respectively, and the terms *upstream* and *downstream* are used to refer to nucleobases that lie further in the direction of the $5'$ end or $3'$ end, respectively, w. r. t. to some reference nucleobase.

The most notable structural difference between DNA and RNA is that the former usually occurs as a *double-stranded* molecule, i. e. two chains of nucleobases paired with each other wind into a helical structure. The two strands pair in opposite directions such that the first base of the

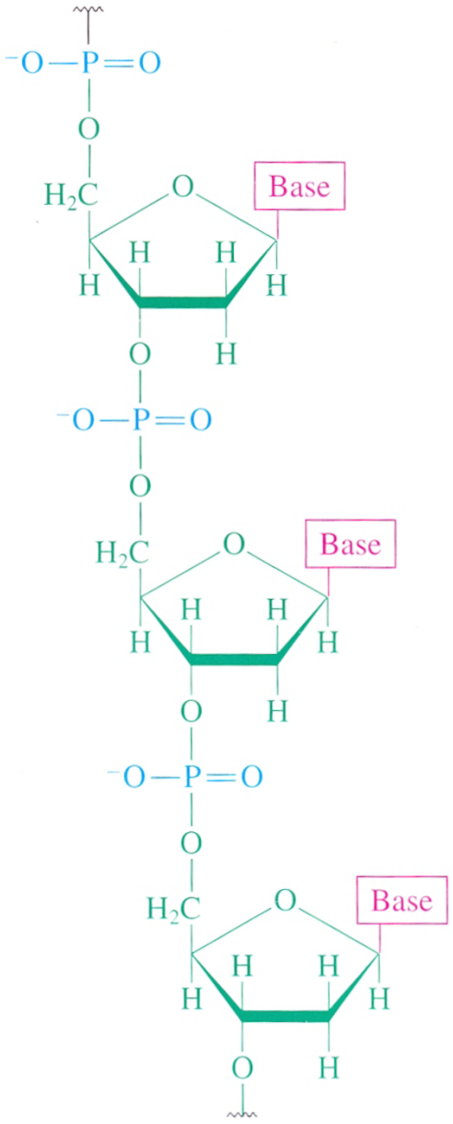


Figure 2.2.: Part of a DNA molecule. The 2-deoxyribose molecules are connected by phosphate groups. Together they form the sugar–phosphate backbone of DNA. The structure of RNAs is similar, but the sugar molecule is ribose instead. The "base" placeholder stands for one of the nucleobases from the bottom of Fig. 2.1 on page 12. The top of the chain represents the 5′-end, the bottom is the 3′-end. Source: McQuarrie and Simon (1997).

first strand pairs with the last base of the second strand, the second base with the second-last one etc. RNA, in contrast, most often has as a single-stranded form. Since the expression of stable base pairs reduces the molecules' free energy and any physical system prefers states of low energy, the nucleobases of the open RNA chain begin to pair with each other within the same molecule. This process is referred to as *folding* of the RNA, leading to a specific combination of base pairs called its *secondary structure*. Every structure the molecule can fold into according to the pairing rules specified above is called a structural *conformation* of this specific RNA. For typical sequences, the combinatorial nature of the structure gives rise to an enormous amount of possible conformations. The entirety of conformations for a given RNA sequence is referred to as this RNA's *structure ensemble*. In a sample of a specific RNA, there are usually many different conformations from the ensemble, though certain structures are very dominant while other ones are extremely rare.

Of course, the given definition of structure and conformation based on the base pairs present within the RNA molecule is a simplification. The folded RNA chain in a solution has a certain three-dimensional structure that can be vital for the functionality it provides. However, there is evidence that the secondary structure is a sensible approximation to describe the RNA structure and reason about its function (Flamm et al. 2000). Often, the secondary structure of related functional RNAs (cf. Section 2.2 on the next page) found in different species is conserved while its sequence differs.

An additional assumption is usually made when dealing with RNA secondary structures. According to the definition above, base pairs in an RNA could also cross each other, e. g. the first nucleobase may pair the fifth one and the third one with the eighth one, forming a knot-like structure called *pseudo-knot*. It is common to exclude conformations containing pseudo-knots from algorithms and methods since they often make them much more difficult, rendering many approaches infeasible for larger molecules. An example is the thermodynamic RNA folding problem (with respect to the usual energy models) that is polynomial for pseudo-knot free structures (Zuker and Stiegler 1981) but becomes NP-hard for arbitrary ones. (Akutsu 2000). Because of that, the same assumption will be made throughout this work, disregarding any structure containing pseudo-knots. Additionally, it is assumed that the minimal loop length in any structure is three, i. e. between any base pair there must lie at least $\varepsilon_{\text{loop}} = 3$ unpaired nucleobases. This is to account for the stiffness of the sugar–phosphate backbone which cannot be bent into arbitrary angles.

2.2. Functions of DNA and RNA

Despite their structural similarities, DNA and RNA perform very different functions in organisms. The purpose of DNA is to store the entirety of information of a cell that is passed on to its ancestors. This information is encoded as a nucleobase sequence on double-stranded DNA molecules. A part of this information encodes *genes*, i. e. sequence snippets that are translated into a protein that performs various functions in the cell. This translation process, more commonly referred to as *expression* of genes, consists of several steps that vary across different types of organisms, for example between procaryotes and eucaryotes. The common steps[1] across all organisms, however, are the *transcription* of the gene into a so-called *messenger RNA* or *mRNA*, and the *translation* of this mRNA into the protein. The *translation* is performed by a protein–RNA complex called *ribosome* that successively reads the mRNA chain 5′-to-3′ direction. Thereby it translates the nucleobase sequence of the mRNA into a sequence of amino acids that form the protein using a coding scheme that is almost generic to all living organisms. The *transcription* of the DNA is performed by an enzyme called *RNA polymerase*. It reads off the sequence of one DNA strand in the same direction and produces a *complementary* sequence of RNA that carries this information from the DNA to the ribosome, explaining the name "messenger RNA". In the bacterium *Escherichia coli*, the transcription proceeds with a rate of about 50 nucleobases per second (Bremer and Dennis 2008). During the translation, the ribosome reads off the mRNA at a similar rate (ibid.). However, since always three nucleotides encode a single amino acid, the "output rate" of the ribosome is about three times lower.

Beside its role as mRNA, there are numerous other classes of *non-coding* RNAs (ncRNAs) that are also transcribed from the DNA and perform a broad spectrum of tasks, many of which are related to the regulation of gene expression. This is achieved e. g. by direct or indirect degradation of specific mRNA transcripts, alteration of the *splicing* process[2] in eucaryotes, or direct interaction with the RNA polymerase or the ribosome. An example for the latter type of RNA are *riboswitches*, which are the central type of RNA that this work is about. They are introduced in greater detail in Section 2.3 on the following page. Because of their functions, non-coding RNAs are also referred to as *regulatory* RNA. Other important types of

[1] The following description is strongly simplified and limited to those aspects that are relevant to this work.

[2] The term "splicing" means the removal of parts of a gene from the mRNA transcript.

non-coding RNA are rRNAs, which serve as components of the ribosome, as well as tRNAs, which provide the ribosome with the amino acids used to build a protein during the translation of mRNA.

While the role that gene expression plays in the metabolism of cells is known for decades, it has not been until recently that the importance of non-coding RNAs (ncRNAs) was realized. In the past, the *intergenic* areas, i. e. the parts of the DNA that are not transcribed into mRNA, have been referred to as "junk DNA". Today it is known that the biggest part of the mammalian genomes are transcribed into RNA transcripts, however, to which extend these transcripts perform biological functions is still hotly debated (Palazzo and Lee 2015)

2.3. Riboswitches

A very interesting class of regulatory RNA elements are so-called *riboswitches*. As their name suggests, they act as a kind of "switch" that can turn on or off the expression of a gene or another non-coding RNA placed immediately downstream of the switch (Breaker 2011). This function is performed by direct interaction of the RNA with a another molecule, the *ligand*. The effect is that the riboswitch inhibits or enables either the *translation* through an interaction with the ribosome, or the *transcription* by interacting with the DNA polymerase. This difference gives rise to the classification of this RNA type into *transcriptional* and *translational* riboswitches. Another type of riboswitch acts as a *ribozyme*, i. e. a RNA molecule catalyzing a certain reaction, in this case a self-cleavage that degrades the mRNA transcript.

As mentioned, riboswitches act as "genetic switches". They are controlled by direct interaction with an external *ligand* molecule that specifically binds a certain structural area of the riboswitch. This area is called the ligand's *binding pocket* on the riboswitch, and the part of the riboswitch containing the binding pocket is also referred to as its *sensor domain*. As described in more detail in Section 3.6 on page 38, the formation of a ligand–RNA dimer complex is an energetically favorable reaction. As a result, in the presence of the ligand the conformations that possess the ligand's binding pocket dominate the RNA's structure ensemble. In absence of the ligand, however other structures become more likely and the binding pocket is not expressed very often.

Beside the sensor domain, riboswitches contain a second structural element: an *actuator domain* containing a *terminator* that is capable of

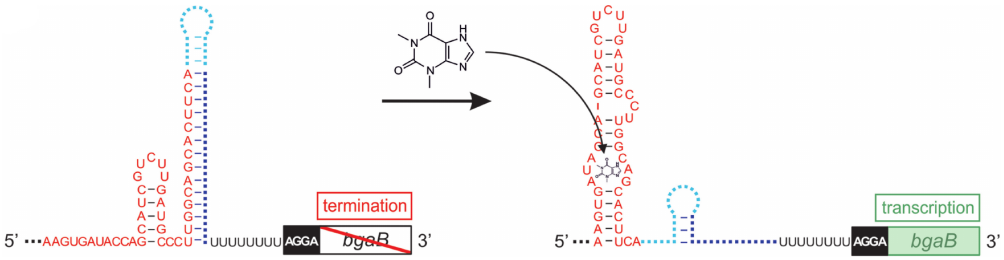


Figure 2.3.: Example of a transcription-regulating riboswitch controlled by theophylline (*top middle*). The red sequence part forms aptamer domain, and the blue parts constitute the terminator. Note that the aptamer and the terminator overlap. *Left:* In absence of the ligand, the terminator hairpin forms. *Right:* The ligand binds into the binding pocket and, thereby, suppresses the terminator formation. Source: Wachsmuth et al. (2013).

interrupting the DNA polymerase during transcription or the ribosome during translation, depending on its type. In transcriptional riboswitches, for example, this process is mediated by a mechanism called *Rho-independent* or *intrinsic termination* (Wachsmuth et al. 2013), i. e. a hairpin loop (formally defined in Definition 17 on page 30) forms on the mRNA transcript immediately after it has been transcribed by the RNA polymerase. This so-called *terminator hairpin* is followed by a *poly-U stretch*, a part of the sequence consisting only of uracil nucleobases, e. g. eight bases in Wachsmuth et al. (ibid.). Figure 2.3 gives an example of such a riboswitch. As the RNA polymerase proceeds to transcribe the poly-U stretch, its binding to the nucleic acid sequences is less strong. The terminator hairpin, which has already formed by this time, is able to interact with the polymerase such that it releases the incomplete mRNA transcript. Note that this process is highly time-critical: if the RNA polymerase has already passed over the poly-U stretch before the terminator has formed, it can no longer be interrupted and the transcription continues even if, in the long run, the terminator dominates the structure ensemble.

The actual switching function of a riboswitch is mediated by an interplay of the formation of the sensor and the actuator domain. Often, the two domains are overlapping and competing in the sense that either only the binding pocket or only the terminator structure can be present at a time. For example, in an *on* switch, the terminator forms in the absence of the ligand, because its energetic properties are better than that of the binding pocket. Binding the ligand molecule, however, grants an energy bonus to

the—by itself unfavorable—binding pocket. This way, the ligand overturns the dominance of the terminator in the structure ensemble and "switches on" the transcription by preventing the formation of the terminator. Again, it is important that the switching process happens quick enough, such that the formation of the terminator is interrupted before it can interact with the RNA polymerase.

2.4. The design of synthetic riboswitches

Though riboswitches are naturally present gene regulatory elements in procaryotes and, to some extent, in eucaryotes (Breaker 2011), a very interesting idea is the design of *synthetic* riboswitches. Given an arbitrary ligand molecule, one wants to be able to engineer a sequence that, when inserted into the 5′-UTR of a gene, acts as specific type of riboswitch. To achieve that, several problems have to be solved:

1. finding an *aptamer* for the given ligand, i. e. a RNA sequence that acts as a sensor expressing a binding pocket the ligand can bind to,

2. finding an actuator structure that can interrupt or enable the transcription or translation, depending on the type of riboswitch that is to be designed,

3. combining sensor and actuator such that they can act as the intended riboswitch type, and

4. finding a RNA sequence that folds into the required secondary structures depending on the presence of the ligand.

Step 1 cannot currently be performed *in silico*. Therefore, an experimental method called *SELEX* (Tuerk and Gold 1990) is utilized to find the desired sensor. Once an aptamer for a given ligand has been found, it can be re-used for the design of all types of riboswitches. Repeating the experiment is only necessary when changing the ligand.

SELEX stands for "systematic evolution of ligands by exponential enrichment". In short, this method starts out with many different RNAs and filters off those that are unable to bind the ligand. The remaining RNAs are amplified and mutated, and the filtering process is repeated with a higher intensity. This process is repeated several times until the strongest-binding aptamer is found. The method can be enhanced to also

ensure a high *specificity* of the ligand. That means that the ligand binds only to the ligand and not to another substance of choice.

Step 2 can be performed by analyzing natural riboswitches and has to be performed only once for each class of riboswitch. Note that this step yields a RNA *structure*, while the previous step yields a *sequence*.

Step 3, the combination of sensor and actuator domain, is a challenging task. Both domains need to overlap in such a way that binding the ligand switches the actuator on or off, depending of the type of riboswitch. Though computational tools can aid this process, it is often performed by hand relying on expert knowledge and intuition. The quality of the result.

Finally, in *Step 4* a sequence needs to be found that folds into the designed structures. This task is known as the *inverse folding problem* and is addressed by a number of methods, e. g. the tool `RNAdesign` (Siederdissen et al. 2013). Since the folding process is supposed to depend on the presence of the ligand, additional design goals are required to make sure the riboswitch works as intended (cf. Flamm et al. 2001; Wachsmuth et al. 2013, 2015).

After one or more riboswitches of the desired type have been designed, it is required to verify that they are fully functional. Of course, a final judgment can only be rendered after an *in vivo* experiment in the target organism, as there are simply much more aspects involved deciding over the functionality of a design than could be taken into account by any feasible design method. The problem, however, is that such experiments are time-consuming and cost a large amount of money, such that it is highly desirable to identify malfunctioning designs before wasting resources on their experimental evaluation. *In silico* approaches can help to achieve this goal, and one important step is the analysis of the folding kinetics of the designed riboswitches. Therefore, this work aims at developing a tractable model of their folding process that can help to decide whether the produced sequence will indeed perform as intended.

Chapter 3.

Mathematical Preliminaries

This chapter introduces the reader to the basic mathematical terms and ideas used across the chapters of this thesis. Constructions that are specific to a certain part of this work are explained in that respective section.

3.1. General energy landscapes

This section formally describes the terminology of energy landscapes used throughout the rest of this work.

Definition 1 (Moves and move sets)**.** *Let S be an arbitrary set. A* move set $\mathcal{M} = \{\mu_1, \mu_2, \ldots\}$ *w. r. t. S is a set of* partial *functions*

$$\mu_k : S \dashrightarrow S, \quad k = 1, 2, \ldots$$

that map none of the elements of S to itself, i. e. $\forall s \in S : \mu_k(s) \neq s$ for all k. These functions are referred to as moves*.*

A move μ is valid *w. r. t. to an element $s \in S$ if $\mu(s)$ is defined.*

Moves are called that way since, given an element $s \in S$, the application of the function μ can be interpreted as a move from s to $\mu(s)$. Move sets can be symmetric or asymmetric:

Definition 2 (Symmetry of move sets)**.** *A move set $\mathcal{M}$ on a set S is called* symmetric *if, for any move $\mu \in \mathcal{M}$, there is an* inverse *move $\mu^{-1} \in \mathcal{M}$ such that*

$$\forall \mu(s) \in \mu(S) : \quad \mu^{-1}\Big(\mu(s)\Big) = s,$$

where $\mu(S) = \{\mu(s) \,|\, s \in S \text{ and } \mu(s) \text{ is defined}\}$ is the image of μ on S. Otherwise, it is called asymmetric*.*

Move sets induce a notion of neighborhood among the elements of S:

Definition 3 (Neighborhood and adjacency)**.** *Let $\mathcal{M}$ be a move set w. r. t. a set S. The* neighborhood *of an element $s \in S$ w. r. t. $\mathcal{M}$ is the set of all elements that can be reached from s by applying a single move from $\mathcal{M}$ to it. It is denoted as*

$$N_{\mathcal{M}}(s) = \{\mu(s) \,|\, \mu \in \mathcal{M}\}\,.$$

If the used move set can be inferred from the context, it is left off the index.

Two elements $s_1, s_2 \in S$ are called adjacent *to each other w. r. t. $\mathcal{M}$ if $s_1 \in N_{\mathcal{M}}(s_2)$ or $s_2 \in N_{\mathcal{M}}(s_1)$. If $\mathcal{M}$ is symmetric,*

$$s_1 \in N_{\mathcal{M}}(s_2) \;\Leftrightarrow\; s_2 \in N_{\mathcal{M}}(s_1).$$

In that case, define the symmetric neighborhood relation $\mathcal{N}_{\mathcal{M}}$ such that $s_1 \,\mathcal{N}_{\mathcal{M}}\, s_2$ if and only if s_1 and s_2 are adjacent w. r. t. $\mathcal{M}$.

Using the definitions above, energy landscapes can be defined easily:

Definition 4 (Energy landscape)**.** *An* energy landscape *$L = (X, f, \mathcal{M})$ is a tuple consisting of:*

1. *a finite[1] set of states X,*
2. *an energy function $f : X \to \mathbb{R}$ mapping each state $x \in X$ to the real number that represents the state's* energy*, and*
3. *a* move set *$\mathcal{M}$ w. r. t. X.*

In a physical system with states that can be associated with a certain energy, one usually assumes that states of lower energy are visited with a higher probability, i. e. a low energy is a favorable property. If an energy landscape is used to model such a system, it is therefore interesting to know the state or states of minimal energy. Beside this global minimum, local extrema can be defined in a natural way, too, and are an important characteristic:

Definition 5 (Extrema)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape. A state $x \in X$ is called a* local minimum (local maximum*, resp.) if for all $y \in N(x)$ the inequation $f(x) \leq f(y)$ ($f(x) \geq f(y)$, resp.) holds.*

The global *minima (*global *maxima, resp.) are the local minima (local maxima, resp.) with the lowest (highest, resp.) energy.*

[1] Energy landscapes can also be defined to allow for infinite state sets, however, this will not be required in the context of this work.

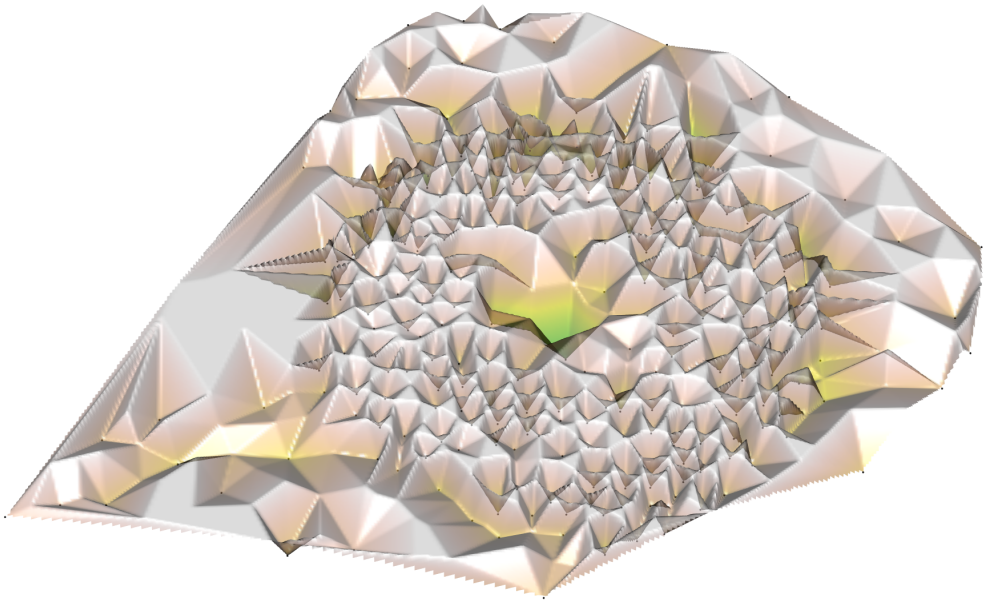


Figure 3.1.: 3-dimensional visualization of an energy landscape. The valleys correspond to the local minima and the peaks between them to the rate between the states. The depth of the valley represents the energy of its local minimum.

The term "landscape" is inspired by some parallels of this theoretical construct to real, natural landscapes. Each state represents a spot on the ground. One can move from one state to any of its neighbors induced by the move set, jumping between different energy levels, just as one would walk up and down when following a trail to cross a hill. The minima and maxima correspond to valley bottoms and mountain peaks, respectively. Figure 3.1 attempts to visualize an energy landscape in three dimensions. It has been constructed by interpreting the local minima of the landscape as nodes of a graph and the rates between them as weighted edges. A 2-dimensional, force-directed layout has been applied to the graph structure and the third dimension was used to encode the energy of the local minima.

Though extrema always exist since X is finite, it can happen that multiple extrema with equal energies are adjacent to each other. Such groups of states will be referred to as *plateaus*. They are an example for artifacts that one often wants to avoid when talking about landscapes, since they cause corner cases in algorithms which can be handled but complicate their description. It is therefore convenient to introduce some more restrictive properties for landscapes (Flamm et al. 2002) that will

lead to a simplified presentation of the following ideas.

Definition 6 (Degeneracy, local invertibility and non-neutrality)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape. L is called ...*

- non-degenerate *if f is injective, i. e. for all $x, y \in X$, $f(x) = f(y)$ implies $x = y$. Otherwise, L is* degenerate*.*
- locally invertible *if the implication $f(x) = f(y) \Rightarrow x = y$ holds for all $x, y, z \in X$ with $x, y \in \{z\} \cup N(z)$.*
- non-neutral *if $f(x) = f(y)$ and $y \in N(x)$ implies $x = y$ for all $x, y \in X$.*

Obviously, non-degeneracy implies local invertibility: if the energy function takes different values for all states, it especially takes different values for the states x, y within the neighborhood of z. Also, local invertibility implies non-neutrality as the special case $x = z$. Note here the subtle difference in the definition of local invertibility in contrast to the cited reference, which arises from the fact that, in this work, the definition of neighborhood of a state x does not include x itself. Further, it is clear that there are no plateaus in a non-neutral landscape, since adjacent states have distinct energies.

A common term to describe natural landscapes is "valley", which refers to the part of the landscape between a single valley bottom and the surrounding peaks. The formalization of this property is slightly more complicated but turns out to be rather useful. First, a notion of walking downhill is required.

Definition 7 (Path)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape. A* path *$P_{x \to y}$ from x to y is a sequence of states $x = x_1, x_2, \ldots, x_{n-1}, x_n = y \in X$ with the following properties:*

1. *$x_k \mapsto x_{k+1}$ is a valid move from $\mathcal{M}$ for all $k \in \{1, \ldots, n-1\}$, and*
2. *$x_i \neq x_j$ for all distinct $i, j \in \{1, \ldots, n\}$.*

As a side note, the former definition induces a notion of connectedness:

Definition 8 (Connectedness)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape and $x, y \in X$. The structures x and y are called* connected *if there are paths $P_{x \to y}$ and $P_{y \to x}$ in L connecting x with y and y with x, respectively.*

The landscape *L is called* connected *if any two structures $x, y \in X$ are connected.*

If $\mathcal{M}$ is a symmetric move set, $P_{y \to x}$ can, of course, be obtained by inverting $P_{x \to y}$.

Returning to the formalization of valleys, one is now interested in paths that descend towards a certain local minimum. The following definition gives rise to a mapping of each state to a local minimum by associating it with a certain path:

Definition 9 (Gradient walk)**.** *Let $P = x_1, \ldots, x_n$ be a path in an energy landscape with energy function f. Then P is called a* gradient walk *if*

1. *x_n is a local minimum, and*
2. *for each move $x_k \mapsto x_{k+1}$ in P,*

$$f(x_{k+1}) = \min \{ f(x) \mid x \in \{x_k\} \cup N(x_k) \}$$

holds.

The second property from the definition means that each move targets one of the neighbors with the lowest energy, and that the energy of the current structure must never increase during the walk. The path follows the direction of the steepest descent, which is also called the *gradient*.

Lemma 1. *Let L be a locally invertible energy landscape. Then all gradient walks are uniquely determined by their initial state.*

In such a landscape, the function that maps each state $x \in X$ to the local minimum determined by the gradient walk starting in x is denoted as $\gamma : X \to M$, where M is the set of local minima of L.

Proof. Since no two neighbors of a structure can have the same energy, the neighbor of minimal energy and therewith the next step in the gradient walk is uniquely determined. Since there are no plateaus, the gradient walk ends in the first local minimum it encounters. Thus, γ is well-defined. □

The preceding lemma is almost trivial, however, it finally allows the formalization of valleys:

Definition 10 (Gradient basin)**.** *Let $L = (X, f, \mathcal{M})$ be a locally invertible energy landscape and $x \in X$. Further, let M be the set of local minima of L and $y \in M$ be the uniquely determined local minimum in which the gradient walk starting in x ends, i. e. $\gamma(g) = y$. The* gradient basin *of x, denoted as $B(x)$, is the set of all structures whose gradient walk also ends in y, i. e.*

$$B(x) = \{ z \in X \mid \gamma(z) = y \}.$$

Definition 10 on page 25 has an important consequence: each structure of a *locally invertible* landscape can be assigned to one and only one gradient basin (or *basin* for short). This allows a *coarse graining* of the landscape that may significantly reduce the number of states while retaining its qualitative properties. This is an advantage if the number of states of a landscape is so large that computations on it would otherwise become infeasible. The following definitions are helpful to formalize the general concept of coarse graining.

Definition 11 (Power set, partition)**.** *Let S be a set. The* power set *of S is the set containing all possible subsets of S, denoted as*

$$\mathcal{P}(S) = \{S' \,|\, S' \subseteq S\}\,.$$

A partition *Ξ of S is a set of subsets of S such that the elements of Ξ are disjoint and their union is S. More formally, $\Xi \subset \mathcal{P}(S)$ with*

$$\forall \alpha, \beta \in \Xi :\ \alpha \cap \beta = \emptyset \ \vee\ \alpha = \beta$$

and

$$\bigcup_{\alpha \in \Xi} \alpha = S.$$

Definition 12 (Macrostates and microstates)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape and $\Xi \in \mathcal{P}(X)$ a partition of X. Then the elements of Ξ are called* macrostates *of L, whereas the elements of X (i. e. the states of L) are also referred to as* microstates*.*

Even though, in general, any partition of a landscape's state set can be understood as a macrostate set, one is usually interested in partitions that are considered to be "sensible". In other words, it should be plausible from a physical point of view that the chosen approach retains the qualitative properties of the landscape. Ideally, this hypothesis is verified later on.

It is also necessary to transform the move set in a proper way to define transitions between macrostates. Given macrostates $\alpha, \beta \in \Xi$, the canonical extension of $\mathcal{M}$ is to allow a move from α to β if, and only if, there are microstates $x \in \alpha$ and $y \in \beta$ such that $x \mapsto y$ is a valid move in L (i. e. if $\exists \mu \in \mathcal{M} :\ \mu(x) = y$).

The following lemma allows the application of the notion of basins to define a set of macrostates for an arbitrary energy landscape:

Lemma 2. *Let $L = (X, f, \mathcal{M})$ be a locally invertible energy landscape and M the set of its local minima. Then the set of all basins*

$$\Xi = \{B(x) \,|\, x \in M\}$$

is a partition of X and therewith a macrostate set of L.

Proof. Due to Lemma 1 on page 25, the gradient basins are well-defined. Since all gradient walks must end in a local minimum $y \in M$ and all gradient walks are uniquely determined, each structure $x \in X$ is contained in exactly one basin. Thus, Ξ is a partition. □

Using gradient basins as macrostates is a good choice for several reasons. The move set can be extended in the canonical way. Since, as mentioned before, a physical system prefers states of lower energy, a gradient walk represents the most likely path in the landscape when starting in an arbitrary microstate. Thus, one can assume that states within the same gradient basin all tend to move towards their associated local minimum, making it a sensible *representative* of this basin. This justifies the definition of the energy of the basin $B(y)$ as the energy E_y of its local minimum y:

Definition 13 (Gradient-induced coarse graining)**.** *Let $L = (X, f, \mathcal{M})$ be a locally invertible energy landscape. Then the* gradient-induced coarse graining *of L yields an energy landscape $\hat{L} = (\Xi, \hat{f}, \hat{\mathcal{M}})$ where:*

1. *$\Xi = \{B(x) \mid x \in M\}$ is the macrostate set induced by the notion of basins associated with the local minima M of L,*
2. *$\hat{f} : \Xi \to \mathbb{R},\ B(x) \mapsto f(x)$ is the energy function for Ξ that maps each basin to the energy of its local minimum x, and*
3. *$\hat{\mathcal{M}}$ is a move set that, for $\alpha, \beta \in \Xi$, allows a transition from α to β if and only if there are microstates $x \in \alpha$ and $y \in \beta$ for which there is a move $\mu \in \mathcal{M}$ with $\mu(x) = y$.*

The definitions above can also be extended for general degenerate landscapes. In that case, the gradient walks are no longer unique, since a state could have several neighbors of the same energy. This ambiguity can be overcome by defining an arbitrary total order on the state set which defines the preferred one of several neighbors with the same energy. In practice, the order of reading from an input file or a lexicographical ordering may be used. Another problem is that there may be several adjacent local minima. Strictly applying Definition 13 would mean to split the associated valley into multiple basins, each containing the states from which a gradient walk descends into the same minimum. Since this is not desired, the basins of adjacent minima are merged and a single representative is chosen among the minima, e. g. by applying the same ordering as described above.

This chapter introduced the notion of *energy landscapes* and their elementary properties such as *extrema*. Further, some restricting terms such as *degenerate* have been defined that simplify the description of methods involving these landscapes by preventing corner cases which are usually easy to handle in practice. A *gradient-based coarse-graining approach* has been used to reduce the number of states of a landscape by partitioning these into *macrostates*. This is accomplished by performing *gradient walks* on all *microstates* and grouping together the states that share the same target minimum.

3.2. RNA energy landscapes

In this section, the definitions from Section 3.1 on page 21 will be applied to the special case of RNA secondary structures to construct a model to analyze RNA folding kinetics.

At first, a formalization of RNA is needed. Since this work is mainly concerned with RNA at the level of secondary structures, all that is needed is an ordered representation of the nucleobases it contains:

Definition 14 (RNA sequence)**.** *A* RNA sequence $s = s_1 \cdots s_n$ *of length* n *is a string over an alphabet representing the nucleobases adenine, uracil, guanine and cytosine, i. e.* $s_i \in \{A, U, G, C\}$ *for all* $i \in \{1, \ldots, n\}$*. Thereby,* s_1 *is the* $5'$*-end and* s_n *is the* $3'$*-end of the RNA.*

As described in Section 2.2 on page 15, a RNA sequence folds up by forming base pairs which constitute the secondary structure of this RNA. Thus, from a formal perspective, a secondary structure is but a set of pairs of sequence indices at which base pairs form. Not all such sets are valid structures, however, as there are additional constraints involved.

Definition 15 (RNA secondary structure)**.** *Let* $s = s_1 \cdots s_n$ *be a RNA sequence. A* RNA secondary structure $x = \{(i_1, j_1), \ldots, (i_m, j_m)\}$ *is a set of ordered pairs of indices with* $i_k, j_k \in \{1, \ldots, n\}$ *and* $i_k < j_k$ *for which the following properties hold:*

1. $\{s_{i_k}, s_{j_k}\}$ *is either one of the* Watson-Crick *base pairs, namely* $\{A, U\}, \{G, C\}$*, or the wobble base pair* $\{G, U\}$*,*

2. i_k *and* j_k *pair only with each other and no other index, i. e.*

$$\nexists (l_1, l_2) \in x : \quad (l_1 \neq i_k \ \wedge \ l_2 = j_k) \ \vee \ (l_1 = i_k \ \wedge \ l_2 \neq j_k),$$

and

3. no other base pair is crossing $\{s_{i_k}, s_{j_k}\}$*, i. e.*

$$\forall (l_1, l_2) \in x : \quad i_k \leq l_1 < l_2 \leq j_k \ \vee \ l_1 \leq i_k < j_k \leq l_2$$

. . . for all $k \in \{1, \ldots, m\}$*. For convenience, each pair of indices* $(i, j) \in x$ *is identified with its associated base pair* $\{s_i, s_j\}$ *and* x *is interpreted as a* set of base pairs*.*

The third property from Definition 15 on page 28 distinguishes *secondary* structures from *tertiary* structures. A structure that satisfies this property is called *pseudo-knot* free; structures that violate it are said to contain pseudo-knots. Though pseudo-knots do appear *in vivo* and are even known to be of functional importance (e. g. as "kissing hairpin" loop, cf. Chang and Tinoco Jr. 1997), they are hard to handle from a computational viewpoint. At least in the very general case, they cannot be considered since they render problems like RNA folding intractable (Akutsu 2000). Nevertheless, secondary structures have proven to be a useful simplification. They are evolutionary conserved and can be used to infer the function of a given RNA or to search for unknown functional RNAs in genome, e. g. using a tool like `Infernal` (Nawrocki and Eddy 2013).

As stated in Definition 15 on page 28, this work mostly uses a notation that treats secondary structures as sets of ordered base pairs. For example, the insertion of a base pair (i, j) into a structure x is expressed as $x \cup \{(i, j)\}$ or, more briefly, as $x \cup (i, j)$. Analogously, removing it would be written as $x \setminus (i, j)$. To denote that a base pair (i, j) or a structural motif y is contained in x, the notations $(i, j) \in x$ or $y \subseteq x$ would be used, respectively. The *open chain*, i. e. the structure that does not contain any base pairs, is represented by the empty set, denoted as $\emptyset$.

Identifying base pairs with indices induces spatial relations among them.

Definition 16 (Spatial relations)**.** *Let* x *be a secondary structure and* $(i, j), (k, l) \in x$ *be base pairs. If* $i < k < l < j$ *holds, then* (i, j) *is said to* enclose (k, l)*, while* (k, l) is enclosed by (i, j)*. If* (i, j) *encloses* (k, l)*, one synonymously says that* (k, l) *lies* inside *of* (i, j)*. If* (i, j) *does* not *enclose* (k, l)*, then* (k, l) *is located* outside *of* (i, j)*.*

Base pair (k, l) *is called* directly enclosed *by* (i, j) *if* $i + 1 = k$ *and* $j - 1 = l$*. In that case, the base pairs* (i, j) *and* (k, l) *are called* adjacent*.*

If $i < j < k < l$ *holds for base pairs* $(i, j), (k, l)$*, then* (i, j) *is said to lie to the* left *or* upstream *of* (k, l) *and,* vice versa*,* (k, l) *is said to lie to the* right *or* downstream *of* (k, l)*.*

Though a secondary structure can be interpreted as a set of base pairs, it is often necessary to describe more complex structures like loops and stems, which may perform biological functions. Using the introduced notation, these can be formalized.

Definition 17 (Structural RNA elements)**.** *Let x be a secondary structure of a RNA $s = s_1 \cdots s_n$ of length n. A* maximal *set of adjacent base pairs $(i+1, j-1), (i+2, j-2), \ldots, (i+k, j-k) \in x$ is referred to as a* stem *of length $k > 0$. Here, "maximal" means that neither (i, j) nor $(i+k+1, j-k-1)$ are present in x. The enclosed part $s_{i+k+1} \cdots s_{i-k-1}$ of s is called a* loop*. If a loop consists only of unpaired nucleobases, it is called a* hairpin *loop.*

It is common to define additional types of loops, however, this is not required for this work.

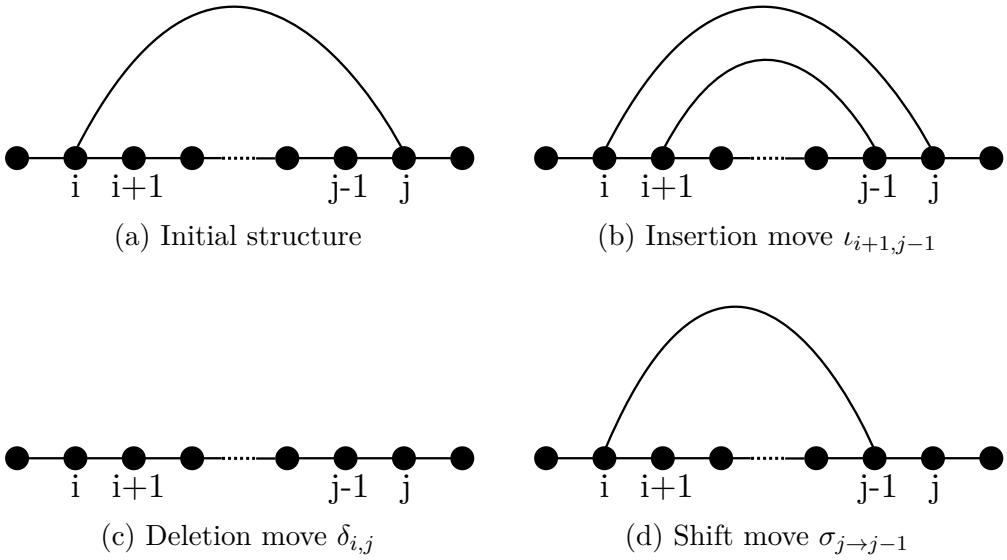
As mentioned before, there is a vast number of possible structures for each RNA sequence. Energy landscapes (Definition 4 on page 22) are a formalism that can be used to model RNA folding kinetics, i. e. to make predictions about the distribution of all possible structures and its development in the course of time. To begin with, the components of the landscape need to be defined in the context of RNAs. First, possible moves for secondary structures are defined.

Definition 18 (Elementary RNA moves)**.** *Let $s = s_1 \cdots s_n$ be a RNA sequence of length n with conformation space X and $x \in X$ a secondary structure of s. The set of* insertions $\mathcal{I} = \{\iota_{ij} \mid 1 \leq i < j \leq n\}$ *is defined as a set of partial mappings $\iota_{ij} : X \dashrightarrow X$ with $x \mapsto x \cup (i, j)$ if both i and j are unpaired and (i, j) is a valid base pair in s. Else, ι_{ij} is undefined on x, i. e. the move is invalid.*

The set of deletions $\mathcal{D} = \{\delta_{ij} \mid 1 \leq i < j \leq n\}$ *is defined as a set of partial mappings $\delta_{ij} : X \dashrightarrow X$ with $x \mapsto x \setminus (i, j)$ if $(i, j) \in x$. Else, δ_{ij} is undefined on x.*

The set of shifts $\mathcal{S} = \{\sigma_{i \to j} \mid 1 \leq i, j \leq n,\ i \neq j\}$ *is defined as a set of partial mappings $\sigma_{i \to j} : X \dashrightarrow X$ that are undefined if i is unpaired in x. If, however, position i is paired with some position k and j and k form a valid base pair in s, then $\sigma_{i \to j}(x) = (x \setminus (i, k)) \cup (j, k)$. If j and k do not form a valid base pair, $\sigma_{i \to j}$ is also undefined on x.*

Examples for each type of move are given in Fig. 3.2 on the next page. Now, it is easy to adapt the notion of energy landscapes to RNAs.



(a) Initial structure

(b) Insertion move $\iota_{i+1,j-1}$

(c) Deletion move $\delta_{i,j}$

(d) Shift move $\sigma_{j \to j-1}$

Figure 3.2.: Elementary RNA moves

Definition 19 (RNA landscapes)**.** *A* RNA landscape $L_s = (X, f, \mathcal{M})$ *for a given RNA sequence s is an energy landscape where*

1. *the state set X is a set of secondary structures of s, usually all structures compatible with s,*
2. *the energy function f is a RNA energy model with $f(\emptyset) = 0$, usually the* Turner *energy model, and*
3. *the move set $\mathcal{M}$ is a RNA move set, usually allowing insertions, deletions and shift moves of single base pairs.*

The global minimum of f on X is called the minimum free energy *of s w. r. t. f.*

Unless explicitly stated otherwise, RNA landscapes are meant to have the default properties mentioned in the definition above. The energy function is usually defined to assign the open chain an energy value of zero. This choice is arbitrary but a general convention. Common alterations of this definition are additional structure constraints or alternative move sets. For example, sometimes shift moves are not wanted as they are more complicated than simple insertions and deletions and also because they void the equivalence of move distance and base pair distance (cf. Definition 32 on page 52). Structural constraints can be used for different purposes. A

functional reason might be the consideration of bound molecules which force a certain structure on a part of the sequence, as is the case for riboswitches (cf. Section 2.3 on page 16). A more pragmatic reason is the reduction of the number of microstates to reduce the computational complexity and thus the computation time. Of course, such constraints are an approximation that reduces the accuracy of the model. Therefore, it must be justified why their use is sensible and their side effects need to be considered carefully. A simplification used extensively throughout this work is to avoid structures that contain so-called *lonely base pairs*, as described in Chapter 4 on page 41. RNA landscapes may be coarse-grained using the gradient-based approach described in Definition 13 on page 27.

3.3. The probability of RNA secondary structures

As mentioned before, secondary structures that possess a low energy are energetically favored, i. e. the RNA molecule is more likely to fold into a structure that yields a lower energy. Also, these structures are more stable and so refolding into another structure is more unlikely.

According to the Turner energy model (Mathews et al. 2004), a widely accepted model to reliably predict the energy of RNA molecules of moderate length, the energy of the molecule is approximately the sum of the energy contributions of its structural compounds. Stabilizing structural elements yield a negative energy while unfavorable ones have positive energy contributions. One example are base pair stems, which increasingly stabilize the structure with increasing length. In contrast, small hairpin loops have a stabilizing contribution, while larger ones destabilize the molecule. Since there is no absolute reference value for the energy of a molecule, the energy model is normalized such that an open RNA chain is associated with an energy of zero.

When a RNA is transcribed from DNA (Section 2.2 on page 15), it is synthesized as an unfolded, open chain. However, since this conformation is energetically unfavorable, the RNA immediately begins to fold itself while it is still being transcribed by the RNA polymerase. Even though the *folding path*, i. e. the exact sequence of structures the RNA will adopt during a certain time, is random, neighboring structures with a lower energy are adopted much more frequently since the activation energy for this refolding reaction is lower. Assuming that one initially has an infinite

or at least very large amount of molecules, one can therefore describe the distribution of the RNA molecules across all possible structures in the course of time. This function, which is specific for each RNA sequence s, fully describes the *folding kinetics* of s. After an infinite amount of time, the distribution of the molecules in each state does not change anymore and the *dynamic equilibrium* is reached. The distribution in this state is called the *equilibrium distribution* of the secondary structures of s. The equilibrium is called "dynamic" since there are still transitions from each conformation to its neighbors, however, the rate of reaction is equal for the forward and backward reaction such that the net change of concentrations is zero. In practice, due to the high rate of RNA folding (cf. Section 6.1 on page 70) the equilibrium is reached quite fast for many sequences. However, it is possible that RNAs form suboptimal intermediate structures on their folding path that are very stable compared to their immediate neighbors. Such structures are called *kinetic traps*, since it can take a long time until a molecule is able to leave this conformation. In the presence of kinetic traps the equilibration process might take so long that in practice the RNA is degraded before it is completed. Also, time-critical processes involving RNAs may also be effected by kinetic effects. Therefore, a *thermodynamical* analysis of the structure space, i. e. one based merely on the equilibrium distribution, is not sufficient to fully explain the functionality of RNA in biological systems.

The equilibrium distribution of a RNA sequence can be calculated directly from the energies of the all its secondary structures.

Definition 20 (Boltzmann weight, partition function)**.** *Let X be a set of RNA secondary structures and $x \in X$ be a structure of energy E_x. Then the term*

$$w(x) = \exp\left(-\frac{E_x}{RT}\right)$$

is called the Boltzmann weight *of x, where R is the universal gas constant and T is the absolute temperature.*

The sum

$$\mathrm{Z}[\,X\,] = \sum_{x \in X} w(x)$$

is called the partition function *of X.*

The following theorem is an elementary result of physical chemistry (McQuarrie and Simon 1997):

Theorem 1. *Let X be the structure ensemble of a given RNA sequence s and $Y \subseteq X$. Then the probability that a RNA molecule of sequence s randomly chosen from an infinite, equilibrated population has adopted a structure $x \in S$ is given by*

$$\Pr[\, Y \mid X \,] = \frac{\mathrm{Z}[\, Y \,]}{\mathrm{Z}[\, X \,]}$$

In the equilibrium state, the RNA molecules are said to follow the BOLTZMANN *distribution*. As can be observed in Definition 20, the BOLTZMANN weight and therewith the probability of a structure decreases exponentially with its associated energy.

3.4. Enumeration of RNA landscapes

Performing computations on a RNA landscape $L_s = (X, f, \mathcal{M})$ for some RNA sequence s of length n requires knowledge about the state set X and the moves that are valid in L. The number of different secondary structures of s, however, grows exponentially in n (Hofacker et al. 1998) such that an exhaustive enumeration of all possible microstates is infeasible even for short RNAs. Note that a gradient-induced coarse graining (Definition 13 on page 27) of the landscape is not helpful here as it requires the microstates to be computed beforehand.

There are different approaches to alleviate this problem. A first measure is to reduce the number of microstates by additional structural constraints. As described in more detail in Chapter 4 on page 41, structures containing isolated base pairs are usually energetically unfavorable when compared to their related structure without any isolated base pairs. It is possible and practiced throughout the implementations of this work to only include structures in X that do not contain isolated base pairs, so called *canonical structures* (Definition 23 on page 41). Of course, this also requires an adaptation of the move set (cf. Section 4.2 on page 42). Even though this approach dramatically reduces the size of X, there are still too many structures too generate all of them.

Since structures with very high energies are not likely to form at all (cf. Section 3.3 on page 32), a radical approach to reduce the size of X is to simply discard any structures that have an energy exceeding a certain threshold ΔE above the minimum free energy of s.

Definition 21 (ΔE-pruned state set)**.** *Let $L = (X, f, \mathcal{M})$ be an energy landscape with global energy minimum $E_{\min} = \min\{f(x) \mid x \in X\}$ and $\Delta E > 0$ a positive energy value. Then*

$$X|_{\Delta E} = \{x \in X \mid f(x) \leq E_{\min} + \Delta E\}$$

is called the ΔE-pruned state set *of L and ΔE is referred to as its* exploration threshold*.*

Since the energies of secondary structures are, at least asymptotically and under simplifying assumptions, normally distributed (Clote et al. 2009), this measure is very effective and leads to a feasible size of $X|_{\Delta E}$, depending on how ΔE is chosen. It is also very efficient since $X|_{\Delta E}$ can be fully enumerated using a dynamic programming algorithm as described by Wuchty et al. (1999) and implemented in the tool `RNAsubobt`. This method is used throughout this work, though it *does* have a major drawback. Due to computational limitations, ΔE may need to be chosen so small that interesting conformations are pruned. This problem is solved in Chapter 5 on page 59.

Of course, there are numerous other possible approaches that were not used in this work. One example is based on the concept of a *basin hopping graph* (Kucharík et al. 2014). In this stochastic method, samples are drawn from the set of local minima which can be enumerated efficiently. Then, the minima are connected by using an iterative direct path heuristic, yielding a graph-like structure. Another concept is the *shape* abstraction of structures that does not consider single base pairs, but more abstract structural features like helices and loops without specifying their exact position in the structure (Giegerich et al. 2004; Huang et al. 2012).

3.5. Basic natural laws and principles

This work relies on some fundamental natural laws and principles, e. g. to describe the speed of chemical reactions. They are summarized in this section.

3.5.1. The principle of detailed balance

The principle of *detailed balance* is a general property of many systems that are composed of elementary processes, e. g. for reversible, *elementary* chemical reactions (McQuarrie and Simon 1997). Here “elementary” means

that the reaction occurs in a single step, without the formation of any intermediate products. It states that, once the system is equilibrated, the net transition rate between any two states of the system is zero, i. e. the transition rate from state a to state b is equal to the transition rate from state b to a. More formally,

$$\Pr[\, a \mid t \to \infty \,] \cdot r_{b \leftarrow a} = \Pr[\, b \mid t \to \infty \,] \cdot r_{a \leftarrow b} \tag{3.1}$$

for any two states a and b of the system, where $\Pr[\, \circ \mid t \to \infty \,]$ is the probability of a state after infinite time (i. e. at the equilibrium) and $r_{\circ \leftarrow \circ}$ is the transition rate coefficient for the respective transition.

3.5.2. The rate laws and their coefficients

The kinetics of chemical reactions can be described with simple relationships called *rate laws* (McQuarrie and Simon 1997; Mortimer 2002). More precisely, the *speed* of a given reaction can be inferred from the concentration of the reaction's reactants, its *stoichiometry* and a *reaction rate constant*. In general, however, this is only possible for *elementary* reactions. A reaction is called elementary if no intermediate products arise. Given a number of reactant species $\mathrm{R}_1, \ldots, \mathrm{R}_\mathrm{n}$, the speed or *rate* of reaction $\mathfrak{r}_i$ for each species can be quantified as the change of its concentration $[\mathrm{R_i}]$ over the time t, i. e. $\mathfrak{r}_i = {}^{d[\mathrm{R_i}]}/_{dt} = [\dot{\mathrm{R}}_\mathrm{i}]$.

Stated in its general form, i. e. for m reactions, the rate law is given by

$$[\dot{\mathrm{R}}_\mathrm{i}] = \sum_{j=1}^{m} \nu_{ij} r_j \prod_{k=1}^{n} [\mathrm{R_k}]^{-\nu_{kj} \cdot I_{<0}(\nu_{kj})}, \tag{3.2}$$

using the notation described above and, additionally, the constant rate coefficients r_j for reaction j as well as the *stoichiometric numbers* ν_{kj}, i. e. the number of molecules of species k in a "single reaction event" of reaction j. By convention, the stoichiometric numbers of the products are positive numbers while the ones of the reactants are negative. The indicator function

$$I_{<0}(\nu) = \begin{cases} 1 & \text{if } \nu < 0, \\ 0 & \text{else,} \end{cases}$$

ensures that only the concentrations of actual reactants of that specific reaction are taken into account. Equation (3.2) looks quite complex, but for simple reactions like e. g.

$$\mathrm{A} \longrightarrow \mathrm{X} \qquad\qquad \mathrm{B} + \mathrm{C} \longrightarrow \mathrm{Y}$$

where all stoichiometric numbers are one, the rate law reduces to

$$[\dot{\mathrm{A}}] = r_1 \cdot [\mathrm{A}] \qquad\qquad [\dot{\mathrm{B}}] = [\dot{C}] = r_2 \cdot [\mathrm{B}] \cdot [\mathrm{C}]$$

respectively. For a single reaction, the absolute value of the sum of all the reactants' stoichiometric numbers is referred to as the *order* of that reaction. In the examples above, the order of the reactions are one and two, respectively.

One aspect that still needs to be considered is the form of the rate coefficients. For elementary reactions, they can be described using the Arrhenius *equation*

$$r = c \cdot \exp\left(-\frac{E_A}{RT}\right), \tag{3.3}$$

where R is the universal gas constant, T is the absolute temperature, $E_A > 0$ is the *activation energy* of the reaction and c is the *pre-exponential factor*. The activation energy of the energy is the external energy required to perform the necessary conformational changes, e. g. to break up chemical bonds. If $E_A = 0$, then no additional energy is required to start the reaction. Put that way, the exponential term of Eq. (3.3) can be interpreted as a probability that the reaction occurs when the required species' particles collide in an appropriate orientation. With higher activation energies, the probability of the reaction decreases exponentially, e. g. because the speed of the colliding particles is too low. The pre-exponential factor, can be interpreted as the rate at which particles of the reactant species collide, multiplied with a *steric factor*. It accounts for the fact that particles also need to collide with a specific orientation and at certain domains of the reactants. Both the collision rate and the steric factor are very hard to estimate theoretically. Therefore, in practice, the pre-exponential factor is simply used as a "fudge factor" to fit the Arrhenius equation to the experimentally measured values.

Substance concentrations have the unit *molar* ($1\,\mathrm{M} = 1\,\mathrm{mol\,L^{-1}}$). As the reaction rate is the derivative of a concentration w. r. t. time, it therefore needs to have the unit $\mathrm{M\,s^{-1}}$. However, for reactions of different orders, the unit of the product of concentrations will differ. Therefore, the unit of the rate coefficient has to be chosen adequately to cancel out the unwanted units. This seemingly odd practice can be explained when deriving the rate law using the more general concept of *activities* from chemical thermodynamics instead of simple concentrations. This extends the applicability of the rate law to special cases not covered by Eq. (3.2) on page 36, however, is not necessary here and beyond the scope of this work.

3.5.3. The law of mass action

Consider a reversible chemical reaction

$$\alpha \cdot \mathrm{A} + \beta \cdot \mathrm{B} + \ldots \rightleftharpoons \sigma \cdot \mathrm{S} + \tau \cdot \mathrm{T} + \ldots$$

of reactants $\mathrm{A}, \mathrm{B}, \ldots$ into products $\mathrm{S}, \mathrm{T}, \ldots$ with stoichiometric numbers $\alpha, \beta, \ldots$ and $\sigma, \tau, \ldots$, respectively. Both the forward and the backward reaction have a specific reaction rates r_{for} and r_{back} which depend on the concentration of the reactants and products, respectively. After a certain amount of time, the reaction reaches a so-called *dynamic equilibrium*, i. e. a state where $r_{\text{for}} = r_{\text{back}}$ and therefore the concentrations of the species participating in the reaction remain constant. In this state, the *law of mass action* dictates

$$K = \frac{[\mathrm{A}]^{\alpha} \cdot [\mathrm{B}]^{\beta} \cdots}{[\mathrm{S}]^{\sigma} \cdot [\mathrm{T}]^{\tau} \cdots}, \tag{3.4}$$

where K is a temperature dependent constant that is characteristic for each specific reaction (McQuarrie and Simon 1997). Further, K may also be influenced by other conditions, e. g. the concentration of other substances, even if they are not actively participating in the reaction.

3.6. Calculation of the ligand binding bonus energy

The atoms in a molecule or chemical compound are being held together by different kinds of chemical bonds between their atoms, e. g. covalent bonds, ionic bonds or hydrogen bonds. To decompose a compound into its components, these bonds have to be broken up which requires a certain amount of energy. In that sense, chemical bonds store potential energy and therewith stabilize the compound. Also, the overall energy of a compound is often lower than the sum of energies of its components. In the case of a riboswitch and its ligand (Section 2.3 on page 16), the formation of a dimer complex stabilizes the riboswitch by adding a certain fixed, negative energy contribution $\theta_L < 0$ to the RNA's free energy. Thus, for some secondary structure x containing the ligand's binding pocket, the energy of the dimer complex Lx is given by

$$E(Lx) = E(x) + \theta_L.$$

To perform computations on the riboswitch, θ_L needs to be determined first. As described in Section 2.3 on page 16, a riboswitch consists of an

aptamer and terminator sequence. For synthetic riboswitches and a given ligand, the aptamer is often sought using an experimental approach called *SELEX* (cf. Section 2.4 on page 18).

To obtain the bonus energy of binding, the aptamer can be mixed with the ligand and the resulting concentrations of the aptamer [A], the ligand [L] and the dimer complex [LA] be measured. From these values, the *dissociation constant* K_D can be calculated by applying the law of mass action (Eq. (3.4) on page 38):

$$K_D = \frac{[\mathrm{A}][\mathrm{L}]}{[\mathrm{LA}]} \tag{3.5}$$

The dissociation constant is the inverse of the reaction's equilibrium constant and measures how fast the ligand dissociates from the aptamer. To obtain the binding energy, previous approaches directly used the measured K_D to obtain θ_L by setting

$$\theta_L^{\mathrm{old}} = RT \ln K_D.$$

However, a more careful approach considers that only a part of the ensemble of the aptamer does in fact contain the binding pocket that is necessary to bind the ligand. Write A^+ for the aptamer species that contains the binding pocket p and identify A and A^+ with the respective sets of secondary structures. Then, Eq. (3.5) can be rewritten as

$$\begin{aligned} K_D &= \frac{\mathrm{Z}[\,A\,] \cdot \mathrm{Z}[\,L\,]}{\mathrm{Z}[\,LA^+\,]} \\ &= \frac{\mathrm{Z}[\,A\,]}{\mathrm{Z}[\,A^+\,] \cdot \exp(-\mathfrak{b}\theta_L)}. \end{aligned}$$

This equality follows from the fact that, after equilibration, the structures are distributed according to their Boltzmann weight (cf. Section 3.3 on page 32). Furthermore, the partition function of the ligand is one since it is assumed that there is only a single ligand conformation, which is a simplification to keep the model simple. By rewriting the last equation, one obtains

$$\theta_L = RT \left(\ln \frac{\mathrm{Z}[\,A^+\,]}{\mathrm{Z}[\,A\,]} + \ln K_D \right),$$

where $\mathrm{Z}[\,A^+\,]/\mathrm{Z}[\,A\,] = \Pr[\,p \mid A\,]$ is the probability of the binding pocket p in the ensemble of the aptamer. If this probability is low, then the ligand has to bind the remaining structures containing p stronger to to achieve

the same K_D and thus the binding energy θ_L has to be higher. In practice however, the difference between θ_L^{old} and θ_L is small if the aptamer is selected by SELEX since this method maximizes the probability that the aptamer binds the ligand and thus $\Pr[\, p \mid A \,] \approx 1$.

Chapter 4.

Canonical RNA Landscapes

This chapter covers *canonical* RNA landscapes, a special type of RNA landscape with additional structural constraints for its state set. These constraints are based on the notion of isolated base pairs (Bompfünewerer et al. 2007) which, according to the TURNER energy model (Turner and Mathews 2009), result in an energy penalty, i. e. adding them to a structure increases its energy.

4.1. Preliminaries

The following definitions precisely describe when a base pair is said to be isolated.

Definition 22 (Inside-loneliness, outside-loneliness, loneliness)**.** *Let x be a RNA secondary structure. A base pair $(i, j) \in x$ is called* inside-lonely *if $(i+1, j-1) \notin x$. Further, $(i, j) \in x$ is called* outside-lonely *if $(i-1, j+1) \notin x$. If (i, j) is both inside-lonely and outside-lonely, it is called* lonely*.*

A base pair $(i, j) \in x$ is said to grow lonely *w. r. t. to the deletion of one of its neighbors $(i-1, j+1)$ or $(i+1, j-1)$ if (i, j) is lonely in $x \setminus (i-1, j+1)$ or $x \setminus (i+1, j-1)$, respectively.*

Informally, a base pair is lonely if it is not adjacent to another base pair on either of its sides. This definition extends to secondary structures and RNA landscapes as follows:

Definition 23 (Canonical secondary structures)**.** *Let x be a RNA secondary structure. x is called a* canonical *structure if it does not contain any lonely base pairs. A structure that contains at least one lonely base pair is said to be* non-canonical*.*

For a set of secondary structures X, its associated set of canonical structures is denoted as $X|_{can} = \{x \in X \mid x \text{ is canonical}\}$.

Definition 24 (Canonical RNA landscape)**.** *Let $L_s = (X, f, \mathcal{M})$ be an RNA landscape of some RNA sequence s. Then L_s is called a* canonical RNA landscape *if X exclusively contains canonical structures.*

The motivation of this definition is to reduce the conformation space by removing structures that are energetically unfavorable. The assumption is that a non-canonical structure has merely a transient role during the folding process from one canonical structure to another. It can therefore be neglected without too much impact on the qualitative properties of the landscape.

A huge fraction of the structures of a general RNA landscape are non-canonical. Clote et al. (2009) proved that the asymptotic number of canonical secondary structures is only $2.1614 \cdot n^{-3/2} \cdot 1.96798^n$. The asymptotic number of general secondary structures, in contrast, is given by $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$, where n is the sequence length. Thus, excluding non-canonical structures from RNA landscapes allows feasible computations on much larger RNA and is therefore of high interest.

4.2. Symmetrical canonical move sets

Definition 24 looks very similar to the definition of general RNA landscapes introduced earlier (cf. Definition 19 on page 30). However, there is an inconspicuous yet highly significant detail to be considered here. The move set $\mathcal{M}$ needs to be defined in a manner that it only generates canonical structures. Simply defining all moves that generate lonely structures to be invalid is no sufficient solution since then e. g. the open chain is an isolated structure. That is because the only possible moves are insertions of single base pairs which lead to non-canonical structures. Any sensible move set must yield a connected landscape as a RNA molecule can fold into any conformation, so another approach is necessary.

Existing move sets for canonical landscapes, e. g. as implemented in current versions of the tool `barriers` (Flamm et al. 2002), have the described problem and also the additional issue of non-symmetry. The latter is the result of how `barriers` handles the enforcement of the canonicality constraint. To be able to remove any stems at all, any time a deletion generates a lonely pair, it is removed from the structure, too. Since a deletion can generate up to two lonely base pairs, this move set may remove up to three base pairs with a single move. However, there is no counterpart for this transition that would allow the insertion of two or even three base pairs. This asymmetry destroys the property of detailed balance (cf.

Section 3.5.1 on page 35) that a system of reversible chemical reactions like RNA folding should have. This follows immediately from Eq. (3.1) on page 36: if a move from a structure x to a structure y is valid, but the reversed move from y to x is not, then the rate coefficient $r_{x \leftarrow y}$ vanishes but $r_{y \leftarrow x}$ does not. Since in a connected RNA landscape, the equilibrium probability of all structures is greater than zero, the equation obviously does not hold.

4.2.1. Defining a symmetric, canonical move set

Let $s = s_1 \cdots s_n$ be a RNA sequence of length n and $L_s = (X, f, \mathcal{M})$ its RNA landscape. Building on Definition 18 on page 30, a *canonical move set*, i. e. a move set mapping canonical structures to other canonical structures only, is now defined.

Definition 25 (Canonical restriction)**.** *For some $x \in X$, let $\mu \in \mathcal{M}$ be a move $\mu : X \dashrightarrow X,\ x \mapsto \mu(x)$. Its* canonical restriction *is defined as the move $\mu|_{can} : X|_{can} \dashrightarrow X|_{can}$ with μ*

$$\mu|_{can}(x) = \begin{cases} \mu(x) & \text{if } \mu(x) \text{ is defined and } x, \mu(x) \in X|_{can}, \\ \text{undefined} & \text{else.} \end{cases}$$

As mentioned, only restricting the standard moves to canonical landscapes does not yield a connected landscape. Therefore, the following definitions adds some additional possibilities to remedy the issue. However, care is taken to allow these additional moves only in cases where they are necessary to minimize the influence on qualitative properties of the landscape.

Definition 26 (Canonical insertions)**.** *Let $\mathcal{I}$ be the set of insertions on s. The set of* canonical insertions *is defined as*

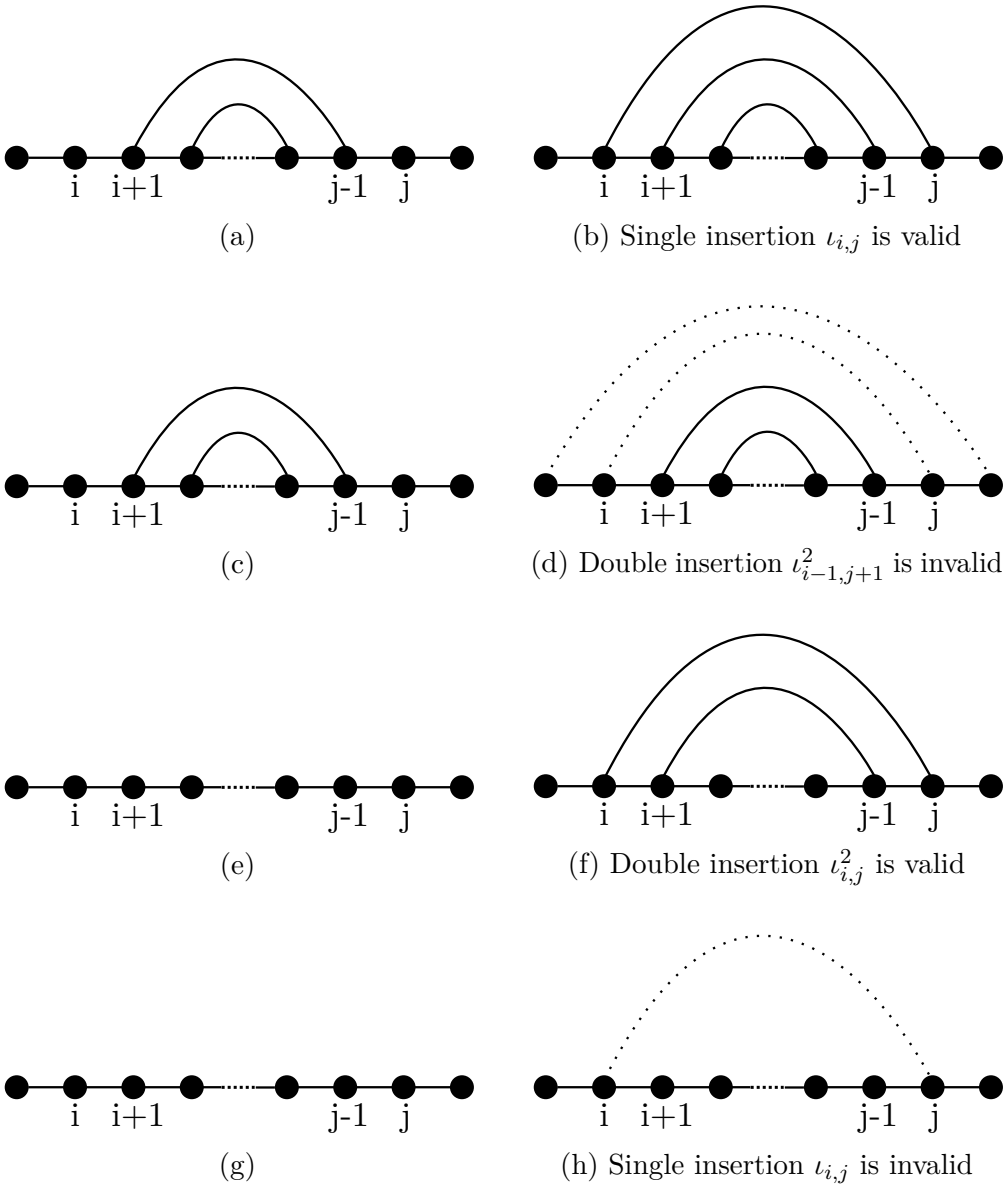
$$\mathcal{I}_{can} = \left\{ \iota_{ij}|_{can} \,\middle|\, \iota_{ij} \in \mathcal{I} \right\} \cup \mathcal{I}_2,$$

where

$$\mathcal{I}_2 = \left\{ \iota^2_{ij} \,\middle|\, 1 \leq i < j < n \right\}$$

denotes the set of canonical double insertions*. These are partial mappings $\iota^2_{ij} : X|_{can} \dashrightarrow X|_{can}$ with $\iota^2_{ij}(x) = \iota_{ij}(\iota_{i+1,j-1}(x))$, if the following conditions hold:*

1. *$\iota_{i,j}$ and $\iota_{i+1,j-1}$ are* valid *moves on x, and*
2. *$\iota_{i,j}|_{can}$ and $\iota_{i+1,j-1}|_{can}$ are* invalid *moves on x.*



(a)

(b) Single insertion $\iota_{i,j}$ is valid

(c)

(d) Double insertion $\iota^2_{i-1,j+1}$ is invalid

(e)

(f) Double insertion $\iota^2_{i,j}$ is valid

(g)

(h) Single insertion $\iota_{i,j}$ is invalid

Figure 4.1.: Canonical insertion moves. In each row, a structure is shown before (*left*) and after (*right*) applying a move.

Otherwise, $\iota^2_{ij}(x)$ is undefined.

Basically the above definition restricts arbitrary insertions to those generating canonical structures. Additionally, it allows double insertions, i. e. insertions of two consecutive base pairs $(i,j),(i+1,j-1)$ at once. However, by property 2, these are valid only if this stack of length two will be isolated, i. e. if (i,j) will be outside-lonely and $(i+1,j-1)$ will be inside-lonely. Figure 4.1 on page 44 gives some examples of canonical insertions. Canonical *deletions* can be defined in a similar way:

Definition 27 (Canonical deletions)**.** *Let $\mathcal{D}$ be the set of deletions on s. The set of* canonical deletions *is defined as*

$$\mathcal{D}_{can} = \left\{ \delta_{ij}|_{can} \,\middle|\, \delta_{ij} \in \mathcal{D} \right\} \cup \mathcal{D}_2,$$

where

$$\mathcal{D}_2 = \left\{ \delta^2_{ij} \,\middle|\, 1 \leq i < j < n \right\}$$

denotes the set of canonical double deletions*. These are partial mappings $\delta^2_{ij} : X|_{can} \dashrightarrow X|_{can}$ with $\delta^2_{ij}(x) = \delta_{ij}(\delta_{i+1,j-1}(x))$, if the following conditions hold:*

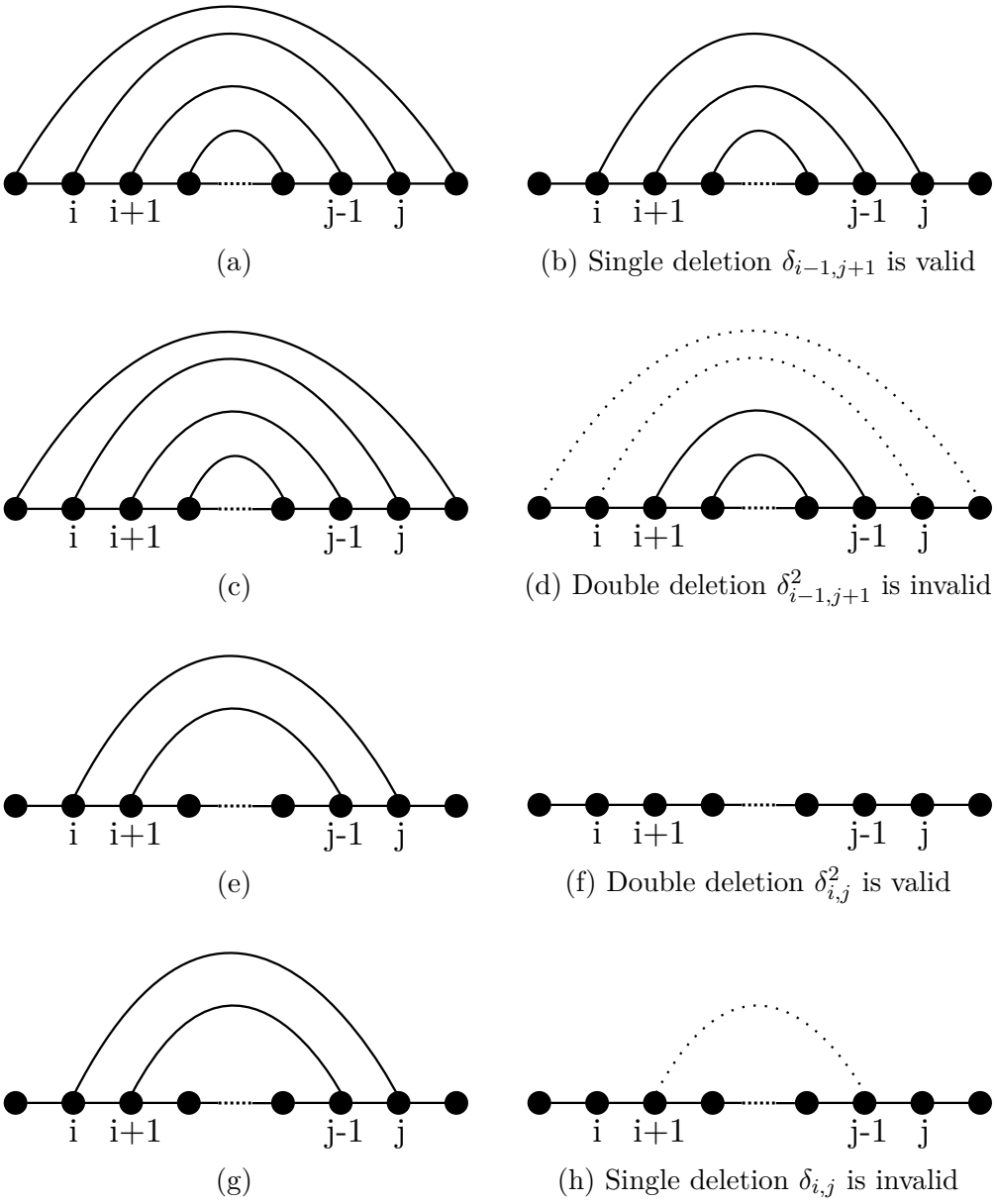
1. *$\delta_{i,j}$ and $\delta_{i+1,j-1}$ are* valid *moves on x, and*
2. *$\delta_{i,j}|_{can}$ and $\delta_{i+1,j-1}|_{can}$ are* invalid *moves on x.*
3. *$\delta_{ij}(\delta_{i+1,j-1}(x)) \in X|_{can}$*

Otherwise, $\delta^2_{ij}(x)$ is undefined.

Analogously, in addition to restricting the deletions to canonical structures, the given definition allows the deletion of lonely stacks of length two. Additionally, property 3 is required to prevent a removal of $(i,j),(i+1,j-1)$ in the middle of a stack of size four, in which case $(i-1,j+1)$ and $(i+2,j-2)$ would grow lonely. Examples are given in Fig. 4.2 on the next page. Finally, canonical *shift moves* can be defined easily by simply restricting the ordinary shifts to canonical landscapes.

Definition 28 (Canonical shifts)**.** *Let $\mathcal{S}$ be the set of shifts on s. The set of* canonical shifts *is defined as*

$$\mathcal{I}_{can} = \left\{ \sigma_{ij}|_{can} \,\middle|\, \sigma_{ij} \in \mathcal{I} \right\}.$$



(a)

(b) Single deletion $\delta_{i-1,j+1}$ is valid

(c)

(d) Double deletion $\delta^2_{i-1,j+1}$ is invalid

(e)

(f) Double deletion $\delta^2_{i,j}$ is valid

(g)

(h) Single deletion $\delta_{i,j}$ is invalid

Figure 4.2.: Canonical deletion moves. In each row, a structure is shown before (*left*) and after (*right*) applying a move. Note that the bases $i-2$ and $j+2$, which are not depicted in the figures above, are assumed to be unpaired.

As a side note, restricting the shifts to canonical structures reduces the number of possible neighbor structures (w. r. t. shift moves) of a structure $x|_{\text{can}}$ to at most six. This is because of the constraint of pseudo-knot absence in secondary structures. Given a base pair (i, j), one can only ...

- shift i to the left (i. e. decrease i) to the next enclosing base pair $(k, j+1)$, $k < i$, if it exists (*outside shift*),
- shift i to the right (i. e. increase i), but keep $i' < j$, to the left side of an enclosed base pair $(k, j-1)$, $k > i$, if it exists (*inside shift*),
- shift i to the right of j to the next enclosing base pair $(i-1, k)$, $k > j$, if it exists (*cross shift*), or
- shift j analogously.

These possibilities are shown in Fig. 4.3 on the following page.

Combining all these definitions, one finally gets a complete move set.

Definition 29 (Canonical move set)**.** *Let $\mathcal{I}_{can}, \mathcal{D}_{can}$ and $\mathcal{S}_{can}$ be the sets of canonical insertions, deletions and shifts on a RNA sequence s, respectively. Then the* canonical move set $\mathcal{M}_{can}$ *for s is given by their union:*

$$\mathcal{M}_{can} = \mathcal{I}_{can} \cup \mathcal{D}_{can} \cup \mathcal{S}_{can}.$$

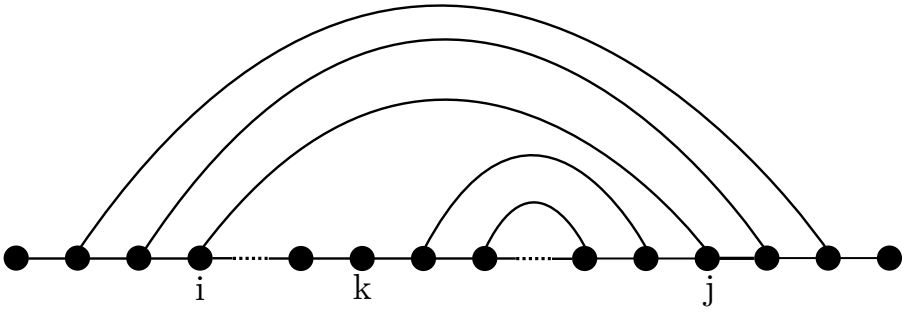
4.2.2. Properties of the canonical move set

In this section, the symmetry of the canonical move set will be proved. Further, it will be shown that the canonical landscape with its associated canonical move set is connected. But initially, it will be shown that the canonical move set does indeed only generate canonical structures. Throughout the section, let $\mathcal{M}_{\text{can}}$ be the canonical move set of an arbitrary RNA sequence s with conformation space X.

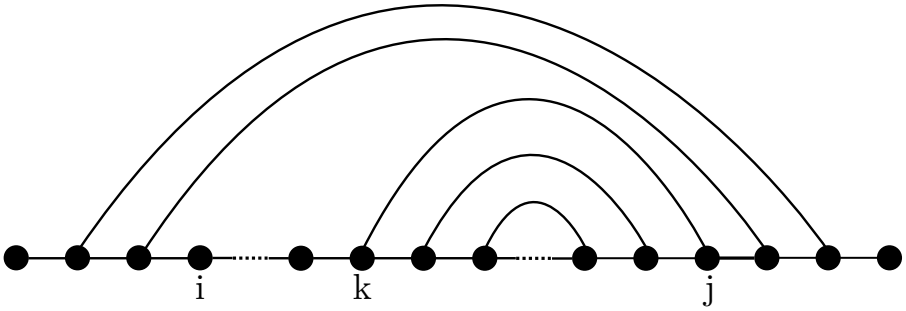
Lemma 3. *The canonical move set $\mathcal{M}_{can}$ exclusively generates canonical structures when applied to $X|_{can}$, i. e. $\forall \mu \in \mathcal{M}_{can} : \ \mu(X_{can}) \subseteq X|_{can}$.*

Proof. The canonical restriction of any move can by definition only generate canonical structures. Therefore, the claim is trivial for canonical shift moves and single base pair insertions or deletions and remains to be shown only for canonical double insertions ι^2_{ij} and double deletions δ^2_{ij}.

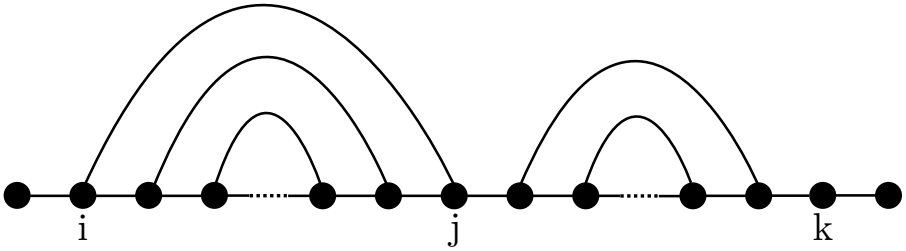
Since the insertion of two consecutive base pairs cannot generate a lonely pair, the claim holds for ι^2_{ij}. By definition, for any structure x, $\delta^2_{ij}(x)$ is defined only if $\delta_{ij}(\delta_{i+1,j-1}(x)) \in X_{\text{can}}$. □



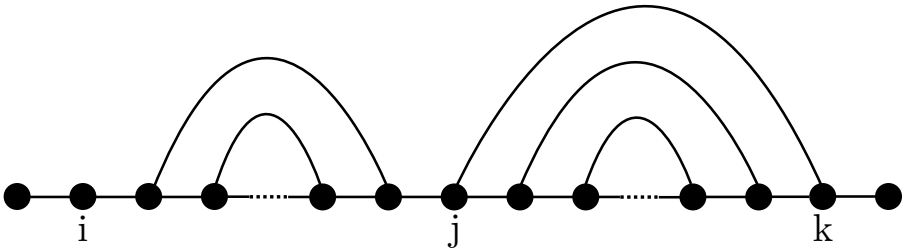
(a) Application of $\sigma_{k\to i}$ to the structure below



(b) Application of $\sigma_{i\to k}$ to the structure above



(c) Application of $\sigma_{k\to i}$ to the structure below



(d) Application of $\sigma_{i\to k}$ to the structure above

Figure 4.3.: Canonical shift moves

Now, it is shown that $\mathcal{M}_{\text{can}}$ is a symmetric move set (Definition 2 on page 21).

Theorem 2 (Symmetry)**.** *For all structures $x \in X|_{can}$ and any valid move $\mu \in \mathcal{M}_{can}$ on it, there is an* inverse move *$\mu^{-1} \in \mathcal{M}_{can}$ such that $\mu^{-1}(\mu(x)) = x$.*

Proof. Let $y = \mu(x)$. By Lemma 3 on page 47, $y \in X|_{\text{can}}$.

For $\mu = \iota_{ij}|_{\text{can}}$, $y = x \cup (i,j)$. Obviously, the inverse move is $\mu^{-1} = \delta_{ij}|_{\text{can}}$ since $\delta_{ij}|_{\text{can}}(x \cup (i,j)) = x$. Of course, $(\mu^{-1})^{-1} = \mu$, so $\delta_{ij}|_{\text{can}}^{-1} = \iota_{ij}|_{\text{can}}$.

If $\mu = \sigma_{i \to j}$, then[1] $y = (x \setminus \{i,k\}) \cup \{j,k\}$, where k is the position of the base in x pairing with position i. The shift can be reverted by applying another shift $\sigma_{j \to i}$ since

$$\sigma_{j \to i}(y) = \Big(\big((x \setminus \{i,k\}) \cup \{j,k\} \big) \setminus \{j,k\} \Big) \cup \{i,k\} = x.$$

For $\mu = \iota^2_{ij}$, $y = x \cup \{(i,j),(i+1,j-1)\}$. It is clear that the inverse move is $\mu = \delta^2_{ij}$ since

$$\delta^2_{ij}(y) = \Big(x \cup \big\{ (i,j),(i+1,j-1) \big\} \Big) \setminus \big\{ (i,j),(i+1,j-1) \big\} = x,$$

however, it is not obvious that δ^2_{ij} is valid on y. To verify that, consider that ι^2_{ij} is valid on x only if $\iota_{i,j}|_{\text{can}}$ and $\iota_{i+1,j-1}|_{\text{can}}$ are *invalid* on x, and so $(i-1,j+1),(i+2,j-2) \notin x,y$. Therefore, $\delta_{i,j}|_{\text{can}}$ and $\delta_{i+1,j-1}|_{\text{can}}$ are invalid on y since removing one of the base pairs $(i,j),(i+1,j-1)$ would result in the other one growing lonely. Thus, δ^2_{ij} is valid on y. □

This result is not only a necessary condition for detailed balance, but also significantly simplifies the proof of the next result. First, recall the definitions of connected landscapes and paths (Definitions 7 and 8 on page 24).

Theorem 3 (Connected, canonical landscapes)**.** *Let $L = (X|_{can}, f, \mathcal{M}_{can})$ be the canonical landscape of a RNA s. Then L is connected, i. e. for all $x, y \in X|_{can}$ there is a path $P_{x \to y}$ in L connecting them.*

Proof. By Theorem 2, $\mathcal{M}_{\text{can}}$ is symmetric on $X|_{\text{can}}$. Therefore, if $P_{x \to y}$ is a valid path in L, there also is a valid path $P_{y \to x}$ in L. Also, two paths $P_{x \to z}$ and $P_{z \to y}$ can be connected to a single path $P_{x \to z} \cdot P_{z \to y} = P_{x \to y}$. It is

[1] The curly braces used to denote the base pairs are supposed to indicate that here it is not implied that $i < k$.

therefore sufficient to proof the claim that for each x there is a path $P_{x\to\emptyset}$ from x to the open chain $\emptyset$. Then, the desired path can be constructed as $P_{x\to\emptyset} \cdot P_{\emptyset\to y}$.

The claim is proved by induction over the number of base pairs $|x|$. For $|x| = 0$, $x = \emptyset$ and the path is trivial. Let $|x| > 0$. Then there is a left-outermost base pair $(i, j) \in x$ (i. e. i is minimal over all base pairs) which is by definition outside-lonely. Therefore, also $(i+1, j-1) \in x$ since x is canonical.

Now, either $(i + 1, j - 1)$ is inside-lonely ($(i + 2, j - 2) \notin x$) or it is not ($(i + 2, j - 2) \in x$). In the first case, $\{(i, j), (i + 1, j - 1)\}$ is a lonely stack of length two and can be removed by applying δ^2_{ij} to x. In the second case, $(i + 1, j - 1)$ will not grow lonely when removing (i, j) and so δ_{ij} can be applied instead. This reduces the number of base pairs in the resulting structure x' by two and one, respectively. By induction, there is a path $P_{x'\to\emptyset}$, and so $x \cdot P_{x'\to\emptyset}$ is a path from x to $\emptyset$. □

4.2.3. Implementation

The described move set has been implemented in a prototypical form as a *Perl* module. Further, it has been incorporated into the tool `barriers`, which is written in *C*. The latter implementation also includes unit tests to verify its correctness. The tests are implemented using the *Check* framework and have been integrated into the *Autotools* configuration of the project to automate the building of the testing binaries. If *Check* is not found on the system, the generation of tests is skipped.

As can be seen in Fig. 4.4 on the next page, the new move set outperforms the old one despite the more complex logic. The reason is that it significantly reduces the number of neighbors of each structure since many moves, which were valid in the old move set, are invalid in the new one.

4.3. Direct canonical paths

It is sometimes useful to find a path connecting two structures x, y in an RNA landscape, e. g. to estimate the energy barrier barrier between them or to connect them in kinetic simulation as practiced in this work. This task is nontrivial because of the enormous amount of secondary structures for usual RNAs.

A well-known approach to reduce the number of paths is to only consider the ones that run *directly* from the start structure x to the target structure

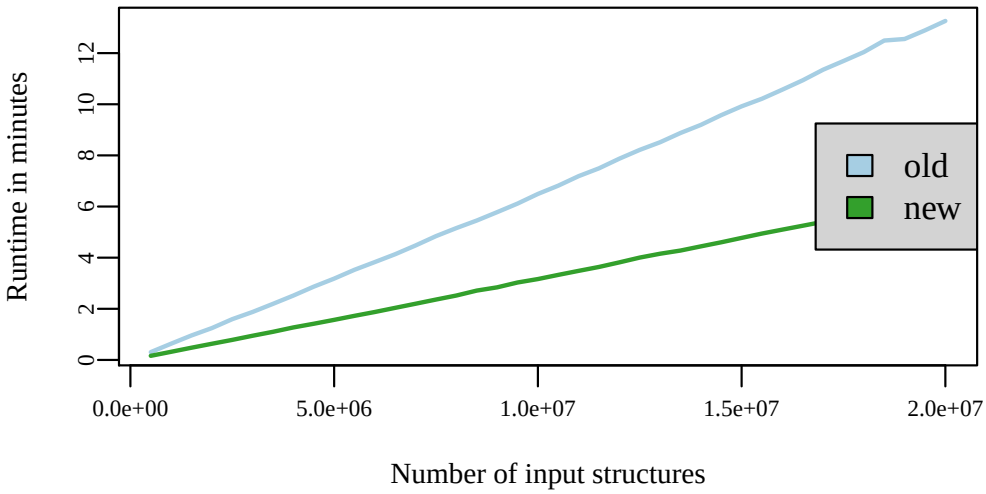


Figure 4.4.: Comparison of the runtime of the old and the new canonical move set of barriers for an increasing number of input structures.

y, i. e. in any step, the distance of the current structure to the target structure is reduced. A more formal definition is given in the following section. The notion of these *direct paths* is used, for example, in the Morgan-Higgs algorithm (Morgan and Higgs 1998) to estimate energy barriers between different states of a RNA landscape. A refined version called `findPath` developed by Flamm and Hofacker (2008) is included in the `ViennaRNA` package. It is a heuristic that, starting at x, greedily chooses the n energetically best neighbor structures on direct paths to y and adds them to a list. Then, each path from the list is processed analogously, reducing the total number of structures in the next step to n if necessary. This process is iterated until the target structure is reached. After finding an initial upper bound, n is increased and the entire process is repeated, rejecting paths containing structures with energies higher than the current optimum.

The biological justification for the approximation of barrier heights only by considering direct paths is that if structures differ only in a certain sub-component, then the re-folding between these structures will likely only change this sub-component and leave other structural elements unchanged. From a practical point of view, directs paths are useful because there are much less direct paths than arbitrary ones. Nevertheless, the number of direct paths is still high such that an exhaustive enumeration is often

infeasible even for smaller RNAs. Therefore one often resorts to heuristics as the ones described.

There is, however, a problem with this direct path approaches: they are based on the classical RNA move set and are therefore often generates non-canonical structures. This section aims at an adoption of the concept of direct paths that makes it usable for canonical RNA landscapes.

4.3.1. Definitions

Let $L_s = (X, f, \mathcal{M})$ be a RNA landscape of some RNA s with an *arbitrary* move set $\mathcal{M}$. The moves $\mu \in \mathcal{M}$ consist of the addition or removal of one or many base pairs. For some $x \in X$, let $B^+_\mu(x) = \{(i_1, j_1), \dots (i_p, j_p)\}$ and $B^-_\mu(x) = \{(k_1, l_1), \dots (k_q, l_q)\}$ be the sets of base pairs that are added or removed, respectively, when the move μ is applied to the structure x. Further, for another $y \in X$, let $B^+(x \to y) = y \setminus x$ and $B^-(x \to y) = x \setminus y$, i. e. the sets of base pairs that need to be added or removed, respectively, to move from x to y.

Definition 30 (Directed moves and neighbors)**.** *A valid move $\mu \in \mathcal{M}$ on x is called* y-directed *if $B^+_\mu(x) \subseteq y$ and $B^-_\mu(x) \cap y = \emptyset$, i. e. any added base pair is present in y, but none of the removed ones is. The neighbor structures of x that can be reached by a single y-directed move are called the* y-directed neighbors *of x.*

Using the notion of directed moves, direct paths can be defined formally, extending Definition 7 on page 24.

Definition 31 (Direct path)**.** *A path $x = x_1, x_2, \dots, x_n, x_{n+1} = y$ from x to y w. r. t. L is called* direct*, if $\forall i \in \{1, \dots, n\}$, x_{i+1} is a y-directed neighbor of x_i.*

The definition of direct paths is intuitively related with a notion of distance between the structures x and y. This can be formalized as follows.

Definition 32 (Base pair distance)**.** *Let $x, y \in X$ be secondary structures. The* base pair distance *$d_{BP}(x, y)$ of x and y is defined as the number of base pairs present in either x or y, but not in both. More formally,*

$$d_{BP}(x, y) = \Big|(x \setminus y) \cup (y \setminus x)\Big| .$$

Walking along a direct path from x to y, the base pair distance to x increases, while the distance to y decreases by at least one with every step[2]. Therefore, the length of a direct path cannot exceed the base pair distance $d_{\mathrm{BP}}(x, y)$.

In a canonical landscape, one is usually only interested in paths that run only through canonical structure. Therefore, the concept of canonical paths is introduced.

Definition 33 (Canonical path)**.** *A path between two canonical structures $x, y \in X|_{can}$ w. r. t. $\mathcal{M}$ is called* canonical*, if all structures on the path are canonical.*

When doing computations on canonical RNA landscapes, *direct canonical paths* are the natural extension of direct paths to canonical landscapes. Note that paths are, in general, only valid for a suited move set.

4.3.2. Existence of direct canonical paths

Now that direct canonical paths are formally defined, the question arises whether they actually exist between arbitrary canonical structures. This section gives a positive answer and proves the result. Let $L_s = (X|_{\mathrm{can}}, f, \mathcal{M}_{\mathrm{can}})$ be the canonical landscape of a RNA s. The following lemma is required for the proof.

Lemma 4. *Let $x, y \in X|_{can}$ and $(i, j) \in B^-(x \to y)$. Let (i, j) not be directly enclosed[3] by an outside-lonely base pair. Then a valid, y-directed move from the canonical move set can be applied to the part of x enclosed by (i, j) or to (i, j) itself.*

Note that the preconditions of the lemma are chosen such that by deleting (i, j) no lonely pairs can arise outside of (i, j).

Proof. By induction over the number n of bases enclosed by (i, j). Using the abbreviated notation $B^+ = B^+(x \to y)$ and $B^- = B^-(x \to y)$, the following cases can be distinguished.

Case 1. $n = \varepsilon_{\mathrm{loop}} = 3$, the minimal number of bases in a hairpin loop. There are no further base pairs inside (i, j) and (i, j) is not enclosed by a base pair that could grow lonely when removing

[2]For the move distance w. r. t. $\mathcal{M}$, this is in general not the case if different moves change the base pair distance by different amounts.

[3]Recall Definition 16 on page 29.

(i, j) so that removing (i, j) is a valid and, since $(i, j) \in B^-$, y-directed move.

Case 2. $n > \varepsilon_{\text{loop}}$. If $(i+1, j-1) \notin x$, the removal of (i, j) is valid since no lonely pairs arise. Else, $(i+1, j-1) \in x$. Now, two cases can be distinguished:

i. $(i+2, j-2) \in x$. In this case, (i, j) can be removed from x since $(i+1, j-1)$ will not grow lonely.

ii. $(i+2, j-2) \notin x$. It follows that $(i+1, j-1)$ would grow lonely when removing (i, j), rendering this move invalid. Further distinction is needed:

a. $(i+1, j-1) \in B^-$, so, depending on whether the surrounding base pair $(i-1, j+1)$ is present, either removing (i, j) or the lonely stack (i, j), $(i+1, j-1)$ is valid and y-directed.

b. $(i+1, j-1) \notin B^-$. Figure 4.5 on the facing page gives visual overview of this case. Since y is canonical and $(i, j) \notin y$, it follows that $(i+2, j-2) \in B^+$, otherwise $(i+1, j-1)$ would be lonely in y. If both $i+2$ and $j-2$ are unpaired positions in x, adding $(i+2, j-2)$ is a valid and y-directed move.

Else, at least one of the bases $i+2$ or $j-2$ is paired. If $(i+2, k) \in x$, where necessarily $i+2 < k < j-2$, then $(i+2, k)$ is found in B^- since it is conflicting with the required base pair $(i+2, j-2)$. Because of the base pair $(i+1, j-1) \in x$, base pair $(i+2, k)$ is not directly enclosed by another base pair: it fulfills the preconditions of this lemma and by induction a valid and y-directed move can be applied to the enclosed part of the sequence, which is also enclosed by (i, j). The case $(l, j-2) \in x$, $i+2 < l < j-2$, is analog.

□

Using this lemma, the existence of direct canonical paths can be proved.

Theorem 4 (Existence of canonical direct paths)**.** *Let $x, y \in X|_{can}$. Then there is a direct canonical path from x to y w. r. t. the canonical move set.*

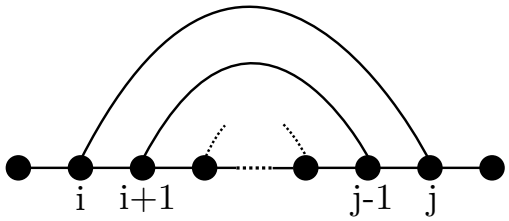


Figure 4.5.: Last case of the proof of Lemma 4 on page 53.

Proof. By induction over the base pair distance $n = d_{BP}(x, y)$ of the structures. If $n = 0$, then $x = y$. For $n > 0$, it suffices to show that a valid, y-directed move can be applied to x since the resulting structure will have a base pair distance of at most $n - 1$ w. r. t. y. The shorthand notation from the previous lemma will be used here, too: let $B^+ = B^+(x \to y)$ and $B^- = B^-(x \to y)$ such that $n = |B^+| + |B^-|$. The following cases can be distinguished:

Case 1. $B^- = \emptyset$, i. e. $x \subseteq y$. Since $n > 0$, there is a left-outermost base pair $(i, j) = \min_i\{(i, j) \in B^+\}$ that needs to be added to x. The bases i and j must be unpaired in x since $B^- = \emptyset$. Further distinction is needed:

i. (i, j) is *not* lonely in $x \cup (i, j)$. This insertion cannot violate the "no pseudo-knot" condition: assume there is a conflicting base pair $(k, l) \in x$. Then, because $B^- = \emptyset$, (k, l) is also present in y, so y contains a pseudo-knot. This is contradiction, since y is a secondary structure, so there cannot be a base pair (k, l) that conflicts with the insertion of (i, j), and so the insertion is valid and y-directed.

ii. (i, j) *is* lonely in $x \cup (i, j)$. Since y is canonical, another base pair in B^+ needs to be added adjacent to (i, j) such that (i, j) is no longer lonely. Because of the minimality of i, this has to be on the inside, so $(i + 1, j - 1) \in B^+$. If $(i + 1, j - 1)$ is inside-lonely, inserting the lonely stack $(i, j), (i + 1, j - 1)$ is a valid move, if not, one can add only $(i + 1, j - 1)$. Pseudo-knots cannot arise due to the same argument as in the previous case.

Case 2. $\exists(i, j) = \min_i\{(i, j) \in B^-\}$. If removing (i, j) does not yield a lonely pair, remove it. Else, there is a lonely-growing base pair.

i. $(i - 1, j + 1)$ is not in x or not outside-lonely. By Lemma 4

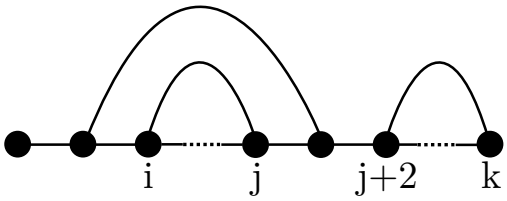


Figure 4.6.: Last case of the proof of Theorem 4 on page 54.

on page 53, there is a valid, y-directed move that can be applied to the substructure enclosed by (i, j).

ii. $(i-1, j+1) \in x$ is outside-lonely. Since (i, j) needs to be removed, $(i-2, j+2) \in B^+$ since otherwise $(i-1, j+1)$ would be lonely in y. The nucleobase at position $i-2$ is unpaired because of the minimality of i. Nucleobase $j+2$ can be paired or unpaired in x:

   a. If $j+2$ is unpaired, $(i-2, j+2)$ can be inserted and is not lonely in $x' = x \cup (i-2, j+2)$ since $(i-1, j+1) \in x'$.

   b. If $j+2$ is paired, it must pair with a $k > j+2$ (cf. Fig. 4.6). The reason is that the pair $(j+2, k)$ conflicts with the necessary insertion of $(i-2, j+2)$, so $(j+2, k) \in B^-$, and i is minimal in this set. Further, $(j+2, k)$ is not enclosed by a base pair since $j+1$ pairs to the left. Therefore, the preconditions of Lemma 4 on page 53 are fulfilled and a y-directed move can be applied to the substructure of x enclosed by $(j+2, k)$.

□

4.3.3. Re-implementation of `findPath`

According to Lemma 4 on page 53, direct canonical paths exist between any two canonical structures. Therefore, direct paths in canonical RNA landscapes can be utilized for the same purposes as in general RNA landscapes. Since it is useful for connecting basins to each other, the `findPath` heuristic has been re-implemented as a *Perl* function, including support for canonical paths and shift moves that can be toggled on or off independently.

The implementation uses a *max* queue to keep track of the n energetically best structures. Additionally, a hash of all structures currently in the queue

is used to prevent that the same structure, reached by a different path, is inserted multiple times. In such cases, only the best structure is kept in the queue.

The performance of the *Perl* implementation is, of course, much lower than the *C* variant. Given the same input, it requires about ten times as long. When performing the direct path search on many basins this becomes a bottleneck in the kinetics pipeline implemented in this work. Therefore, a re-implementation in *C* or another fast compiled language would be a beneficial future task.

Chapter 5.

Partial Exploration of Energy Landscapes

In this chapter, the partial exploration of energy landscapes is discussed. More specifically, what follows is a generic description of an approach to, starting at a specific state of the landscape, enumerate neighbor states of this element and find one or multiple paths to the low energy states of the landscape. This method is then applied to RNA energy landscapes and has also been implemented as a ready-to-use *Perl* script.

5.1. Motivation

To analyze the folding kinetics of a RNA sequence s, it is necessary to construct its RNA energy landscape L_s (cf. Definition 19 on page 30). To keep the size of the structure space feasible, the full space X can be pruned to an energy band of width ΔE as described in Section 3.4 on page 34, yielding a microstate set $X|_{\Delta E}$. This, however, might have the consequence of pruning away interesting structures as e. g. the open chain, which is usually assumed to be the start structure of a kinetic simulation. For example, the synthetic riboswitch RS4 designed by Wachsmuth et al. (2013), which has a minimum free energy of $-26.70\,\mathrm{kcal\,mol^{-1}}$ at 37 K, yields

$$\left|X|_{25\,\mathrm{kcal\,mol^{-1}}}\right| = 2{,}034{,}217{,}895.$$

This is by far infeasible and does not even include the open chain.

The example above raises the question of how to include additional structures to a pruned energy landscape. Note that it is insufficient to simply add additional candidate structures to $X|_{\Delta E}$ since, in general, they will not have any neighbors and, thus, be isolated. In a kinetic simulation, the isolated state would never be visited or left. One therefore needs to connect the candidate structure to the enumerated part of the landscape.

There are different solutions to this problem. An easy way would be to use a direct path heuristic as implemented in `findPath` (Flamm et al. 2000), which, however, has several drawbacks. Direct paths give only a rough estimate of the actual barrier height, introducing errors into the rate computation. This effect aggravates in the context of a gradient-based coarse graining (cf. Definition 13 on page 27) since a direct path might introduce several new basins consisting of very few or even only a single structure. To attenuate the error, several direct paths to different target structures need to be computed, reducing the performance of this approach. It is also necessary to repeat the entire process for each additional structure that is supposed to be connected to the landscape.

This chapters aims at introducing a general algorithm for the partial exploration of energy landscapes based on the notion of basins (Definition 10 on page 25) and gradient walks (Definition 9 on page 25), which tries to avoid the problems mentioned above.

5.2. Concepts and definitions

Throughout this section, let $L = (X, f, \mathcal{M})$ be an energy landscape. Further, let $\Delta E > 0$ an exploration threshold and $x \in X$ a candidate structure that is not found in the ΔE-pruned state set $X|_{\Delta E}$ (cf. Definition 21 on page 34).

The heuristic is based on the goal to enumerate structures in the basin of x, searching for structures leading to other basins and following them until the explored part $X|_{\Delta E}$ of the landscape is reached. This is made more precise with the following definitions.

Definition 34 (Contact state)**.** *Let $x \in X$ be a local minimum of L and $B(x)$ its associated basin. A state $y \in X$ is called a* contact state *w. r. t. $B(x)$ if $B(y) \neq B(x)$ and there is a structure $x' \in B(x)$ such that y and x' are neighbored. Let $\mathcal{C}(B(x))$ denote the set of all contact states of basin $B(x)$.*

The idea now is to explore *a part of* the basin of x and repeat this process for the basins of all encountered contact states that lead to basins of lower energy. However, since states with a high energy are less likely, it is sensible to only flood each basin up to a certain energy threshold η. Therefore, the following definition, related to concepts in Flamm et al. (2002) is useful:

Definition 35 (η-basin)**.** *Let $x \in X$ be a local minimum of L and E_x be its energy. Then the* η-basin $B_\eta(x)$ *of x for some energy value $\eta > 0$ is the set of structures from $B(x)$ with an energy not more than $E_x + \eta$, i. e.*

$$B_\eta(x) = \{y \in B(x) \,|\, E_y \leq E_x + \eta\}\,.$$

Note that B_η is always connected: if $y \in B(x)$, then there is a gradient walk $P = y, z_1, \ldots, z_k, x$ leading from y to x. Of course, $z_1, \ldots, z_k \in B(x)$ since P is also a gradient walk from each these states. Since P is a gradient walk, $E_y \leq E_{z_1} \leq \cdots \leq E_{z_k} \leq E_x$. Therefore, if $E_y \leq E_x + \eta$, then $E_{z_i} \leq E_x + \eta$ for all $1 \leq i \leq k$, so all z_i are also found in the η-basin and P connects y to x. The notion of contact states naturally extends to η-basins.

To improve the heuristic's ability to overcome basins that are surrounded only by basins with higher minimal energies, the notion of an *extended basin* is introduced. It also includes states from neighboring basins up to the same energy value η.

Definition 36 (extended η-basin)**.** *Let $x \in X$ be a local minimum of L and E_x be its energy. Then the* extended η-basin *of x for some energy value $\eta > 0$ is the set*

$$B^*_\eta(x) = \bigcup_{\substack{y \in \mathcal{C}(B_\eta(x)) \cup x \\ E_{\gamma(y)} \geq E_x}} B_{\eta - (E_y - E_x)}(y),$$

where E_y and $E_{\gamma(y)}$ are the energy of state y and the energy of the target state of the gradient walk starting in y, respectively.

5.3. Description of the algorithm

Building on Definition 36, the heuristic can finally be described precisely. The extended η-basin of the initial structure x is flooded for increasing values of η. For all encountered contact states, the procedure is repeated until the explored part of the landscape $X|_{\Delta E}$ is reached. This way, a significant part of the structures in the encountered basins are enumerated while excluding structures the lie in a different part of the landscape.

The algorithm is given as pseudo-code listing (cf. Algorithm 1 on page 64). It begins by performing a gradient walk from x, yielding the local minimum $x' = \gamma(x)$. Next, the extended E_{inc}-basin $B(x')$ is flooded up to an energy level of $E_{x'} + E_{\text{inc}}$, storing all encountered contact states in a list. If less

than $n_{\min}$ contact states y' with $\gamma(y') < E_{x'}$ where found, increase E_{inc} and repeat the process until either enough contact states where found or some $E_{\max}$ is reached and the heuristic bails out for this basin. For each contact state, continue if $\gamma(y') \leq E_{\text{thresh}}$, where $E_{\text{thresh}} = \Delta E + E_{\text{mfe}}$, else add $\gamma(y')$ to a queue. Until this queue is empty, repeat the described process for each state in it. At all times, keep a hash of visited states to prevent processing parts of the landscape more than once. The algorithm returns a list of each state it has encountered, which can then be added to the pruned state set $X|_{\Delta E}$. Figure 5.1 on the next page shows an example run of the algorithm.

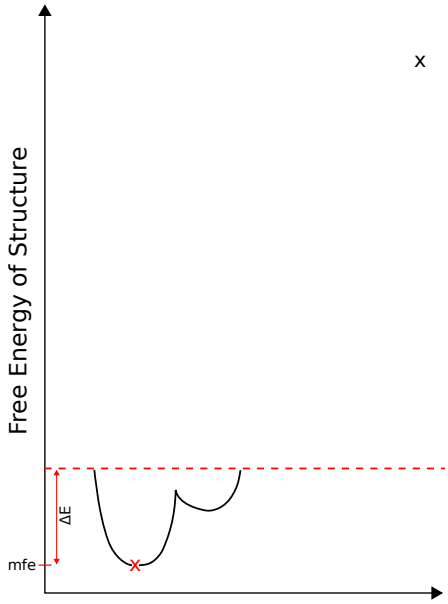
To flood an extended basin $B^*_\eta(x)$ to an energy level of η, initialize a queue Q with the local minimum $x' = \gamma(x)$. During the flooding, mark any structures that are pushed into the queue and prevent inserting them a second time. While the queue is not empty, pop an element x from Q and compute all neighbors $N(x)$ with an energy of $E_{x'} + \eta$ or less. For each such neighbor y, check whether it lies in the same extended basin as x, and if so, push it to the queue. Else, add y to the returned list of contact states. A pseudo-code representation is given in Algorithm 2 on page 65.

The described method can easily be extended for more than a single candidate state. Therefore, they simply need to be added to the queue. Since the algorithm remembers any visited state, multiple similar states will be processed quickly.

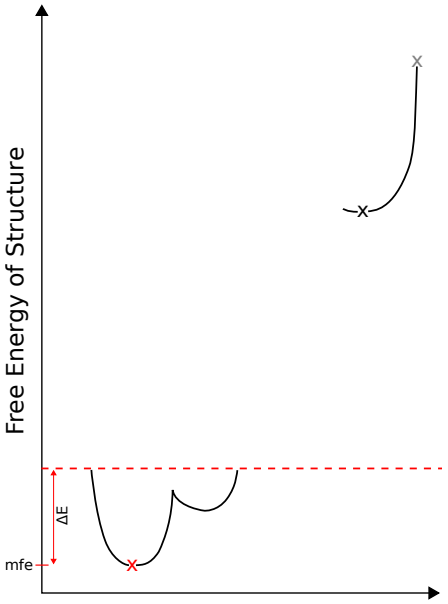
5.4. Implementation and application

The described algorithm has been implemented prototypically for RNA secondary structures. It was implemented as a *Perl* module, relying on the `ViennaRNA` *Perl* bindings to compute the energy of secondary structures. The program expects a RNA sequence and one or more secondary structures in dot-bracket notation as input. If no structures are provided, the open chain is used. Of course, the parameters from the algorithmic description can be specified by the user via command line arguments. This allows for an easy adaption of the program for different energy landscapes in the case that a run should take too long or that no path down to structures from the pruned state set are found. The following options are available:

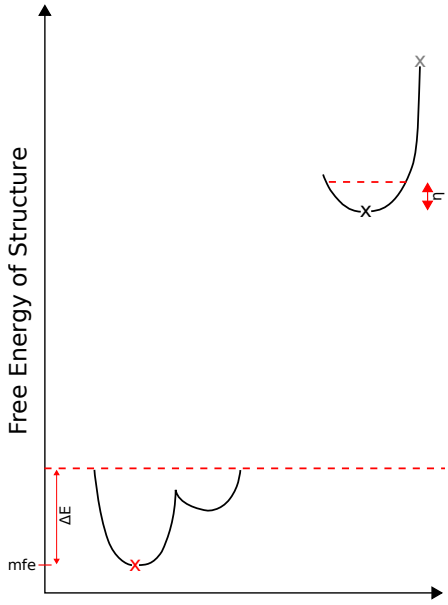
`-T`: folding temperature in °C, which is required to compute the energy of secondary structures



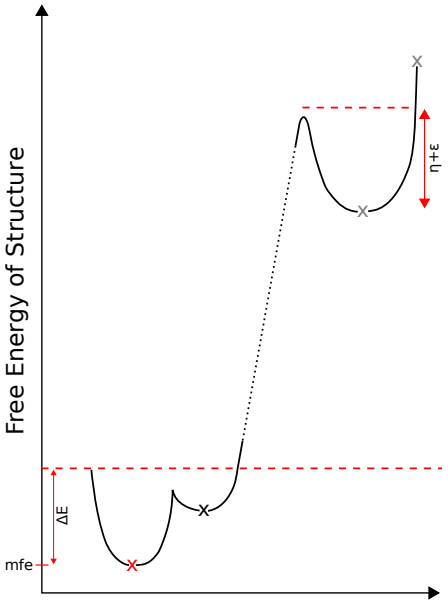
(a) Explored part of the landscape and candidate structure x.



(b) A gradient walk starting in x is performed.



(c) The basin is flooded to a level of η.



(d) The flooding level is increased until contact states are found.

Figure 5.1.: Example run of one iteration of Algorithm 1 on the next page.

Algorithm 1: Partially explore an energy landscape.

Input : structures of interest $x_1, x_2, \ldots$; absolute exploration threshold E_{thresh}, flooding energy increment value E_{inc}, maximum per-basin flooding energy $E_{\max}$, minimum number of contact states $n_{\min}$

Output : list of all visited structures

Data: integer n, queue Q, list L, hash of queued structures H

$Q \leftarrow (x_1, x_2, \ldots)$

while $Q \neq \emptyset$ **do**

  $x \leftarrow \texttt{pop}(Q)$

  $x' \leftarrow \gamma(x)$

  if $E_{x'} < E_{thresh}$ *or* $H[x'] =$ *"queued"* **then**

    continue

  $n \leftarrow 0$

  repeat

    $n \leftarrow n + 1$

    $(L, H_{\text{basin}}) \leftarrow \texttt{floodBasin}(x', E_{x'} + n \cdot E_{\text{inc}}, n_{\min}, E_{\max})$

  until $|L| \geq n_{\min}$ *or* $n \cdot E_{inc} \geq E_{\max}$

  foreach $z \in \{z' \in L \mid H[z'] \neq$ *"queued"*$\}$ **do**

    $\texttt{push}(z \rightarrow Q)$

  merge H_{basin} into H

return *keys of* H

Function 2: floodBasin: Flood extended basin of a given structure up to a specified energy level.

Input : structure x, absolute exploration threshold E_{thresh}, minimum number $n_{\min}$ of contact states to find, maximum flooding level $E_{\max}$

Output : list L of contact states

Data: queue Q, local hash H to mark queued structures

$x' \leftarrow \gamma(x)$

$\texttt{push}(x' \rightarrow Q)$

$H[x'] \leftarrow$ "queued"

while $Q \neq$ **do**

  $y \leftarrow \texttt{pop}(Q)$

  foreach *neighbor* z *of* y *with* $E_z \leq E_{thresh}$ **do**

    if $H[z] \neq$ *"queued"* **then**

      $H[z] \leftarrow$ "queued"

      if $z \in B(x')$ *or* $E_{\gamma(z)} \geq E_{x'}$ **then**

        $\texttt{push}(z \rightarrow Q)$

      else

        $\texttt{push}(z \rightarrow L)$

return L, H

-E: exploration threshold in kcal mol^{-1} above the minimum free energy in which the landscape is already enumerated ($E_{\text{thresh}} - E_{\text{mfe}}$)

-m: maximum flooding level for a single basin (in kcal mol^{-1}) ($E_{\max}$)

-i: flood energy increment value in kcal mol^{-1} (E_{inc})

-n: minimum number of contact states to search for ($n_{\min}$)

-l: allow lonely pairs, i. e. do not exclude non-canonical structures (cf. Section 4.2 on page 42)

-o: output file name in which all encountered structures and their respective energies are stored

-v: enable verbose output, printing out progress for each basin and encountered contact states

-q: be quiet, do not print found structures that lie below exploration threshold

-h: show the program help

For the synthetic riboswitches designed by Wachsmuth et al. (2013), each of 70 to 80 nucleotides long, flooding from the open chain until an exploration threshold of 15 kcal mol^{-1} above the minimum free energy was reached took approximately 20 minutes on an Intel Core i7-4770 CPU (4×3.4 GHz). Thereby, each basin was flooded up to 5 kcal mol^{-1}, searching for at least five contact states. The folding temperature was set to 37 °C.

5.5. Discussion

Obviously, the runtime strongly depends on the length and the composition of the input sequence as well as on the specified structures of interest and the passed parameters. In the worst case, the entire landscape has to be flooded and so the runtime is exponential in the sequence length. On the other hand, the algorithm will never enumerate structures with an energy higher than $E_{\text{in}} + E_{\max}$, where $E_{\text{in}} = \max\{E_{x_1}, E_{x_2}, \ldots\}$ is the maximum energy of all input structures. An upper bound for the value of $E_{\max}$ required to be able reach the pruned part of the landscape may be estimated with a direct path search (Flamm et al. 2000).

Until now, the implementation is not parallelized. Using a thread-safe implementation of the hash data structure, this should be possible without difficulties, yielding significant performance improvements. For this work, the current performance was sufficient since the computation of the additional structures needed to be performed only once for each riboswitch at any given temperature. Therefore, the improvement of this valuable tool is an unresolved task that remains for future work.

Chapter 6.

Folding Kinetics of Riboswitches

As explained in Section 2.4 on page 18, the folding kinetics of riboswitches are important for their analysis and design. This section presents a tractable modeling approach that allows the prediction of concentrations of different conformations for a given riboswitch during the course of time.

To obtain an approximate but tractable model of RNA folding, Wolfinger et al. (2004) present a coarse-graining approach as described in Definition 13 on page 27. Following e. g. Flamm et al. (2000), they analyze the folding process on the energy landscape of conformations, i. e. secondary structures, R_i of a RNA R. Conformation change is modeled by elementary moves (base pair insertion or deletion, Definition 18 on page 30) endowed with reaction rates that follow the Arrhenius rule (cf. Section 3.5.2 on page 36) and thus depend on the energy barrier between the source and target conformations. In the approximation of RNA secondary structures, activation energies for opening/closing of single base pairs are approximately constant. The energy barrier thus effectively depends only on the energy difference between source and target (Wolfinger et al. 2004). This defines a Markov *process* on the state space of all secondary structures, which is too large to make it possible to analyze it by diagonalizing the corresponding rate matrix, i. e. by integrating the master equation as described in section Section 6.7. To effectively reduce the state space, Wolfinger et al. (ibid.) combine states into basins that consist of all conformations that are connected to the same local minimum by their gradient walk on the energy landscape. Since gradient walks connect states to their lowest energy neighbors, they correspond to the fastest folding paths from a state into a local minimum. This provides the rationale for approximating the full process by the *macro*process on gradient basin *macro*states, which are assumed to be equilibrated. Consequently, the rates between the macrostates are canonically derived as weighted sums of *micro*rates of the original process. The macroprocess is finally solved by diagonalization. The approach described here re-uses ideas of this coarse-graining, which also allows re-using several tools for single-molecule RNA kinetics (`RNAsubopt`

(Lorenz et al. 2011), `barriers` (Flamm et al. 2002), `treekin` (Wolfinger et al. 2004)).

6.1. RNA ligand interaction model

The reaction system of the RNA R (a RNA molecule, given by its sequence of nucleotides $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$) with the ligand L is described at the level of RNA and binding complex conformations, such that the kinetics of association, dissociation, and conformation changes can studied. For simplicity, it is assumed that there is only a single ligand conformation (also denoted L). Like the RNA, the complex of RNA and ligand adopts various conformations LR_i; note however that only a subset of the RNA conformations binds the ligand. Thus, the system of consideration consists of the reactions

$$\mathrm{R_i} \longrightarrow \mathrm{R_j} \qquad (6.1a) \qquad\qquad \mathrm{LR_i} \longrightarrow \mathrm{L} + \mathrm{R_i} \qquad (6.1c)$$

$$\mathrm{L} + \mathrm{R_i} \longrightarrow \mathrm{LR_i} \qquad (6.1b) \qquad\qquad \mathrm{LR_i} \longrightarrow \mathrm{LR_j}, \qquad (6.1d)$$

for all $i, j \in \{1, \ldots, N\}$, $i \neq j$. According to the rate laws for elementary reactions (cf. 3.5.2), the rates of each of these reactions depend on specific rate constants and the concentrations of the molecules. The Reactions 6.1a and 6.1d only have non-zero rate constants, if the RNA conformations R_i and R_j are related by an elementary move such as the insertion or deletion of a base pair. Moreover, L and R_i can interact only for the subset of the R_i that form an appropriate binding pocket; otherwise, the complex LR_i is deemed unstable and thus excluded from the model.

Since RNA conformations correspond to RNA secondary structures, the energies of monomer states can be calculated from the Turner energy model (Turner and Mathews 2009). For dimer states, the aptamer-ligand-specific binding energy is added. For the exemplary studied riboswitch RS3, this energy can be derived from the empirical dissociation constant (Wachsmuth et al. 2013) as described in Section 3.6 on page 38.

Finally, the rate constants are derived as Metropolis rates with appropriate pre-exponential factors that can be estimated from empirical rates. Note that the rates of base pair opening k^- and closing k^+ are directly related by the energy change ΔG due to the closing. Concretely, $k^-/k^+ = \exp(\Delta G/RT)$ for $\Delta G < 0$. Experimental values are available for the zippering rate, which corresponds to the rate of closing the last hairpin in a helix. A careful analysis in Kuznetsov and Ansari (2012) yields a value in the range $4.7 \times 10^7\,\mathrm{s}^{-1}$ to $1 \times 10^9\,\mathrm{s}^{-1}$ roughly consistent with

earlier estimates (Cocco et al. 2003; Pörschke 1974; Zhang and Chen 2002). In principle, a kinetic constant can be derived for the closing of first base pair in a loop from a worm-like chain model Kuznetsov and Ansari (2012) and Toan et al. (2008); following earlier work on RNA kinetic models (Wolfinger et al. 2004), a single kinetic parameter k^+ for all base pairs is used here.

An empirical rate of one specific theophylline aptamer association was reported as $600\,\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ (Latham et al. 2009), which may serve as rough estimate for comparable systems. Note that Latham et al. (ibid.) measured the macroscopic *apparent rate* that depends on the rate of *dimerization*, i. e. the formation of a RNA–ligand dimer complex, as well as the rate of refolding into structures with theophylline binding pocket.

While the Reactions 6.1*a*, 6.1*c*, and 6.1*d* are of first order, the second order association in Reaction 6.1*b* introduces non-linearity into the system. Assuming *ligand excess*, which is a very plausible assumption for small molecular ligands, however, it is possible to devise a *pseudo-first order approximation* of the system.

Even if the reaction equations above appropriately model the RNA–ligand interaction, this system is still computationally intractable for typical riboswitch sizes. As a remedy, a coarse-grained process based on separate gradient basins for the monomer and dimer states is constructed. The monomer states with suitable binding pocket are connected to dimer states, Fig. 6.1 on the next page. Importantly, there is no direct mapping from monomer macrostates to dimer macrostates of the coarse-grained system because conformations without binding pockets are absent from the "dimer world". The upper basin in the "monomer world" of Fig. 6.1 on the following page is subdivided into two basins in the dimer world; conversely, the middle and lower monomer basins correspond to a single basin of the dimer world.

6.2. Contributions

Section 6.3 on page 73 discusses the general macroprocess of RNA ligand interaction based on gradient basin macrostates and the derivation the corresponding rate constants. This original description of the specific macrostate system is a fundamental prerequisite for RNA ligand interaction kinetics based on gradient basin coarse graining. Subsequently, this system is discussed under the assumption of excessive ligand concentrations, which is valid for a wide spectrum of biological systems. On this basis, this chapter

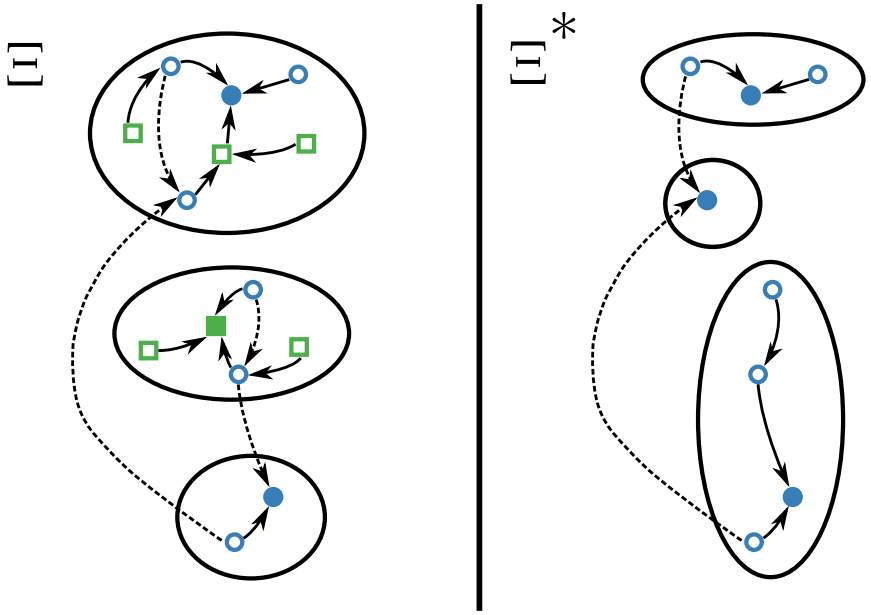


Figure 6.1.: Correspondence between the energy landscapes of the monomers (*left*) and the dimers (*right*). The dimer landscape is obtained from the monomer landscape by only retaining the structures possessing the binding pocket (*blue circles*) while removing the other ones (*green squares*). As the removed structures might lie on a gradient walk (*solid arrow*), rendering that path invalid in the dimer landscape, formerly suboptimal moves (*dashed arrows*) become gradient walks and new local minima (*filled squares and circles*) may arise. These effects might alter the mapping of the structures to their gradient basin.

devises the first analytical approach for RNA ligand interaction kinetics enabling the computation of time-dependent macrostate probabilities based on solving the master equation of the interaction process. Finally, the interaction kinetics of the artificially designed theophylline riboswitch RS3 (Wachsmuth et al. 2013) at different concentrations is discussed and the effect of co-transcriptional interaction in the model is studied.

6.3. Macrostate kinetics of RNA–ligand interaction

6.3.1. Preliminaries and basic notation

Consider the fixed interaction system of the *RNA* R and the *ligand* L. Let X denote the set of all *monomer microstates*, $X = \{R_i \mid i = 1, \ldots, N\}$; in this setting the R_i are the secondary structures of a given RNA sequence. The subset $X^+ \subseteq X$ comprises the conformations that can bind the ligand. Here X^+ contains all states with a specific binding pocket. Furthermore, define X^* as the set of *dimer microstates* LR_i, $X^* = \{LR_i \mid R_i \in X^+\} \subseteq \{LR_i \mid i = 1, \ldots, N\}$.

A monomer microstate $LR_i \in X^*$ has the energy $E(R_i) + \theta_L$, where $\theta_L < 0$ denotes the *binding energy* of R and L. The *inverse temperature* is $\mathfrak{b} = \frac{1}{RT}$, where T is the *absolute temperature* and R is the *universal gas constant*. For a set $S \subseteq X \cup X^*$ of microstates, $\mathrm{Z}[\,S\,]$ denotes the partition function of S (cf. Definition 20 on page 33) The *probability of a microstate* x in S is denoted by $\Pr[\,x \mid S\,]$ (Ibid.). Let $x, y \in X \cup X^*$ be microstates. The *microrate constant* from x to y is denoted $k(x \to y)$ (or $k(y \leftarrow x)$.)

On microstates, define the symmetric *neighborhood relation* $x \mathcal{N} y$ w. r. t. a given symmetric move set. Only neighbors have non-zero transition rates between microrates. All microrate constants are defined by the Metropolis rule, i. e. for $x, y \in X \cup X^*$, $x \neq y$

$$k(x \to y) = c(x \to y) \begin{cases} \exp(-\mathfrak{b}[\max\{E(x), E(y)\} - E(x)]) & \text{if } x \mathcal{N} y, \\ 0 & \text{else,} \end{cases} \tag{6.2}$$

where $c(x \to y)$ denotes the reaction-specific *pre-exponential factor*. Here it is assumed that this factor depends only on the type of reaction and the factors for conformation change in monomers and dimers are equal. Thus, one can distinguish the factors c_{a} for association, c_{d} for dissociation, and c_{R} for conformation changes of the RNA secondary structure. As is shown shown later in Eq. (6.11) on page 85, $c_{\mathrm{a}} = c_{\mathrm{d}}$ due to *detailed balance* (cf. Section 3.5.1 on page 35).

Denote the *power set* of a set S by $\mathcal{P}(S)$ (cf. Definition 11 on page 26). A *monomer (dimer) macrostate* is a set of monomer (dimer) microstates, i. e. an element of $\mathcal{P}(X)$ ($\mathcal{P}(X^*)$), respectively. Denote the *(macro)rate* constant from macrostate α to β by $r(\alpha \to \beta)$ (alternatively, $r(\beta \leftarrow \alpha)$). Macrorate constants are defined by microrate constants and state proba-

bilities as

$$r(\alpha \to \beta) = \sum_{x \in \alpha, y \in \beta} \Pr[\, x \mid \alpha \,] k(x \to y).$$

Note that the term macrostates is used freely to denote general sets of microstates. Only when specific partitions of the microstates into macrostates are introduced in Section 6.4 on page 76, it makes sense to distinguish *represented* macrostates of the specific coarse-grained system from other sets of microstates.

6.3.2. Rate constants between dimer states

For a microstate $x \in X^+$, let Lx denote its corresponding dimer microstate (after binding to L), i. e. for $x = R_i$, $Lx = LR_i$. This notation is raised to sets of microstates by defining $L\alpha = \{Lx \mid x \in \alpha\}$. Lemma 5 below asserts that the rate constants between dimer microstates and macrostates can be computed exactly like rate constants of monomer states.

Lemma 5. *For $x, y \in X^+$, $k(Lx \to Ly) = k(x \to y)$. Furthermore, for all $\alpha \in \mathcal{P}(X^+)$, $\Pr[\, Lx \mid L\alpha \,] = \Pr[\, x \mid \alpha \,]$. Finally, $r(L\alpha \to L\beta) = r(\alpha \to \beta)$, for all macrostates $\alpha, \beta \in \mathcal{P}(X^+)$.*

Proof. The individual claims follow easily from the definitions:

$$\begin{aligned}
k(Lx \to Ly) &= c(Lx \to Ly) \exp(-\mathfrak{b}[\max\{E(Lx), E(Ly)\} - E(Lx)]) \\
&= c(x \to y) \exp(-\mathfrak{b}[\max\{E(x) + \theta_L, E(y) + \theta_L\} - (E(x) + \theta_L)]) \\
&= c(x \to y) \exp(-\mathfrak{b}[\max\{E(x), E(y)\} + \theta_L - E(x) - \theta_L]) \\
&= \exp(-\mathfrak{b}[\max\{E(x), E(y)\} - E(x)]) \\
&= k(x \to y).
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\Pr[\, Lx \mid L\alpha \,] &= \frac{\exp(-\mathfrak{b} E(Lx))}{\mathrm{Z}[\, L\alpha \,]} \\
&= \frac{\exp(-\mathfrak{b}[E(x) + \theta_L])}{\sum_{x \in \alpha} \exp(-\mathfrak{b}[E(x) + \theta_L])} \\
&= \frac{\exp(-\mathfrak{b}\theta_L) \exp(-\mathfrak{b} E(x))}{\exp(-\mathfrak{b}\theta_L) \sum_{x \in \alpha} \exp(-\mathfrak{b} E(x))} \\
&= \Pr[\, L \mid \alpha \,].
\end{aligned}$$

Finally,

$$\begin{aligned} r(L\alpha \to L\beta) &= \sum_{\substack{Lx \in L\alpha \\ Ly \in L\beta}} \Pr[\, Lx \mid L\alpha \,] k(Lx \to Ly) \\ &= \sum_{\substack{x \in \alpha \\ y \in \beta}} \Pr[\, x \mid \alpha \,] k(x \to y) \\ &= r(\alpha \to \beta), \end{aligned}$$

where the sum runs over the $Lx \in L\alpha$ and $Ly \in L\beta$. □

The microrate constant from monomer to dimer states is constant, whereas the back rate depends on the binding energy θ_L.

Lemma 6 (Association and dissociation microrate constants)**.** *For $x \in X^+$, the rate of association is $k(x \to Lx) = c_a$ and the dissociation rate is $k(Lx \to x) = c_d \exp(\mathfrak{b}\theta_L)$. All other rates between monomer and dimer microstates are* 0*.*

Proof. By METROPOLIS rule, for $x \in X^+$, with $\eta = \max\{E(x), E(Lx)\} - E(x)$,

$$\begin{aligned} k(x \to Lx) &= c_{\mathrm{a}} \exp(-\mathfrak{b}\eta) \\ &= c_{\mathrm{a}} \exp(-\mathfrak{b}[E(x) - E(x)]) = c_{\mathrm{a}}, \end{aligned}$$

since $E(Lx) = E(x) + \theta_L \leq E(x)$. Analogously, for the inverse microrate,

$$\begin{aligned} k(Lx \to x) &= c_{\mathrm{d}} \exp(-\mathfrak{b}\eta) \\ &= c_{\mathrm{d}} \exp(-\mathfrak{b}[E(x) - E(Lx)]) \\ &= c_{\mathrm{d}} \exp(\mathfrak{b}\theta_L). \end{aligned}$$

□

The association (dissociation) microrates due to Lemma 6 induce corresponding macrorates, which additionally depend on the probability of the associable (dissociable) microstates in the source macrostate, respectively (Lemma 7.)

Lemma 7 (Association and dissociation macrorate constants)**.** *For $\alpha \in \mathcal{P}(X)$ and $\beta \in \mathcal{P}(X^+)$, the dimerization and dissociation reactions rate constants have the forms*

$$r(\alpha \to L\beta) = c_a \frac{\mathrm{Z}[\, \alpha \cap \beta \,]}{\mathrm{Z}[\, \alpha \,]}$$

and

$$r(L\beta \to \alpha) = c_d \frac{\mathrm{Z}[\,\alpha \cap \beta\,]}{\mathrm{Z}[\,\beta\,]} \cdot \exp(\mathfrak{b}\theta_L).$$

Proof. Let $\alpha \in \mathcal{P}(X)$ and $\beta \in \mathcal{P}(X^+)$. Thus

$$\begin{aligned} r(\alpha \to L\beta) &= \sum_{\substack{x \in \alpha \\ Ly \in L\beta}} \Pr[\,x \mid \alpha\,] \cdot k(x \to Ly) \\ &= \sum_{x \in \alpha \cap \beta} \Pr[\,x \mid \alpha\,] \cdot k(x \to Lx) \\ &= c_{\mathrm{a}} \frac{\mathrm{Z}[\,\alpha \cap \beta\,]}{\mathrm{Z}[\,\alpha\,]} \end{aligned}$$

and

$$\begin{aligned} r(L\beta \to \alpha) &= \sum_{\substack{Lx \in L\beta \\ y \in \alpha}} \Pr[\,Lx \mid L\beta\,] \cdot k(Lx \to y) \\ &= \sum_{x \in \alpha \cap \beta} \Pr[\,Lx \mid L\beta\,] \cdot k(Lx \to x) \\ &= c_{\mathrm{d}} \frac{\mathrm{Z}[\,\alpha \cap \beta\,]}{\mathrm{Z}[\,\beta\,] \cdot \exp(\mathfrak{b}\theta_L)}. \end{aligned}$$

□

6.4. A tractable model under ligand excess

For the described coarse-grained RNA ligand interaction process, partition the monomer microstates X and the dimer microstates X^* into respective sets of macrostates Ξ and Ξ^*. For the theoretical discussion, it is required only that Ξ and Ξ^* are partitions of the respective sets X and X^*. Later, during the application, macrostates are defined as gradient basins (within their respective component).

Denote the monomer macrostates in Ξ by $\alpha_1, \ldots, \alpha_n$ and the dimer macrostates in Ξ^* by $\beta_1, \ldots, \beta_m$. Since—by model assumption—the ligand is in large excess, the change of the ligand concentration $[L]$ is essentially negligible in relation to the change of RNA concentrations. Formally, this equates to the assumption $d/dt[L] = 0$, i. e. at all times $[L] = l_0$, for the initial ligand concentration l_0. The change of RNA monomer and RNA ligand dimer concentrations over time is described by a system of

linear, first-order *ordinary differential equations* (ODEs) corresponding to Reactions (6.1a)–(6.1d).

Following first-order rate laws (Section 3.5.2 on page 36), Reaction (6.1a) causes the flows $r(\alpha_i \to \alpha_j)[\alpha_i]$ from α_i to α_j (for $1 \leq i, j \leq n$); Reaction (6.1d), $r(\beta_i \to \beta_j)[\beta_i]$ from β_i to β_j for $1 \leq i, j \leq m$; and Reaction (6.1c), $r(\beta_i \to \alpha_j)[\beta_i]$ from β_i to α_j for $1 \leq i \leq m, 1 \leq j \leq n$. In contrast to these simple first-order transitions, the state changes due to Reaction 6.1b follow second-order rate laws contributing the flow $r(\alpha_i \to \beta_j)[L][\alpha_i]$ from α_i to β_j. Without the assumption $d/dt[L] = 0$, the rate would depend on two variable concentrations, causing the system to be non-linear. However, by the assumption of ligand excess, the concentration $[L]$ is constant.

The change of concentrations is now described by summing over the single contributions, yielding the following system of ODEs:

$$\begin{aligned}
(i = 1, \ldots, n) \quad \frac{d}{dt}[\alpha_i] = & \sum_{\substack{1 \leq k \leq n \\ k \neq i}} r(\alpha_k \to \alpha_i)[\alpha_k] \quad + \sum_{1 \leq k \leq m} r(\beta_k \to \alpha_i)[\beta_k] \\
& - \sum_{\substack{1 \leq k \leq n \\ k \neq i}} r(\alpha_i \to \alpha_k)[\alpha_i] \quad - \sum_{1 \leq k \leq m} r(\alpha_i \to \beta_k)[L][\alpha_i], \\
(j = 1, \ldots, m) \quad \frac{d}{dt}[\beta_j] = & \sum_{1 \leq k \leq n} r(\alpha_k \to \beta_j)[L][\alpha_k] \quad + \sum_{\substack{1 \leq k \leq m \\ k \neq j}} r(\beta_k \to \beta_j)[\beta_k] \\
& - \sum_{1 \leq k \leq n} r(\beta_j \to \alpha_k)[\beta_j] \quad - \sum_{\substack{1 \leq k \leq m \\ k \neq j}} r(\beta_j \to \beta_k)[\beta_j].
\end{aligned}$$

Let $\gamma = (\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m)^t$ and define the $(n+m) \times (n+m)$-matrix $R(l_0)$. Then the entire coarse-grained system under ligand excess can be expressed by the linear ODE

$$\frac{d}{dt}[\gamma] = R(l_0)[\gamma],$$

For this purpose, $R(l_0)$ is constructed as

$$R(l_0) = \begin{pmatrix} A & C \\ l_0 \cdot D & B \end{pmatrix} \tag{6.3}$$

from four submatrices:

A: $n \times n$-matrix with entries $a_{ij} = r(\alpha_i \leftarrow \alpha_j)$ for $1 \leq i, j \leq n$, $i \neq j$;
for $1 \leq i \leq n$, $a_{ii} = -\sum_{1 \leq k \leq n} r(\alpha_k \leftarrow \alpha_i) - \sum_{1 \leq k \leq m} l_0 r(\beta_k \leftarrow \alpha_i)$

B: $m \times m$-matrix with entries $b_{ij} = r(\beta_i \leftarrow \beta_j)$ for $1 \leq i, j \leq m$, $i \neq j$; for $1 \leq j \leq m$, $b_{jj} = -\sum_{1 \leq k \leq n} r(\alpha_k \leftarrow \beta_j) - \sum_{1 \leq k \leq m} r(\beta_k \leftarrow \beta_j)$

C: $n \times m$-matrix with entries $c_{ij} = r(\alpha_i \leftarrow \beta_j)$ for $1 \leq i \leq n, 1 \leq j \leq m$, and

D: $m \times n$-matrix with entries $d_{ij} = r(\beta_i \leftarrow \alpha_j)$ for $1 \leq i \leq m, 1 \leq j \leq n$.

6.5. Equilibrium distribution

The distribution of the RNA molecules to the macrostates in the equilibrium state of the full model can be calculated using the law of mass action (cf. Eq. (3.4) on page 38). Inside the world of monomers and dimers, respectively, the probability of each macrostate can be derived from its Boltzmann weight as described in Section 3.3 on page 32. However, in the presence of the ligand, one has to consider the distribution of the RNA to both worlds. This distribution depends on the ligand concentration and is derived in this section. Therefore, denote by R^+ the species of all RNAs that *can* bind the ligand (i. e. all conformations in X^+), by $\tilde{\mathrm{R}}$ the species of all RNAs that *cannot* bind the ligand (i. e. all conformations in $X \setminus X^+$), by L the ligand and by LR the species of all RNA–ligand dimer complexes. The transition between the monomer and dimer world is described by the reactions

$$\tilde{\mathrm{R}} \rightleftharpoons \mathrm{R}^+, \tag{6.4}$$

which describes the formation of structures with and without the binding pocket inside the monomer world, and the dimerization reaction

$$\mathrm{L} + \mathrm{R}^+ \rightleftharpoons \mathrm{LR}. \tag{6.5}$$

By the law of mass action and the Boltzmann distribution of energies (cf. Section 3.3 on page 32), one gets

$$\tilde{K} = \frac{[\mathrm{R}^+]}{[\tilde{\mathrm{R}}]} \tag{6.6a}$$

$$= \frac{\mathrm{Z}[\,X^+\,]}{\mathrm{Z}[\,X \setminus X^+\,]}$$

$$= \frac{\mathrm{Z}[\,X^+\,]}{\mathrm{Z}[\,X\,] - \mathrm{Z}[\,X^+\,]}, \tag{6.6b}$$

for Reaction (6.4) and, under the assumption of ligand excess,

$$\begin{aligned} K^+ &= \frac{[\mathrm{RL}]}{[\mathrm{R}^+] \cdot [\mathrm{L}]} \\ &= \frac{[\mathrm{RL}]}{[\mathrm{R}^+] \cdot l_0} \end{aligned} \tag{6.7a}$$

as well as

$$\begin{aligned} K^+ &= \frac{\mathrm{Z}[\, LX^+ \,]}{\mathrm{Z}[\, X^+ \,] \cdot \mathrm{Z}[\, \mathrm{L} \,]} \\ &= \frac{\mathrm{Z}[\, X^+ \,] \cdot \exp(-\mathfrak{b}\theta_L)}{\mathrm{Z}[\, X^+ \,]} \\ &= \exp(-\mathfrak{b}\theta_L) \end{aligned} \tag{6.7b}$$

for Reaction (6.5), where $\tilde{K}$ and K^+ are the equilibrium constants of their respective reactions. Note these can readily be computed by using the expressions (6.6b) and (6.7b).

By rewriting Eq. (6.6a) and Eq. (6.7a), one obtains

$$[\mathrm{R}^+] = \tilde{K} \cdot [\tilde{\mathrm{R}}] \tag{6.8a}$$

and

$$[\mathrm{RL}] = K^+ \cdot l_0 \cdot [\mathrm{R}^+]. \tag{6.8b}$$

Additionally, for the known initial RNA concentration r_0, the relation

$$r_0 = [\tilde{\mathrm{R}}] + [\mathrm{R}^+] + [\mathrm{RL}] \tag{6.8c}$$

holds, since any RNA molecule must be in either of these three molecule species. By solving the system of Eqs. (6.8a) to (6.8c), one can now derive the equilibrium concentrations for each species:

$$\begin{aligned} [\tilde{\mathrm{R}}] &= \frac{r_0}{1 + \tilde{K} + \tilde{K} \cdot K^+ \cdot l_0} \\ [\mathrm{R}^+] &= \frac{\tilde{K} \cdot r_0}{1 + \tilde{K} + \tilde{K} \cdot K^+ \cdot l_0} \\ [\mathrm{RL}] &= \frac{K \cdot K^+ \cdot l_0 \cdot r_0}{1 + \tilde{K} + \tilde{K} \cdot K^+ \cdot l_0} \end{aligned}$$

Knowing these, the fractions of RNA monomers and dimers can be calculated as $p_{\mathrm{mon}} = {}^{[\mathrm{R}]}/_{r_0}$, $[\mathrm{R}] = [\tilde{\mathrm{R}}] + [\mathrm{R}^+]$ and $p_{\mathrm{dim}} = 1 - p_{\mathrm{mon}} = {}^{[\mathrm{RL}]}/_{r_0}$,

respectively. Putting things together, the probability of a macrostate α in the equilibrated full system is

$$\Pr[\,\alpha\,] = \begin{cases} \Pr[\,\alpha \mid X\,] \cdot p_{\text{mon}} & \text{if } \alpha \in \Xi, \\ \Pr[\,\alpha \mid X^+\,] \cdot p_{\text{dim}} & \text{if } L\alpha \in \Xi^*. \end{cases} \tag{6.9}$$

To obtain the equilibrium concentration of a certain state, one can multiply its probability with the initial RNA concentration r_0.

6.6. Detailed balance

This section shows that described model is in detailed balance. The full rate matrix (Eq. (6.3) on page 77) incorporates the ligand concentrations. Therefore, for any two macrostates α and β with indices i and j, respectively, denote the *full macrorate coefficient* for the transition from α to β as

$$\hat{r}(\alpha \to \beta) = r_{ji},$$

where r_{ji} is the respective entry of the full rate matrix. Therefore, $\hat{r}(\alpha \to \beta) = l_0 \cdot r(\alpha \to \beta)$ if α is a monomer and β is a dimer macrostate. Else, $\hat{r}(\alpha \to \beta) = r(\alpha \to \beta)$.

Theorem 5 (Detailed balance)**.** *The system described in Section 6.4 on page 76 is in detailed balance.*

Proof. Recalling Eq. (3.1) on page 36, the system is in detailed balance if, in the equilibrated system,

$$\Pr[\,\alpha\,] \cdot \hat{r}(\alpha \to \beta) = \Pr[\,\beta\,] \cdot \hat{r}(\beta \to \alpha)$$

holds for any two macrostates α and β. If the rate coefficients are both zero, the claim is trivial. Else, *both* constants are non-zero, as follows directly from Eq. (6.2) on page 73 and the symmetry of the neighborhood relation $\mathcal{N}$. Therefore, the previous equation is equivalent to

$$\frac{\Pr[\,\alpha\,]}{\hat{r}(\beta \to \alpha)} = \frac{\Pr[\,\beta\,]}{\hat{r}(\alpha \to \beta)}. \tag{6.10}$$

The detailed balance of the system is now shown by proving that Eq. (6.10) holds for all macrostates α and β. Let, as before, Ξ and Ξ^* denote the sets of monomer and dimer macrostates, respectively. Note that the claim is trivial for $\alpha = \beta$, so assume $\alpha \neq \beta$. One can now distinguish several cases.

First, let $\alpha, \beta \in \Xi$. Then, $\hat{r}(\alpha \to \beta) = r(\alpha \to \beta)$. Further, by Eq. (6.9) on page 80,

$$\begin{aligned}\frac{\Pr[\,\alpha\,]}{\Pr[\,\beta\,]} &= \frac{\Pr[\,\alpha \mid X\,] \cdot p_{\mathrm{mon}}}{\Pr[\,\beta \mid X\,] \cdot p_{\mathrm{mon}}} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,X\,]} \cdot \frac{\mathrm{Z}[\,X\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{p_{\mathrm{mon}}}{p_{\mathrm{mon}}} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]}.\end{aligned}$$

Let $E_{x,y}^{\max} = \max\{E(x), E(y)\}$. Inserting the definitions of the macrorate and microrate coefficients, the ratio of the rate coefficients is

$$\begin{aligned}\frac{\hat{r}(\beta \to \alpha)}{\hat{r}(\alpha \to \beta)} &= \frac{\sum_{y\in\beta} \Pr[\,y \mid \beta\,] \sum_{x\in\alpha} k(y \to x)}{\sum_{x\in\alpha} \Pr[\,x \mid \alpha\,] \sum_{y\in\beta} k(x \to y)} \\ &= \frac{\sum_{y\in\beta} \left(\exp(-\mathfrak{b}E(y))/\mathrm{Z}[\,\beta\,]\right) \sum_{x\in\alpha} c_{\mathrm{R}} \exp(-\mathfrak{b}[E_{x,y}^{\max} - E(y)])}{\sum_{x\in\alpha} \left(\exp(-\mathfrak{b}E(x))/\mathrm{Z}[\,\alpha\,]\right) \sum_{y\in\beta} c_{\mathrm{R}} \exp(-\mathfrak{b}[E_{x,y}^{\max} - E(x)])} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{\sum_{y\in\beta} \exp(-\mathfrak{b}E(y)) \sum_{x\in\alpha} \exp(-\mathfrak{b}[E_{x,y}^{\max} - E(y)])}{\sum_{x\in\alpha} \exp(-\mathfrak{b}E(x)) \sum_{y\in\beta} \exp(-\mathfrak{b}[E_{x,y}^{\max} - E(x)])} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{\sum_{y\in\beta} \sum_{x\in\alpha} \exp(-\mathfrak{b}E_{x,y}^{\max})}{\sum_{x\in\alpha} \sum_{y\in\beta} \exp(-\mathfrak{b}E_{x,y}^{\max})} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \\ &= \frac{\Pr[\,\alpha\,]}{\Pr[\,\beta\,]},\end{aligned}$$

so for these transitions the criterion is fulfilled.

Analogously, for $L\alpha, L\beta \in \Xi^*$ one gets

$$\frac{\Pr[\,L\alpha\,]}{\Pr[\,L\beta\,]} = \frac{\mathrm{Z}[\,L\alpha\,]}{\mathrm{Z}[\,L\beta\,]} = \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]}.$$

Since $\hat{r}(L\alpha \to L\beta) = r(L\alpha \to L\beta) = r(\alpha \to \beta)$ (cf. Lemma 5 on page 74) and, equally, $\hat{r}(L\beta \to L\alpha) = r(\beta \to \alpha)$, the ratio of the rate coefficients is also $\mathrm{Z}[\,\alpha\,]/\mathrm{Z}[\,\beta\,]$.

In the third case, $\alpha \in \Xi$ while $L\beta \in \Xi^*$. Now,

$$\begin{aligned}\frac{\Pr[\,\alpha\,]}{\Pr[\,L\beta\,]} &= \frac{\Pr[\,\alpha \mid X\,]}{\Pr[\,L\beta \mid X^*\,]} \cdot \frac{p_{\mathrm{mon}}}{p_{\mathrm{dim}}} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,X\,]} \cdot \frac{\mathrm{Z}[\,X^*\,]}{\mathrm{Z}[\,L\beta\,]} \cdot \frac{1+\tilde{K}}{1+\tilde{K}+\tilde{K}\cdot K^+\cdot l_0} \cdot \frac{1+\tilde{K}+\tilde{K}\cdot K^+\cdot l_0}{\tilde{K}\cdot K^+\cdot l_0} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{\mathrm{Z}[\,X^+\,]}{\mathrm{Z}[\,X\,]} \cdot \frac{1+\tilde{K}}{\tilde{K}\cdot K^+\cdot l_0}.\end{aligned}$$

To further simplify this expression, Eq. (6.6b) on page 78 can be rewritten as

$$\frac{\tilde{K}}{1+\tilde{K}} = \frac{\mathrm{Z}[\,X^+\,]}{\mathrm{Z}[\,X\,]},$$

such that

$$\begin{aligned}\frac{\Pr[\,\alpha\,]}{\Pr[\,L\beta\,]} &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{\tilde{K}}{1+\tilde{K}} \cdot \frac{1+\tilde{K}}{\tilde{K}\cdot K^+\cdot l_0} \\ &= \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{1}{K^+\cdot l_0} \\ &\overset{(6.7\mathrm{b})}{=} \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\beta\,]} \cdot \frac{\exp(\mathfrak{b}\theta_L)}{l_0} \cdot \frac{c_{\mathrm{a}}}{c_{\mathrm{a}}} \\ &\overset{(6.11)}{=} c_{\mathrm{d}} \frac{\mathrm{Z}[\,\alpha\cap\beta\,]}{\mathrm{Z}[\,\beta\,]} \exp(\mathfrak{b}\theta_L) \frac{\mathrm{Z}[\,\alpha\,]}{\mathrm{Z}[\,\alpha\cap\beta\,]} \cdot \frac{1}{l_0\cdot c_{\mathrm{a}}} \\ &\overset{(*)}{=} \frac{r(L\beta\to\alpha)}{r(\alpha\to L\beta)\cdot l_0} \\ &= \frac{\hat{r}(L\beta\to\alpha)}{\hat{r}(\alpha\to L\beta)},\end{aligned}$$

where the equivalence $(*)$ follows from Lemma 7 on page 75. □

6.7. Computing RNA–ligand kinetics

The described ODE system can be solved analytically building on existing software. The entire computation pipeline can be outlined as follows:

1. enumeration of the RNA's structure space
2. computation of the gradient basins and rates between these for
   a) the monomer landscape

b) the dimer landscape

3. computation of the rates between the monomer and dimer basins
4. construction of the full rate matrix $R(l_0)$
5. integration of the linear ODE system

Since an exhaustive enumeration of the structure space is infeasible even for short RNAs, *Step 1* generates only a selected part of all possible secondary structures of the input RNA. For this work, only structures up to a certain energy above the minimum free energy of the sequence as computed by `RNAsubopt` (Lorenz et al. 2011) are considered. To further reduce the number of structures, only canonical structures as described in Chapter 4 on page 41 are considered.

Often, the restriction to low energy structures excludes important microstates of relatively high energy such as the open RNA chain from the model. Simply adding such a structure to the system is insufficient without also including transitional structures that connect it to the remaining states. The solution for this work was to develop a heuristic algorithm to partially explore an energy landscape around a given structure of interest (cf. Chapter 5 on page 59). This approach seems to be more adequate in the context of a gradient basin coarse graining than a direct path heuristic (e. g. `findPath` (Flamm et al. 2000)).

In *Step 2a*, the gradient basins and rates for the monomer landscape from the list of input structures is computed using `barriers` (Wolfinger et al. 2004) (with `minh` heuristic). For *Step 2b*, a list of all input structures that contain the binding pocket is generated with `RNAsubopt`'s constraint folding mode. This enables one to enumerate dimer structures up to a higher energy than possible for the entire landscape, ensuring the dimer world is connected. As shown in Lemma 5 on page 74, the transition rates in the constrained dimer landscape are independent of the ligand's binding energy and thus can be computed exactly like those of the monomer landscape.

In *Step 3*, the transition rates between monomer and dimer macrostates are computed based on Lemma 7 on page 75 using the mapping of the monomer and dimer structures to their respective basins. For this purpose `barriers` has been modified to output the required information.

Step 4 yields the full rate matrix $R(l_0)$ for one set of pre-exponential factors and a certain ligand concentration l_0 by combining the previously computed rate constants. Take note that one can easily compute $R(l_0)$

for different values of l_0, c_a and c_R without repeating the previous, more time-consuming computation steps.

Finally, in *Step 5* the described ODE system is solved exactly with the tool `treekin` (Wolfinger et al. 2004). After diagonalizing $R(l_0)$, it efficiently computes the development of the macrostates' concentrations during an arbitrary time interval.

6.8. Parameters from empirical measurements

The binding energy θ_L can be derived from an empirically measured dissociation constant K_d^A of the aptamer; e. g. in the case of theophylline, Jenison et al. (1994) measure a K_d^A of 0.32 µM for the theophylline aptamer of RS3. From the macroscopic measurement, one can, as described in Section 3.6 on page 38, derive the binding energy as

$$\theta_L = RT_\mathrm{A} \ln\Big(K_\mathrm{d}^\mathrm{A} \cdot \Pr[\,\text{“pocket”} \mid \mathrm{A}, T_\mathrm{A}\,]\Big),$$

where $T_\mathrm{A} = 298\,\mathrm{K}$ is the temperature of the measurement, R is the gas constant, and $\Pr[\,\text{“pocket”} \mid \mathrm{A}, T_\mathrm{A}\,]$ denotes the equilibrium probability of the binding pocket in the aptamer at temperature T_A as calculated in the Turner energy model (cf. Wachsmuth et al. 2013, which neglect the probability). This relation allows calculating the effective dissociation constant at temperature T_R of a theophylline riboswitch like RS3 that contains the aptamer, due to the inverse relation

$$\begin{aligned} K_\mathrm{d}^\mathrm{RS} &= \frac{\exp\left(\frac{\theta_L}{RT_\mathrm{R}}\right)}{\Pr[\,\text{“pocket”} \mid \mathrm{RS}, T_\mathrm{R}\,]} \\ &= \frac{\exp\left(\frac{RT_\mathrm{A}}{RT_\mathrm{R}} \ln\left(K_\mathrm{d}^\mathrm{A} \cdot \Pr[\,\text{“pocket”} \mid \mathrm{A}, T_\mathrm{A}\,]\right)\right)}{\Pr[\,\text{“pocket”} \mid \mathrm{RS}, T_\mathrm{R}\,]} \\ &= \frac{\left(K_\mathrm{d}^\mathrm{A} \cdot \Pr[\,\text{“pocket”} \mid \mathrm{A}, T_\mathrm{A}\,]\right)^{T_\mathrm{A}/T_\mathrm{R}}}{\Pr[\,\text{“pocket”} \mid \mathrm{RS}, T_\mathrm{R}\,]}. \end{aligned}$$

For RS3 at $T_\mathrm{R} = 313.15\,\mathrm{K}$,

$$\Pr[\,\text{“pocket”} \mid \mathrm{A}, T_\mathrm{A}\,] \approx 0.292$$

and

$$\Pr[\,\text{“pocket”} \mid \mathrm{RS}, T_\mathrm{R}\,] \approx 2.59 \times 10^{-11}$$

due to the pocket-constrained and unconstrained ensemble free energies in the TURNER model. Thus,

$$\theta_L \approx RT_\mathrm{A} \ln(0.292 K_\mathrm{d}^\mathrm{A}) \approx -9.59\,\mathrm{kcal\,mol^{-1}}$$

and

$$K_\mathrm{d}^\mathrm{RS3} \approx \frac{(0.292 K_\mathrm{d}^\mathrm{A})^{T_\mathrm{A}/T_\mathrm{R}}}{2.59 \cdot 10^{-11}} \quad \approx 7891\,\mathrm{M}.$$

For relating the rates of the different reaction types, one needs to estimate the pre-exponential factors of all reactions. Commonly, one assumes constant factors for each type of reaction. Furthermore, it is reasonable to equate the factors for monomer and dimer conformation changes.

Given the apparent association rate c_a^m (which we assume to equal the macroscopic pre-exponential factor of dimerization), one can bound the microscopic pre-exponential factor c_a. Assuming that refolding is much slower than dimerization, c_a^m is a product of the microrate and the equilibrium probability of the binding pocket. Conversely, if the refolding is assumed to be much faster, than c_a^m directly measures the dimerization microrate. Thus,

$$c_\mathrm{a}^\mathrm{m} \quad \leq \quad c_\mathrm{a} \quad \leq \quad c_\mathrm{a}^\mathrm{m} \cdot \Pr[\,\text{“pocket”} \mid \text{aptamer}\,]^{-1}.$$

In the case of theophylline, $\Pr[\,\text{“binding pocket”} \mid \text{aptamer}\,] \approx 1$ and consequently, $c_\mathrm{a} \approx c_\mathrm{a}^\mathrm{m}$.

Finally, the pre-exponential factor for dissociation c_d equals c_a. This is a consequence of detailed balance of the dimerization reaction, i. e.

$$k(R_i \to LR_i) \Pr[\,R_i\,] = k(LR_i \to R_i) \Pr[\,LR_i\,],$$

which implies

$$c_\mathrm{a} \Pr[\,R_i\,] = c_\mathrm{d} \exp(\mathfrak{b}\theta_L) \Pr[\,R_i\,] \exp(-\mathfrak{b}\theta_L) = c_\mathrm{d} \Pr[\,R_i\,]. \tag{6.11}$$

6.9. Ligand intake into the cell

Aiming to use designed riboswitches as a tool for *in vivo* experiments, any useful simulation of the folding process needs to consider the physiological conditions inside living cells. As already stated, the ligand concentration is an important parameter to the folding kinetics of riboswitches. In an

in vivo experiment, however, the experimenter adds a certain amount of ligand *to the medium* the cells are growing in. To get to the location of RNA transcription or translation and interact with a riboswitch, the ligand needs to permeate through the lipid bilayer surrounding all cells, following the gradient of concentrations until the concentrations outside and inside of the cell are equal. Depending on how well a certain molecule can pass through these membranes, the equilibration of the concentrations may take some time. For example, polar molecules typically have a lower rate of diffusion through lipid bilayers while water molecules can get past them rather quick. This raises the question how long it actually takes for a given ligand to enter the cells and begin its interaction process.

To answer this question, one needs to resort to FICK's laws of diffusion (Crank 1979) which describe the *flux* J, the flow of particles into the cell, as

$$J = P_L(c_o - c_i), \tag{6.12}$$

where c_o and c_i are the concentrations of the ligand outside and inside the cell, respectively, and P is the *permeability coefficient* for the ligand L of interest. The unit of P is $\mathrm{m\,s^{-1}}$. Assuming that the concentrations are given in their SI default units $\mathrm{mol\,m^{-3}}$, the flux J has the unit $\mathrm{mol\,m^{-2}\,s^{-1}}$, i. e. it describes how many particles enter the cell depending on the size of its surface area A.

The idea now is to model the concentration $c_i = c_i(t)$ as a function of the time. This can easily be done using an ordinary differential equation (ODE): the change of the concentration over time equals the number of particles entering the cell divided by the cell's volume, i. e.

$$\begin{aligned} \dot{c_i}(t) = \frac{dc_i}{dt}(t) &= A \cdot V^{-1} \cdot J \\ &= AV^{-1}P_L(c_o(t) - c_i(t)). \end{aligned}$$

In Eq. (6.13), the concentration c_o also depends on the time. This is correct since any molecule that enters the cell reduces the concentration outside of it. However, since the volume of the medium is huge compared to that of the cell, one can assume $c_o(t) \equiv c_o$ to be constant. Furthermore, c_o is known since it is the initial concentration the experimenter adds to the medium, and $c_i(0) = 0$ since in the beginning there is no ligand inside the cell[1]. By letting $k = AV^{-1}P_L$ and dropping the index of c_i, Eq. (6.13) on

[1] Of course, this is only the case if the ligand is not naturally present in the cell, but in this case, it is not sensible to use it to toggle a riboswitch anyway.

page 86 becomes

$$\dot{c}(t) = k(c_o - c(t)) = kc_o - kc(t) \tag{6.13}$$

which is a linear, inhomogeneous, first-order ODE. This class of ODEs can easily be integrated. The solution has the general form $c(t) = c_h(t) + c_s(t)$, where c_s is a special solution for Eq. (6.13) and c_h is the general solution for the homogeneous ODE

$$\dot{c_h}(t) = -kc_h(t). \tag{6.14}$$

Obviously, the constant function $c_s(t) \equiv c_o$ is a solution to Eq. (6.13) since $\frac{dc_o}{dt} = 0 = kc_o - kc_o$. The general solution for Eq. (6.14) can be obtained by separation of variables:

$$\begin{aligned}
& \frac{dc_h}{dt} = -kc_h \\
\Leftrightarrow \quad & \int \frac{dc_h}{-kc_h} = \int dt \\
\Leftrightarrow \quad & -\frac{1}{k} \ln |kc_h| = t + a \\
\Leftrightarrow \quad & c_h = \frac{1}{k} \exp(-ka) \exp(-kt)
\end{aligned}$$

where a is the integration constant. The general solution for Eq. (6.13) is therefore

$$c(t) = c_h(t) + c_s(t) = \frac{1}{k} \exp(-ka) \exp(-kt) + c_o.$$

Inserting the initial value $c(0) = 0$, one can determine a to solve the initial value problem (IVP):

$$\begin{aligned}
0 &= \frac{1}{k} \exp(-ka) \cdot 1 + c_o \\
\Leftrightarrow \quad \exp(-ka) &= -kc_o \\
\Leftrightarrow \quad -ka &= \ln |-kc_o| \\
\Leftrightarrow \quad a &= -\frac{1}{k} \ln |-kc_o|
\end{aligned}$$

Inserting this constant into $c(t)$, the final solution of the IVP:

$$c_{\mathrm{IVP}}(t) = -c_o \exp(-kt) + c_o = c_o(1 - \exp(-kt)) \tag{6.15}$$

As can be seen directly from Eq. (6.15) on page 87, the concentration inside the cell scales linearly with the initial concentration, i. e. the duration of the equilibration process does not depend on c_o. By setting it to 1, $c_{\mathrm{IVP}}|_{c_o=1}(t)$ therefore describes the fraction of the initial concentration that is diffused into the cell at time t.

To apply Eq. (6.15) on page 87 to the problem of the ligand theophylline diffusing into, for example, an *E. coli* cell, the parameter k needs to be specified. The average cell surface of *E. coli* is $A = 6\,\mu\mathrm{m}^2$, its average volume is $V = 1\,\mu\mathrm{m}^3$ (Gilbert 2009, p. 26). The permeability coefficient is $P_{\mathrm{theophylline}} = 2.9 \times 10^{-6}\,\mathrm{m\,s}^{-1}$ (Gutknecht and Walter 1981). Therefore, $k = 11.6\,\mathrm{s}^{-1}$. A plot of $c_{\mathrm{IVP}}|_{c_o=1,k=11.6\,\mathrm{s}^{-1}}(t)$ is shown in Fig. 6.2 on the next page.

The ligand concentration reaches the equilibrium in less than one second. As a consequence, in practice it can be assumed that the ligand concentration inside the cells of the experiment is equal to the one of the surrounding medium for theophylline and *E. coli*. For larger cells, the equilibration takes longer since the parameter k which linearly scales the time axis in Eq. (6.15) on page 87 grows inversely proportional to the cell diameter. That is because $k = AV^{-1}P$ and, assuming a cell has a spherical shape, its surface area A grows quadratically while its volume V grows cubically in the diameter. However, given that the diameter of e. g. a human cell (Kuse et al. 1985; Luciani et al. 2001) is only about ten to 15 times larger than that of *E. coli*, the concentrations inside and outside of the cell may be assumed to be equal even for large eukaryotic cells.

It should be noted that the calculations in this section are subject to some simplifications. As stated before, the concentration in the medium is assumed to be constant. Also, the flux as calculated by Eq. (6.12) on page 86 does not consider the influence of differences of concentrations of other molecules inside and outside the cell and the resulting chemical potentials. The calculation also assumes that the ligand accumulates in the cell, neglecting degradation as well as efflux from the cell. The latter might be especially important if the ligand can interact with the various active or passive transportation mechanisms of the cell. As a final note, *E. coli*, like the most gram-negative bacteria, has not just one but two cell membranes that hinder the diffusion process (Gupta 2011).

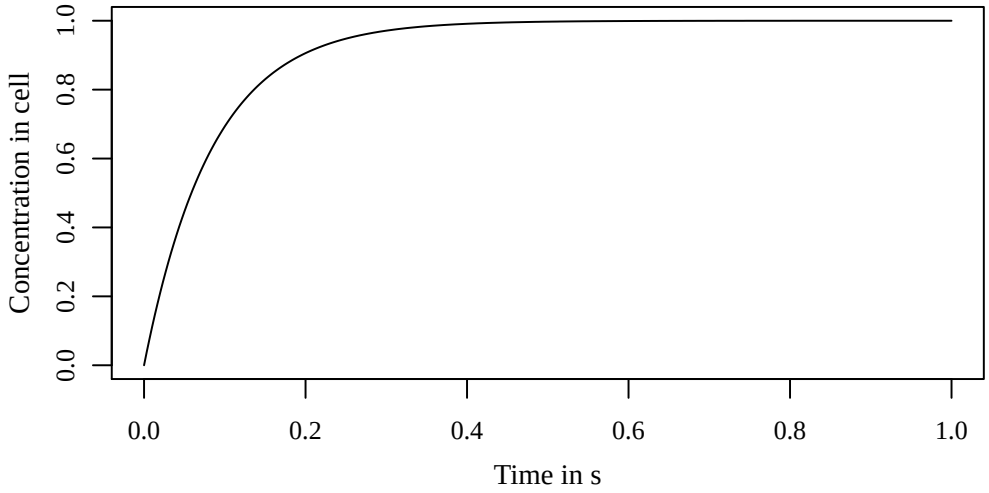


Figure 6.2.: Plot of $c_{\mathrm{IVP}}|_{c_o=1,k=11.6\,\mathrm{s}^{-1}}(t)$, which resembles the diffusion process of theophylline into an average *E. coli* cell. The fraction of the initial concentration in the medium that is diffused into the cell is plotted against the time in seconds. The equilibration process is nearly complete after just 1 s.

6.10. Empirical results

Applying the model to the designed *on* switch RS3 from Wachsmuth et al. (2013) demonstrates the effect of changes in ligand concentrations to the interaction of that riboswitch with its ligand theophylline. Using a prototypical software, the macroprocess for RS3 including rate constants is precomputed in several hours. Subsequently, computing the kinetics for each combination of concentrations and pre-exponential factors takes only a few seconds on off-the-shelf hardware (e. g. Core i5-750 @ 2.67 GHz). Figure 6.3 on page 91 summarizes the results; each subfigure plots the probabilities of prominent monomer and dimer states over time.

The pre-exponential factors are set to the estimations $c_{\mathrm{R}} = 1 \times 10^6\,\mathrm{s}^{-1}$ and $c_{\mathrm{a}} = 600\,\mathrm{M}^{-1}\,\mathrm{s}^{-1}$ as described before. This allows interpreting the time and ligand concentrations in concrete units and relates the speed of folding and dimerization.

Figures 6.3A-C show the results for ligand concentrations 10^4 M, 10^5 M, and 10^6 M. In the RS3 riboswitch, the aptamer domain is fused to a rho-independent terminator at the 3'-end. Thus, during transcription the aptamer is available shortly before the strong terminator stem can be formed and then dominates the entire structure ensemble. Therefore, the

partially transcribed riboswitch RS3 that is shortened by the 3′-half of the terminator stem and the 3' poly-U stretch is studied as well. The kinetics of the shortened riboswitch is shown for concentrations of 10^{-7} M, 10^{-6} M, and 10^{-3} M in respective Figures 6.3D-F. Note that the time scales for interaction of RS3 with theophylline are in accordance with the computed dissociation constant $K_{\mathrm{d}}^{\mathrm{RS3}}$, which implies that the monomer and dimer concentrations are balanced at about 10^4 M ligand concentration. This extreme concentration suggests that the riboswitch would be non-functional without further, probably co-transcriptional, effects. This is a plausible hypothesis since RS3 was designed to regulate at the transcriptional level.

The estimated rates are derived from a small number of empirical measurements at different conditions, such as ion concentrations (100 mM NaCl in Kuznetsov and Ansari (2012), 5 mM $MgCl_2$ and 0.5 M NaCl in Jenison et al. (1994), no Mg^{2+} and 100 mM NaCl in Latham et al. (2009)), temperatures, and actual sequences; hence they are not directly comparable. Nevertheless, they provide reasonable ball park estimates, because one can observe that the qualitative behavior of the system is robust against variations of these parameters by several orders of magnitude.

6.11. Discussion

Several refinements of the model remain for future research. Most importantly, the assumption that there is only a single binding motif is rather stringent. In general, multiple binding motifs with different binding energies are plausible. A corresponding generalization of the model naturally leads to multiple "ligand worlds" for the different binding modes. Furthermore, some ligands, such as Mg^{2+} have multiple binding sites. The current implementation of the Arrhenius approximation of the RNA folding kinetics, finally, is quite simplistic, using only a single kinetic pre-factor for all structural rearrangements. A refined model would presumably distinguishing constants for nucleation, stack extension, base pair sliding, and loop pinching. In particular riboswitches that control at the transcriptional level will strongly depend of the kinetics of transcription, i. e., the growth of the RNA chain itself. After all, growing RNA molecules are known to favor different local minima and thus to refold globally as the chain becomes longer (Hofacker et al. 2010). Conceptually it is not difficult to extend the current framework. However, experimental measurements are required to gauge additional thermodynamic parameters, kinetic pre-factors, and transcriptional speed—and these are very scarce at present.

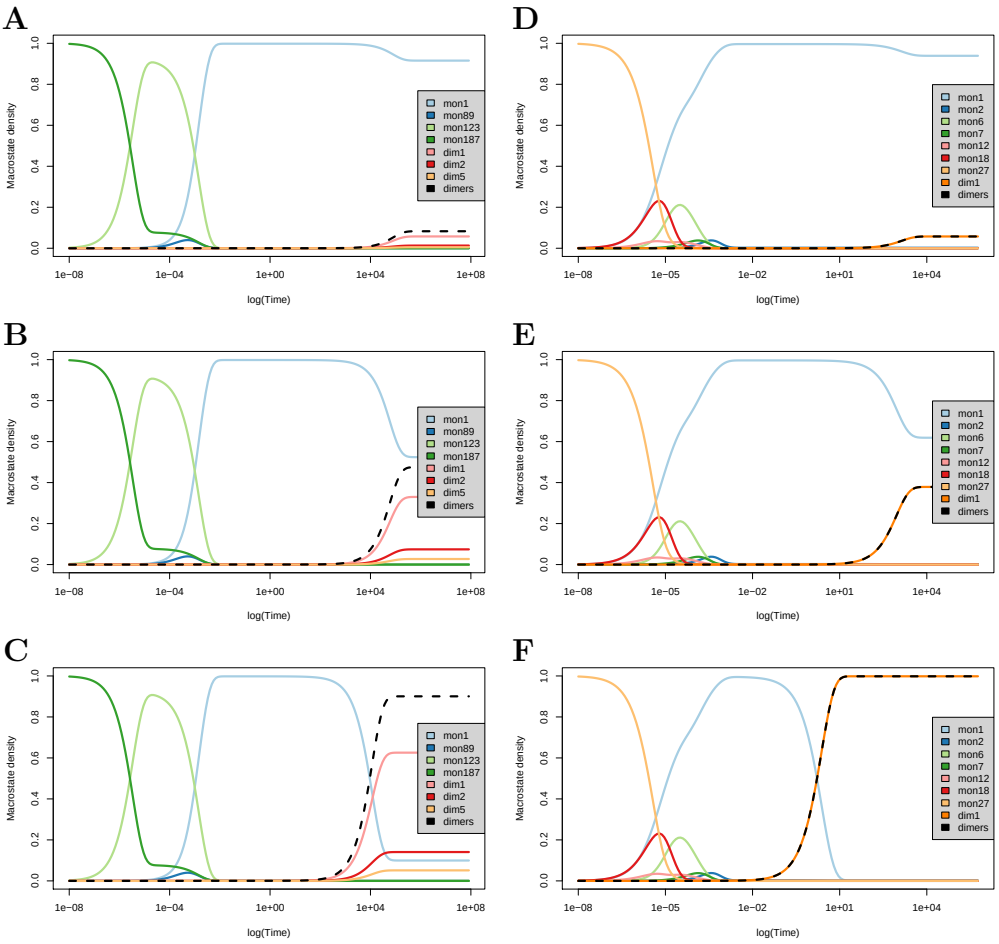


Figure 6.3.: Kinetics plots showing the probabilities of prominent monomer and dimer states over time. **A-C** Complete riboswitch RS3 at (unit-free) concentrations 10^4 (**A**), 10^5 (**B**), and 10^6 (**C**). **D-F** Partially transcribed riboswitch RS3 (without 3'-half of terminator stem) at concentrations 10^{-7} (**D**), 10^{-6} (**E**), and 10^{-3} (**F**). Folding rate: 10^6, dimerization rate: 600

Chapter 7.

Conclusion

In this work, a solid framework to model the folding kinetics of riboswitches has been developed. Utilizing a number of sensible approximations and heuristics, it is feasible to evaluate the simulation for a large number of different parameters and ligand concentrations. It has been shown that the kinetics of riboswitches may be strongly affected by co-transcriptional effects. Furthermore, the theoretical results from Chapters 4 and 5 on page 41 and on page 59 are useful on their own. It is planned that the implementation of the symmetric canonical move set will be included in upcoming versions of `barriers`. A direct path search heuristic for canonical landscapes based on the idea of `findPath` has been developed which can also be used to estimate barrier heights. The partial exploration tool for RNA landscapes is another result of this work.

As already stated in each individual section, there are several possible directions for possible future work. For example, the model may be extended to support multiple binding pocket conformations yielding different energy bonuses when the ligand binds to them. Another possible improvement would be to consider different pre-exponential factors for different types of moves, e. g. for loop formation and helix zippering, when better measurements of such parameters become available. With exception of the move set, which was implemented in *C*, the tools developed in this work are written in *Perl*. This speeded up the development, but results in a worse overall performance compared to an implementation in a faster, compiled language. One could therefore consider a re-implementation of one or several of the tools for optimized performance. Further, a parallelization of the algorithms should not be too hard to accomplish and would further decrease the runtime.

All in all, this work contributes to the better understanding of riboswitches and provides a valuable method to analyze their behavior *in silico*.

Bibliography

Akutsu, Tatsuya (2000). "Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots". In: *Discrete Applied Mathematics* 104.1–3, pp. 45–62. ISSN: 0166–218X. DOI: http://dx.doi.org/10.1016/S0166-218X(00)00186-4 (cit. on pp. 14, 29).

Bernhart, Stephan H., Hakim Tafer, Ulrike Mückstein, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker (2006). "Partition function and base pairing probabilities of RNA heterodimers". In: *Algorithms Mol Biol* 1.1, p. 3. ISSN: 1748-7188. DOI: 10.1186/1748-7188-1-3 (cit. on p. 9).

Bompfünewerer, Athanasius F., Rolf Backofen, Stephan H. Bernhart, Jana Hertel, Ivo L. Hofacker, Peter F. Stadler, and Sebastian Will (2007). "Variations on RNA folding and alignment: lessons from Benasque". In: *Journal of Mathematical Biology* 56.1, pp. 129–144. ISSN: 1432-1416. DOI: 10.1007/s00285-007-0107-5 (cit. on pp. 10, 41).

Breaker, Ronald R. (2011). "Prospects for riboswitch discovery and analysis". In: *Molecular cell* 43.6, pp. 867–879 (cit. on pp. 16, 18).

Bremer, Hans and Patrick P. Dennis (2008). "Modulation of chemical composition and other parameters of the cell at different exponential growth rates". In: *EcoSal Plus* 3.1 (cit. on p. 15).

Chang, Kung-Yao and Ignacio Tinoco Jr. (1997). "The structure of an RNA "kissing hairpin" complex of the HIV TAR hairpin loop and its complement". In: *Journal of Molecular Biology* 269.1, pp. 52–66. ISSN: 0022-2836. DOI: http://dx.doi.org/10.1006/jmbi.1997.1021. URL: http://www.sciencedirect.com/science/article/pii/S0022283697910214 (cit. on p. 29).

Clote, Peter, Evangelos Kranakis, Danny Krizanc, and Bruno Salvy (2009). "Asymptotics of canonical and saturated RNA secondary structures". In: *Journal of bioinformatics and computational biology* 7.05, pp. 869–893 (cit. on pp. 35, 42).

Cocco, S., J. F. Marko, and R Monasson (2003). "Slow nucleic acid unzipping kinetics from sequence-defined barriers". In: *Eur Phys J E Soft Matter* 10, pp. 153–161 (cit. on p. 71).

Crank, J. (1979). *The Mathematics of Diffusion*. Oxford science publications. Clarendon Press. ISBN: 9780198534112. URL: https://books.google.de/books?id=eHANhZwVouYC (cit. on p. 86).

Dimitrov, R. A. and M. Zuker (2004). "Prediction of hybridization and melting for double-stranded nucleic acids". In: *Biophys. J.* 87.1, pp. 215–226 (cit. on p. 9).

Espah Borujeni, Amin, Dennis M. Mishler, Jingzhi Wang, Walker Huso, and Howard M. Salis (2015). "Automated physics-based design of synthetic riboswitches from diverse RNA aptamers". In: *Nucleic Acids Res* 44.1, pp. 1–13. DOI: 10.1093/nar/gkv1289. URL: http://dx.doi.org/10.1093/nar/gkv1289 (cit. on p. 9).

Flamm, C., W. Fontana, I. L. Hofacker, and P. Schuster (2000). "RNA folding at elementary step resolution". In: *RNA* 6.3, pp. 325–38 (cit. on pp. 9, 10, 14, 60, 66, 69, 83).

Flamm, Christoph and Ivo Hofacker (2008). "Beyond energy minimization: approaches to the kinetic folding of RNA". In: *Chemical Monthly* 139, pp. 447–457 (cit. on p. 51).

Flamm, Christoph, Ivo L. Hofacker, Sebastian Maurer-Stroh, Peter F. Stadler, and Martin Zehl (2001). "Design of multistable RNA molecules". In: *RNA* 7.02, pp. 254–265 (cit. on p. 19).

Flamm, Christoph, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger (2002). "Barrier Trees of Degenerate Landscapes". In: *Zeitschrift fuer Physikalische Chemie* 216.2/2002. DOI: 10.1524/zpch.2002.216.2.155. URL: http://dx.doi.org/10.1524/zpch.2002.216.2.155 (cit. on pp. 10, 23, 42, 60, 70).

Giegerich, Robert, Björn Voß, and Marc Rehmsmeier (2004). "Abstract shapes of RNA". In: *Nucleic acids research* 32.16, pp. 4843–4851 (cit. on p. 35).

Gilbert, Robert (2009). "Physical biology of the cell, by Rob Phillips, Jane Kondev and Julie Theriot". In: *Crystallography Reviews* 15.4, pp. 285–288. DOI: 10.1080/08893110903104081 (cit. on p. 88).

Gupta, Radhey S (2011). "Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes". In: *Antonie Van Leeuwenhoek* 100.2, pp. 171–182 (cit. on p. 88).

Gutknecht, John and Anne Walter (1981). "Histamine, theophylline and tryptamine transport through lipid bilayer membranes". In: *Biochimica et Biophysica Acta (BBA)-Biomembranes* 649.2, pp. 149–154 (cit. on p. 88).

Hofacker, Ivo L., Peter Schuster, and Peter F. Stadler (1998). "Combinatorics of RNA secondary structures". In: *Discrete Applied Mathematics* 88.1, pp. 207–237 (cit. on p. 34).

Hofacker, Ivo L., Christoph Flamm, Christian Heine, Michael T. Wolfinger, Gerik Scheuermann, and Peter F. Stadler (2010). "BarMap: RNA folding on dynamic energy landscapes". In: *RNA* 16.7, pp. 1308–16. ISSN: 1469–9001. DOI: 10.1261/rna.2093310 (cit. on pp. 9, 90).

Huang, Jiabin, Rolf Backofen, and Björn Voß (2012). "Abstract folding space analysis based on helices". In: *RNA* 18.12, pp. 2135–2147 (cit. on p. 35).

Jenison, R. D., S. C. Gill, A. Pardi, and B. Polisky (1994). "High-resolution molecular discrimination by RNA". In: *Science* 263.5152, pp. 1425–9. ISSN: 0036–8075 (cit. on pp. 84, 90).

Kucharík, Marcel, Ivo L. Hofacker, Peter F. Stadler, and Jing Qin (2014). "Basin Hopping Graph: a computational framework to characterize RNA folding landscapes". In: *Bioinformatics* 30.14, pp. 2009–2017 (cit. on p. 35).

Kühnl, Felix, Peter F. Stadler, and Sebastian Will (2016). "Tractable Kinetics of RNA–Ligand Interaction". In: *Proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA)*. To appear (cit. on p. 10).

Kuse, R., S. Schuster, H. Schübbe, S. Dix, and K. Hausmann (1985). "Blood lymphocyte volumes and diameters in patients with chronic lymphocytic leukemia and normal controls". In: *Blut* 50.4, pp. 243–248 (cit. on p. 88).

Kuznetsov, Serguei V. and Anjum Ansari (2012). "A Kinetic Zipper Model with Intrachain Interactions Applied to Nucleic Acid Hairpin Folding Kinetics". In: *Biophysical Journal* 102.1, pp. 101–111. DOI: 10.1016/j.bpj.2011.11.4017. URL: http://dx.doi.org/10.1016/j.bpj.2011.11.4017 (cit. on pp. 70, 71, 90).

Latham, Michael P., Grant R. Zimmermann, and Arthur Pardi (2009). "NMR Chemical Exchange as a Probe for Ligand-Binding Kinetics in a Theophylline-Binding RNA Aptamer". In: *J. Am. Chem. Soc.* 131.14, pp. 5052–5053. DOI: 10.1021/ja900695m. URL: http://dx.doi.org/10.1021/ja900695m (cit. on pp. 71, 90).

Lorenz, Ronny, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker (2011). "ViennaRNA Package 2.0". In: *Algorithms Mol Biol* 6, p. 26. ISSN: 1748–7188. DOI: 10.1186/1748-7188-6-26 (cit. on pp. 70, 83).

Luciani, Anna Maria, Antonella Rosi, Paola Matarrese, Giuseppe Arancia, Laura Guidoni, and Vincenza Viti (2001). "Changes in cell volume

and internal sodium concentration in HeLa cells during exponential growth and following lonidamine treatment". In: *European Journal of Cell Biology* 80.2, pp. 187 –195. ISSN: 0171-9335. DOI: http://dx.doi.org/10.1078/0171-9335-00102 (cit. on p. 88).

Mann, Martin, Marcel Kucharik, Christoph Flamm, and Michael T. Wolfinger (2014). "Memory efficient RNA energy landscape exploration". In: *Bioinformatics* 30.18, pp. 2584–2591. ISSN: 1367–4811. DOI: 10.1093/bioinformatics/btu337 (cit. on p. 9).

Mathews, David H., Matthew D. Disney, Jessica L. Childs, Susan J. Schroeder, Michael Zuker, and Douglas H. Turner (2004). "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.19, pp. 7287–7292. DOI: 10.1073/pnas.0401799101. eprint: http://www.pnas.org/content/101/19/7287.full.pdf (cit. on p. 32).

McQuarrie, Donald Allan and John Douglas Simon (1997). *Physical chemistry: a molecular approach*. Vol. 1. Sterling Publishing Company (cit. on pp. 12, 13, 33, 35, 36, 38).

Morgan, Steven R. and Paul G. Higgs (1998). "Barrier heights between ground states in a model of RNA secondary structure". In: *Journal of Physics A: Mathematical and General* 31.14, p. 3153. URL: http://stacks.iop.org/0305-4470/31/i=14/a=005 (cit. on p. 51).

Mortimer, Michael (2002). *Chemical kinetics and mechanism*. Vol. 1. Royal Society of Chemistry (cit. on p. 36).

Nawrocki, Eric P. and Sean R. Eddy (2013). "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29.22, pp. 2933–2935 (cit. on p. 29).

Palazzo, Alexander F. and Eliza S. Lee (2015). "Non-coding RNA: what is functional and what is junk?" In: *Frontiers in genetics* 6 (cit. on p. 16).

Pörschke, Dietmar (1974). "Model Calculations on the Kinetics of Oligonucleotide Double Helix Coil Transitions. Evidence for a Fast Chain Sliding Reaction". In: *Biophysical chemistry* 2, pp. 83–96 (cit. on p. 71).

Proceedings of the International Symposium on Bioinformatics Research and Applications (ISBRA) (2016). To appear (cit. on p. 10).

Siederdissen, Christian Höner zu, Stefan Hammer, Ingrid Abfalter, Ivo L. Hofacker, Christoph Flamm, and Peter F. Stadler (2013). "Computational design of RNAs with complex energy landscapes". In: *Biopolymers* 99.12, pp. 1124–1136 (cit. on p. 19).

Toan, N. M., G. Morrison, C. Hyeon, and D. Thirumalai (2008). "Kinetics of loop formation in polymer chains". In: *J Phys Chem B.* 112, pp. 6094–6106 (cit. on p. 71).

Tuerk, Craig and Larry Gold (1990). "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase". In: *Science* 249.4968, pp. 505–510 (cit. on p. 18).

Turner, Douglas H. and David H. Mathews (2009). "NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure". In: *Nucleic Acids Research* 38.Database, pp. D280–D282. DOI: 10.1093/nar/gkp892. URL: http://dx.doi.org/10.1093/nar/gkp892 (cit. on pp. 41, 70).

Vollhardt, Kurt Peter C. and Neil Eric Schore (2003). *Organic chemistry: Structure and function.* Ed. by Neil Eric Schore. 4. ed. New York: W. H. Freeman. ISBN: 9780716743743. URL: https://katalog.ub.uni-leipzig.de/Record/0002640683 (cit. on p. 11).

Wachsmuth, Manja, Sven Findeiss, Nadine Weissheimer, Peter F. Stadler, and Mario Morl (2013). "De novo design of a synthetic riboswitch that regulates transcription termination". In: *Nucleic Acids Res* 41.4, pp. 2541–51. ISSN: 1362–4962. DOI: 10.1093/nar/gks1330 (cit. on pp. 9, 17, 19, 59, 66, 70, 72, 84, 89).

Wachsmuth, Manja, Gesine Domin, Ronny Lorenz, Robert Serfling, Sven Findeiß, Peter F. Stadler, and Mario Mörl (2015). "Design criteria for synthetic riboswitches acting on transcription". In: *RNA biology* 12.2, pp. 221–231 (cit. on p. 19).

Wolfinger, Michael T., W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L. Hofacker, and Peter F. Stadler (2004). "Efficient computation of RNA folding dynamics". In: *Journal of Physics A: Mathematical and General* 37.17, pp. 4731–4741. URL: http://stacks.iop.org/0305-4470/37/4731 (cit. on pp. 9, 69–71, 83, 84).

Wuchty, Stefan, Walter Fontana, Ivo L. Hofacker, Peter Schuster, et al. (1999). "Complete suboptimal folding of RNA and the stability of secondary structures". In: *Biopolymers* 49.2, pp. 145–165 (cit. on p. 35).

Zhang, W. and S. J. Chen (2002). "RNA hairpin-folding kinetics". In: *Proc Natl Acad Sci USA* 99, pp. 1931–1936 (cit. on p. 71).

Zuker, Michael and Patrick Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information". In: *Nucleic Acids Research* 9.1, pp. 133–148. DOI: 10.1093/nar/9.1.133. eprint: http://nar.oxfordjournals.org/content/9/1/133.full.pdf+html (cit. on p. 14).

Appendix A.

Acknowledgments

At first, I would like to thank my supervisors: Peter for pointing me to this interesting topic and all his useful suggestions, and Sebastian for all the fruitful discussions whenever I needed them.

Without Petra's help, I would still try to find Permit A 38. Thank you. Many thanks to Jens for keeping our systems running, responding quickly to every request and installing dozens of software packages. Per day. Thanks to all my other colleagues who make working at our institute so pleasant, interesting and instructive.

Special thanks go to my girlfriend, Saskia, as well as my parents. Thanks for your support. Thanks for being there for me when I need you.

Appendix B.

Selbstständigkeitserklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe. Insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann.

Leipzig, den

Felix Kühnl