Prediction Algorithms for Restricted RNA Pseudoknots

DISSERTATION

eingereicht von

Mag. Christian Haslinger

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät

der Universität Wien

March 23, 2001

An dieser Stelle möchte ich mich herzlich bei all jenen bedanken, die zum Entstehen der vorliegenden Arbeit beigetragen habe.
Prof. Peter Schuster bot mir die Möglichkeit zur Dissertation und war immer ein sehr kompetenter und unkomplizierter Betreuer.
Allen Kolleginnen und Kollegen vom Institut danke ich für die Hilfsbereitschaft und die angenehme Arbeitsatmosphäre.
Meine Eltern und Großeltern ermöglichten mir dieses Studium und unterstützten mich wo sie nur konnten.
Zum Schluss danke ich meiner Freundin Barbara, die stets für das emotionelle Wohlergehen des Doktoranden sorgte.

Abstract

RNA pseudoknots represent a particularly interesting structural motif in functional RNA molecules. They can be involved in translational and replicational control, or are necessary to form reaction centers of certain ribozymes. The conventional definition of RNA secondary structures excludes pseudoknots explicitly, mainly for computational reasons. In this thesis the RNA secondary structure definition was extended to include the so called h-type pseudoknots. H-type pseudoknots are the most abundant of all known pseudoknots, furthermore they are the only type of pseudoknot for which an energy model exists. Consequently, the structural diversity of h-pseudoknots is restricted to the domain of the energy model. Most of these restrictions are stereo-chemically motivated and comprise almost all experimentally known h-pseudoknots.

As a result, dynamic programming becomes feasible in terms of time and memory demand. A minimum free energy folding algorithm was implemented which requires $\mathcal{O}(mn^3)$ time and $\mathcal{O}(mn^2)$ memory, where n is the length of the molecule and m is a constant depending on the structural freedom approved to the pseudoknot. Additionally the backtracking process was adapted as well, to calculate all suboptimal structures within a given energy range above the minimum free energy. However, kinetically determined phenomena are not directly accessible with this approach but need sufficient knowledge about the RNA energy landscape. Therefore an existing high resolution kinetic folding algorithm was adopted which uses an elementary move set for the inter conversion of RNA secondary structures, consisting of the closing and opening of a single base pair. The pseudoknot-specific part of this algorithm produces a list of neighbors for any given structure, according to this move set. Together with the suboptimal folding algorithm, the neighbor generating function is a prerequisite for another algorithm that provides detailed information about the energy landscape, such as local minima and folding pathways. The combination of all these methods provides a comprehensive tool to study the implications of pseudoknot formation on the energy landscape.

RNA conformational switches are an ideal opportunity to apply the com-

plete set of algorithms developed. They consist of two competing secondary structures that show mutually exclusive base pair patterns but nearly equal free energy. With the help of pseudoknots, the primarily high energy barrier between the two conformations can be lowered significantly by providing alternative folding paths.

Zusammenfassung

RNA Pseudoknoten stellen ein besonders interessantes Strukturmotiv funktioneller RNA Moleküle dar. Sie können an der Translations- und Replikationskontrolle beteiligt sein oder sind notwendig, um Reaktionszentren von Ribozymen zu bilden. Die konventionelle Definition der RNA Sekundärstruktur schließt Pseudoknoten explizit aus. In dieser Arbeit erweitern wir den Begriff der RNA-Sekundärstruktur um die sogenannten h-typ Pseudoknoten. H-typ Pseudoknoten sind bei weitem die häufigsten Pseudoknoten und zudem die einzigen für die ein Energie-Modell existiert. Daher wurde die strukturelle Vielfalt der h-typ Pseudoknoten auf den Wertebereich des Energie-Modells eingeschränkt. Die meisten dieser Beschränkungen sind stereochemisch motiviert, umfassen dennoch fast alle experimentell bekannten h-typ Pseudoknoten.

Daraus resultierend ergeben sich "Dynamic Programming" Algorithmen mit moderater Zeit- und Speicheranforderung. Es wurde ein Algorithmus zur Vorhersage der stabilsten Sekundärstruktur implementiert der $\mathcal{O}(mn^3)$ Zeit- und $\mathcal{O}(mn^2)$ Speicherabhängigkeit aufweist, wobei n die Kettenlänge ist und m eine Konstante die von der Einschränkungen der strukturellen Vielfalt der Pseudoknoten abhängt. Weiters wurde der "Backtracking"- Prozess dieses Algorithmus modifiziert, um zusätzlich alle suboptimalen Sekundärstrukturen zu generieren, die innerhalb eines bestimmten Energiebandes oberhalb der minimalen freien Energie liegen.

Dennoch sind kinetisch bestimmte Phänomene mit diesem Ansatz nicht direkt zugänglich, sondern erfordern ein ausreichendes Wissen über die Energielandschaft einer RNA Sequenz. Deshalb wurde ein hochauflösender kinetischer Faltungsalgorithmus adaptiert. Dieser verwendet elementare Transformationen, um Sekundärstrukturen ineinander umzuwandeln, wie zum Beispiel das Öffnen und Schließen eines Basenpaares. Der pseudoknoten-spezifische Teil dieses Algorithmus produziert, entsprechend dieser Transformationen, für jede gegebene Sekundärstruktur eine Liste von Nachbarn. Zusammen mit den "Dynamic Programming" Algorithmen ist diese Nachbar-generierende Funk-

tion eine Voraussetzung für einen weiteren Algorithmus der detailierte Informationen über die Energielandschaft gibt, wie zum Beispiel lokale Minima und Faltungspfade. Die Kombination aller Methoden ergibt ein umfassendes Werkzeug um zu untersuchen, wie Pseudoknoten die Energielandschaft verändern.

Ein idealer Testfall um alle entwickelten Algorithmen anzuwenden, sind RNA-Konformationsschalter. Diese bestehen aus zwei konkurrierenden Sekundärstrukturen die sich gegenseitig ausschließende Basenpaar-Muster aufweisen, aber fast die gleiche freie Energie haben. Es zeigt sich, dass Pseudoknoten alternative Faltungswege ermöglichen, welche die ursprünglich hohe Energiebarriere zwischen den beiden Konformeren erheblich erniedrigen.

Contents

List of Figures

1 Introduction

1.1 General Context

During the cause of evolution nature has developed the principle of hierarchical organization. In simple words, a confined set of smaller units is build together to give a bigger unit. The resulting set of different bigger units might also be confined and subsequently used as units to construct the next level of organization. Every level shows novel and often unpredictable properties. At the lowest level, organic molecules (monomers) are linked together by covalent bonds to yield biopolymers. Biopolymers might be linear or branched. While the most abundant biopolymers in the bio-sphere are derived from carbohydrates (cellulose), the most essential are nucleic acids (DNA, RNA) and proteins. All three of them are linear biopolymers, which is not a coincidence because DNA is linear and there is a linear transformation from DNA to RNA and the protein. This cascade is actually at the core of every living cell. The DNAs task is to store genetic information (genotype), the proteins task is to build and maintain the organism (phenotype). RNA molecules are all-rounder, they can store genetic information (RNA-viruses), they are inevitable and versatile intermediates between DNA and proteins (mRNA) and last but not least they can act as enzymes (ribozymes) [6, 19] . According to Spiegelman [59] the phenotype of an RNA molecule can be defined as its spatial structure, therefore RNA is capable to carry genotype and phenotype in the same molecule. This multi-functionality supports the idea that and RNA-world stood at an early stage of life [17, 26–29].

The function of a biomolecule is closely linked to its spatial structure, thus even in very different species, molecules with the same function exhibit very similar structure. In fact this assumption justifies all types of sequence comparison and alignment, including the so called “phylogenetic” folding (chapter

3) which gives very reliable structure predictions.
Obviously, when we speak about the spatial structure of a linear biopolymer, we also have to consider the folding process that converts the one-dimensional string into a three-dimensional structure. The process of folding is guided by the molecules tendency to adopt the thermodynamically most stable structure. In a cellular environment an RNA molecule tries to minimize the interactions between the hydrophobic planar bases and the polar solvent. It does so by folding back on itself to form pairs of bases which in turn are stacked on top of each other. The pattern of base pairs is called secondary structure. To be more precise, the fact that the RNA backbone is negatively charged, spoils the tendency to form compact structures. Systems exhibiting such behavior are considered to be energetically "frustrated", in a sense that not all favorable interactions can be satisfied simultaneously. Fortunately the secondary structure accounts for most of the stabilizing interactions, but it has to be mentioned that also not paired bases can stack as well as they can interact with already formed stacks

1.2 Secondary Structures with Pseudoknots

Since not all types of bases can form base pairs, the secondary structure strongly depends on the sequence of bases. There are four types of bases A (adenine), U (uracil), G (guanine), and C (cytosine). The hydrogen bonds which mediate the pairing may be formed between the complementary Watson-Crick pairs A-U, G-C and the slightly less stable G-U wobble pair. Secondary structures are a useful coarse graining of the spatial structure for several reasons:

- they cover most of the free energy of folding
- the secondary structure is conserved during evolution and has been used successfully to interpret RNA function

- they are computationally easy to handle because only discrete base pair patterns (no coordinates) are considered

The common theoretical secondary structure model comprises only a very small subset of all possible base pair patterns. The model excludes by definition all overlapping base pair interactions, subsequently called pseudoknots, mainly for computational reasons. It turns out that algorithms dealing with simple secondary structures can be implemented in a very elegant way, with the help of a method called dynamic programming. In fact, dynamic programming was for a long time considered incompatible with pseudoknots, at least until Rivas and Eddy [51] published an algorithm capable to handle certain types of pseudoknots. However, this dynamic algorithm exhibits a rather prohibitive resource demand ($\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ memory) and neglects stereo-chemical constraints. The used energy model is very dynamic programming friendly but is not known to reflect the real situation. In [51] an estimate of 130-140 bases is mentioned as a rough upper limit for this algorithm.

1.3 Objectives of this Work

Nevertheless it would be desirable to handle longer sequence even at the expense of structural diversity. It is the purpose of this work to give reasonable restrictions for the huge diversity of possible pseudoknots. We mainly concentrate on the so called *hairpin- or h-type pseudoknot*, for which we have at least an approximate energy model [20], and give an outline how to include more complex pseudoknots as well. This limitation allows us to manage sequences of several thousand nucleotides in length within a reasonable time scale by means of dynamic programming. To investigate how pseudoknots change the folding process, it is necessary to include the pseudoknots in all algorithms used to that end, such as the suboptimal folding, the kinetic folding and the barrier algorithm. Suboptimal folding is concerned with calculating all structures within a given energy range above the most stable structure that a specific

sequence can adopt. It is a prerequisite for an algorithm which produces the barrier tree. The barrier tree gives a notion of the folding process including misfolded states, the most stable structure and folding pathways. The kinetic folding algorithm simulates the progression of folding in time and is used to measure folding or re-folding speed.

1.4 Functional Aspects of RNA Pseudoknots

Recent work has indicated that pseudoknots are only marginally more stable than simple secondary structures (although thermodynamic data in this area are still scarce [37, 48]). This observation suggests a role for pseudoknots as conformational switches or control elements in several biological functions [56]. In molecules that lack an overall three-dimensional fold, pseudoknots fold locally and their positions along the sequence reflect their function [36]. For example, pseudoknots that are folded at the 5'-end of mRNAs tend to be involved in translational control whereas those at the 3'-end maintain signals for replication. In molecules with catalytic activities, pseudoknots are located at the core of the tertiary fold and involve nucleotides that are far apart in the sequence (RNAseP). The diversity of molecular biological functions performed by pseudoknots can be subdivided into tree different groups:

(1) **Translational control:** 5'-end pseudoknots appear to adopt two roles in the control of mRNA translation: either specific recognition of a pseudoknot by some protein is required for control, as described for the 5'-end of mRNAs in some procaryotic systems [43, 56]; or, the presence of a folded pseudoknot is necessary with no requirements on the nucleotide sequence [4, 7, 62]. In several viruses, the expression of replicase is controlled either by *ribosomal frame shifting* [4, 7, 8, 10, 62] or by *in-frame read-through* of stop codons [69]. In both cases, pseudoknot formation is necessary [4, 10, 62]. The requirements appear, however, more strict for read-through than for frame shifting. Nevertheless, the correct position of the pseudoknot in the 3' direction with respect to the slip site in riboso-

mal frame shifting, and with respect to the AUG codon in read-through is an absolute requirement [4, 69]. The presence of three pseudoknots in 16S rRNA has been suggested on the basis of comparative sequence analyzes [44]. In general these pseudoknots are assumed to show strong interactions with ribosomal proteins. One pseudoknot is known to be important for the binding of tRNA to the ribosomal A site [40, 70], and was shown to be essential for ribosomal function [47]. These observations are particularly interesting in view of the suggested conformational switch that involves the other two pseudoknots.

(2) **Core pseudoknots:** are necessary to form the reaction center of ribozymes. Most of the enzymatic RNAs with core pseudoknots are involved in cleavage or self-cleavage reactions [5,14,21,39].(see **figure 1.1**)

(3) **3'-end pseudoknots:** replication control is the common function of tRNA-like motifs at the 3'-end of several groups of plant viral RNA genomes [36]. This structural similarity is paralleled in biological function as the tRNA-like motifs are recognized by many tRNA-specific enzymes such as aminoacyl-tRNA synthetases, nucleotidyl transferase, or RNAseP [36]. The tRNA-like structure has been shown to be necessary for the initiation of replication [36]. A telomeric function of the tRNA-like structure was also demonstrated [50] , in agreement with the genomic tag model associated with such 3'-terminal tRNA-like motifs [68]. Recently, the stretch of three pseudoknots preceding the tRNA-like structure in tobacco mosaic virus was shown to act as the functional equivalent of a poly(A) tail, stabilizing a reporter mRNA and increasing gene expression up to 100-fold [16].

1.5 Organisation of the Thesis

The first chapter introduces the concept of RNA secondary structures and its structural diversity. Subsequently h-type pseudoknots are characterized as a

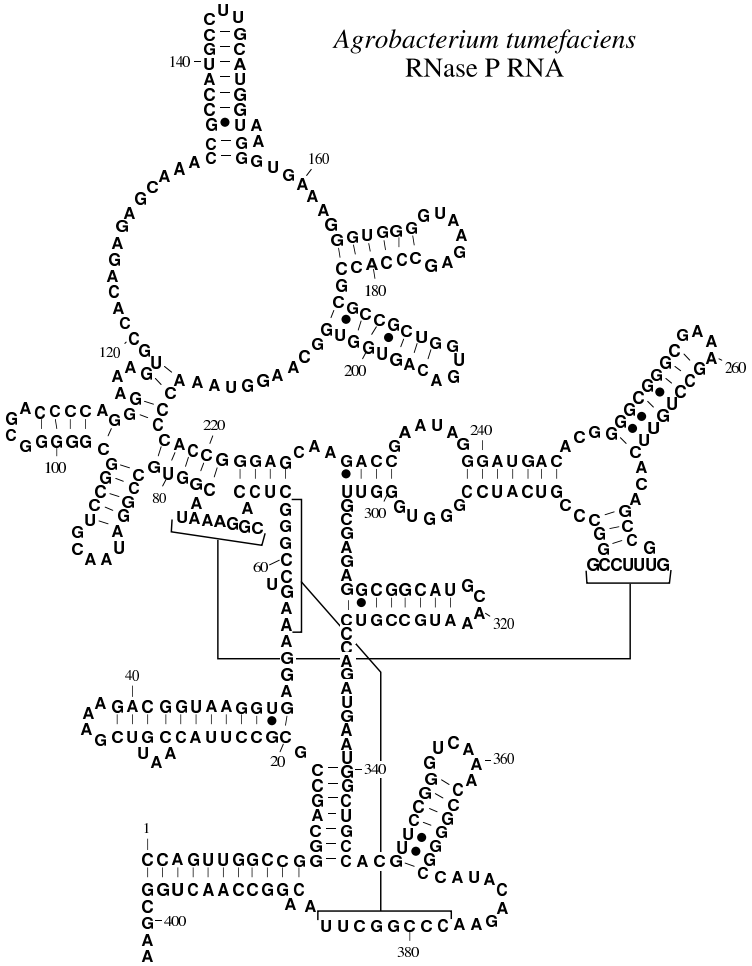


Figure 1.1: RNAseP RNA is a well studied molecule which is found in all cells that carry out tRNA synthesis. It is a processing endonuclease that specifically cleaves precursors of tRNA. In bacteria it is associated with a small protein but is clearly the catalyst. It acts as a true enzyme, in the sense that it reacts with multiple substrates.

strong restriction of the huge diversity of possible pseudoknot interactions. Furthermore, stereo-chemical considerations, which are obligatory when we allow pseudoknots, are given and lead to the energy model. The energy model concludes the chapter and explains in detail how to assign an energy to a base pair pattern and which approximations are applied to deal with pseudoknots.

In chapter 3 we show how to solve several counting and optimization prob-

lems, regarding pseudoknots, by means of dynamic programming. We start with basic enumerations, proceed to maximum matching and finally get to thermodynamic folding. The calculation of suboptimal structures is outlined for the maximum matching as well as for the thermodynamic folding problem.

Chapter 4 discusses a high resolution kinetic folding algorithm that is used to simulate the process of folding. The existing algorithm for secondary structures is extended to handle h-pseudoknots, additionally the canonic move-set is introduced, which facilitates the folding simulation.

In chapter 5 we combine all pseudoknot algorithms and give examples how the consideration of h-pseudoknots changes the RNA folding landscape. The so called barrier trees are used to give a notion about the RNA folding landscape.

2 Basics

2.1 Basic Definitions

Definition 1 A [66] *secondary structure* S is a vertex-labeled graph on n vertices with an adjacency matrix A fulfilling

1. $a_{i,i+1} = 1$ for $1 \leq i < n$
2. For each i there is at most a single $k \neq i-1, i+1$ such that $a_{i,k} = 1$
3. If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$ then $i < l < j$.

We will call an edge $(i,j), |i-j| \neq 1$ a bond or base pair. A vertex i connected only to $i-1$ and $i+1$ will be called unpaired.
A vertex i is said to be *interior* to the base pair (k,l) if $k < i < l$. If, in addition, there is no base pair (p,q) $k < p < q < l$ such that $p < i < q$ we will say that i is *immediately interior* to the base pair (k,l). A base pair (p,q) is said to be (immediately) interior if p and q are (immediately) interior to (k,l).

Secondary Structure of a Sequence
Of course a given RNA-sequence cannot form all secondary structures, since not all nucleotides form base pairs.

Definition 2 Let A be some finite alphabet of size κ, let Π be a symmetric Boolean $\kappa \times \kappa$-matrix and let $\Sigma = [\sigma_1 ... \sigma_n]$ be a string of length n over A. A secondary structure is *compatible* with the sequence Σ if $\Pi_{\sigma_p, \sigma_q} = 1$ for all base pairs (p,q).
Consequently the set of edges consists of two disjoint subsets. One (**definition 1.1**) is common to every secondary structure and represents the covalent

backbone connections while the second set (**definition 1.2**) is sequence specific and represents hydrogen bonds between the bases.

Definition 3 A secondary structure consists of the following structure elements

1. A *stack* consists of subsequent base pairs $(p-k, q+k), (p-k+1, q+k-1), ..., (p, q)$ such that neither $(p-k-1, q+k+1)$ nor $(p+1, q-1)$ is a base pair. $(k+1)$ is the *length* of the stack, $(p-k, q+k)$ is the terminal base pair of the stack. Isolated single base pairs are considered as stacks $(length = 1)$ as well.

2. A *loop* consists of all unpaired vertices which are immediately interior to some base pair (p, q), the "closing" pair of the loop. The number of this vertices is called the size of the loop.

3. An *external vertex* is an unpaired vertex which does not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 1 or n it is a *free end*, otherwise it is called *joint*.

Definition 4 A stack $[(p, q), ..., (p+k, q-k)]$ is called *terminal* if $p-1 = 0$ or $q+1 = n+1$ or if the two vertices $p-1$ and $q+1$ are not interior to any base pair. The sub-structure enclosed by the terminal base pair (p, q) of a terminal stack will be called a *component of* S.

Definition 5 The degree of a loop is given by 1 plus the number of terminal base pairs of stacks which are interior to the closing bond of the loop. A loop of degree 1 is called *hairpin (loop)*, a loop of a degree larger than 2 is called *multi-loop*. A loop of degree 2 is called *bulge* if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of degree 2 is termed *interior loop*. Two stacked base pairs form an interior

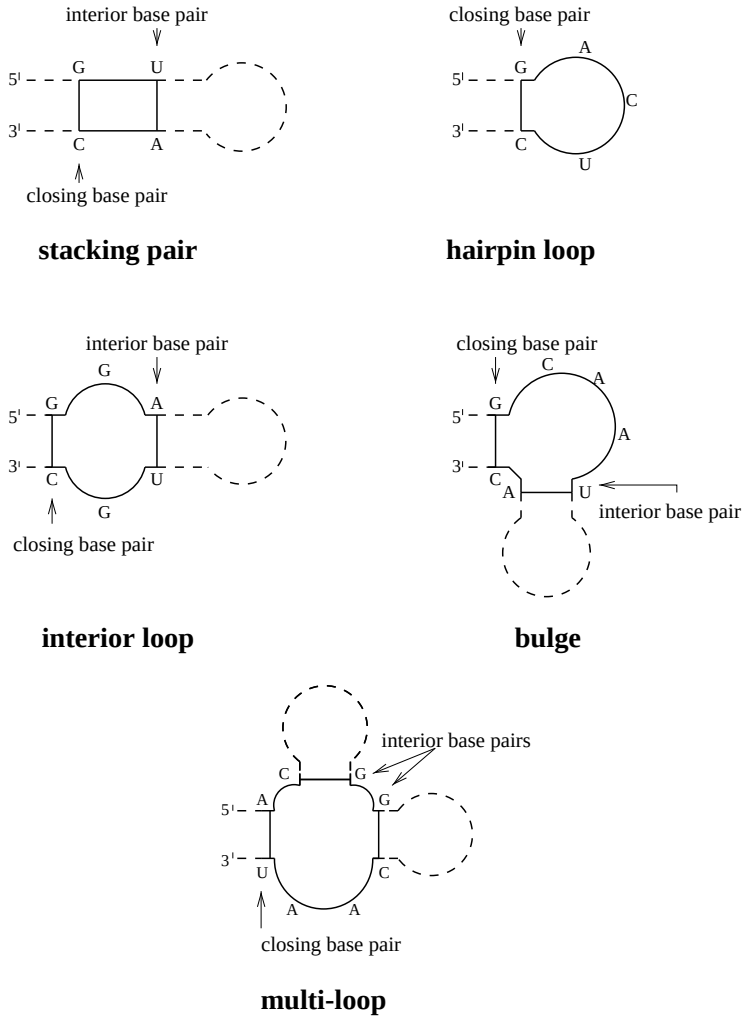


Figure 2.1: Basic loop types

loop with size 0.

.

Representation of Secondary Structures

Up until now we used a graph-representation directly derived from the definition of secondary structures. A string representation $\mathbf{S}$ can be obtained by the following rules:

(1) If vertex i is unpaired then $\mathbf{S}_i =$"."

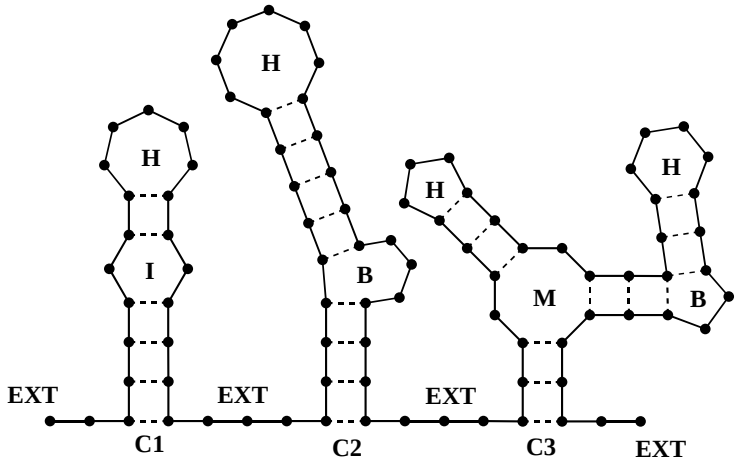


Figure 2.2: Labeled secondary structure: All loops except stacks are labeled. H: hairpin loops, B: bulges, I: interior loop, M: multi loop, EXT external vertices, C component

(2) If $(p, q), p < q$ is a base pair and then $\mathbf{S}_p =$"(" and $\mathbf{S}_q =$ ")"

A linked graph can be viewed as a particular way to draw secondary structure graphs by placing the sequence on a line and connecting the bases with arcs. Especially useful to compare even large structures is the *mountain*-representation (or *mountain*-plot) [24, 31]. The three symbols of the string representation '(',')' and '.' are assigned to three directions 'up','down' and 'horizontal' in the plot. The structural elements of a mountain-plot profile match certain secondary structure features.

- *Peaks* correspond to hairpins. The symmetric slopes represent the stems enclosing the unpaired bases in the hairpin loop, which appear as a plateau.

- *Plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively.

- *Valleys* indicate the unpaired regions between the branches of a multi

loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The height of the mountain at sequence position k is simply the number of base pairs that enclose position k, i.e., the number of all base pairs (i, j) for which $i < k$ and $j > k$. The mountain representation allows straightforward comparison of secondary structures and inspired a convenient algorithm for alignment of secondary structures.

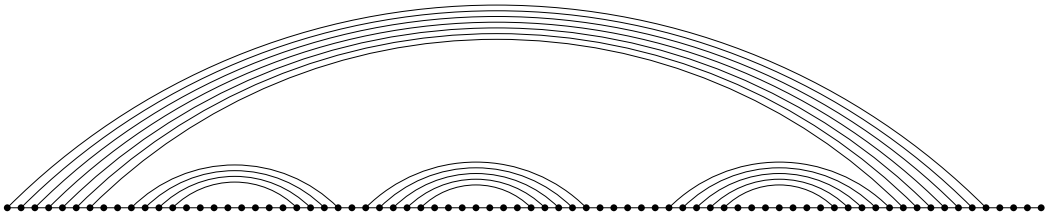


Figure 2.3: The secondary structure of tRNA^{Phe} in linked diagram representation. The same structure in string representation: `(((((((..((((........)))).(((((.......)))))).....(((((.......)))))))))))....`

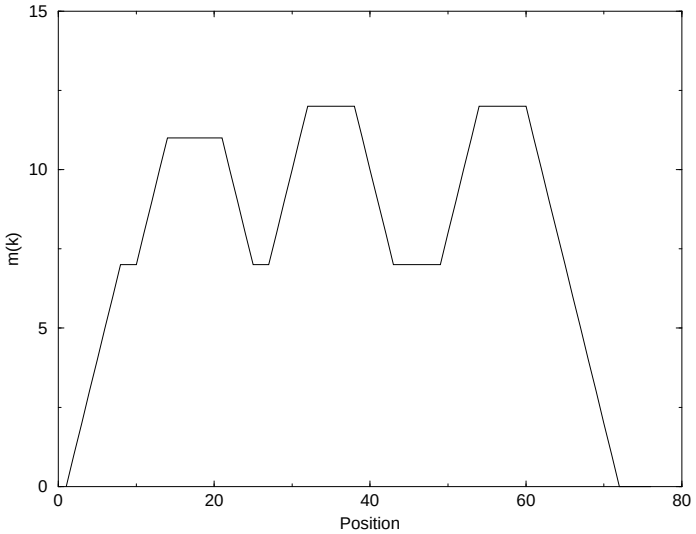


Figure 2.4: The secondary structure of tRNA^{Phe} in mountain plot representation

2.2 H-Type Pseudoknot Definitions

If we violate **definition 1.3** we produce two overlapping base pairs, the result is called *pseudoknot*. The most simple type of pseudoknot is the so called *h-type pseudoknot* or *h-pseudoknot*.

Representation of Secondary Structures with H-type Pseudoknots Some useful representations of secondary structures are not applicable if we include h-pseudoknots in the set of allowed base pair patterns. This is due to the fact that overlapping base pairs do occur. However, we can always represent an h-pseudoknot as a planar graph or as a linked graph.

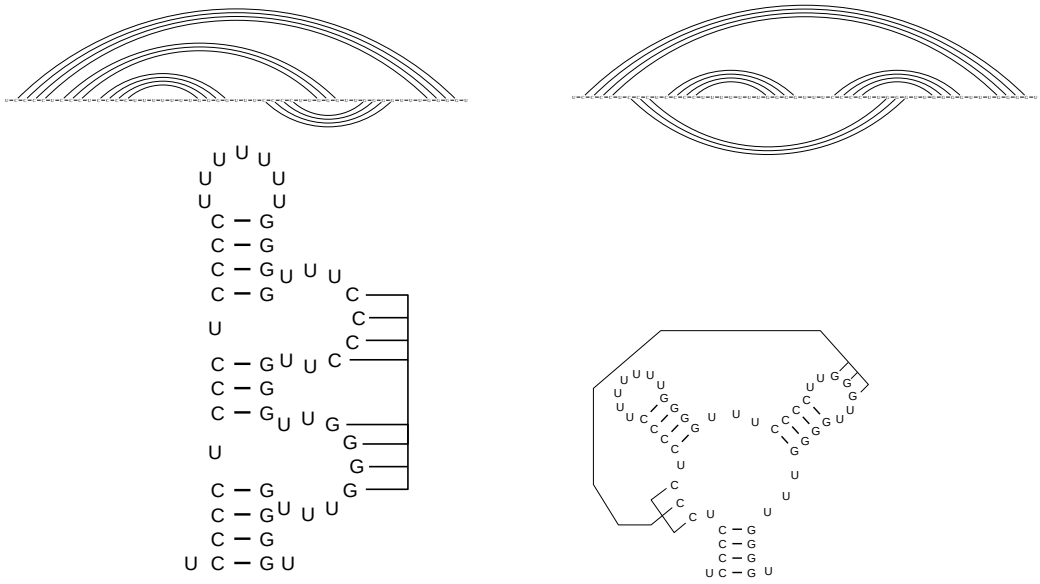


Figure 2.5: Two alternative representations of the same base pair pattern

The consideration of overlapping base pair interactions gives an enormous enlargement of the structure space. The number of possible structures grows faster than exponentially.

We are not able to deal with the resulting complexity by computational means. Additionally a lot of them can never be realized by an RNA sequence (e.g. parallel β-sheets). Thus the term "pseudoknot" gives not at all a sufficient definition to what extend the structure space should be enlarged. Simply calling all structures containing an overlapping base pair a pseudoknot is not of great help. Therefore we start with the most elementary pseudoknot and discuss the more complex types later.

Definition 1 A *building block* $B_{i,k,l,j}, i \leq k < l \leq j$ is a secondary structure on the interval $[i, j]$ with a gap at $[k+1, l-1]$. All loops of this structure, except (k, l), have degree 2, therefore a building block can also be viewed as a set of base pairs fulfilling the condition:

$$B_{i,k,l,j} = \{(p, q) | i \leq p \leq k \text{ and } l \leq q \leq j\} \tag{2.1}$$

Note that, because of its gap, a building block alone is not a valid secondary structure.

Definition 2 Two building blocks $B_{i,k,l,j}$ and $B_{i',k',l',j'}$ are *h-type generating* if their associated intervals are disjoint:

$$\Big\{[i, k] \cup [l, j]\Big\} \cap \Big\{[i', k'] \cup [l', j']\Big\} = \emptyset \tag{2.2}$$

and arranged in an alternating way

$$i < i' : [i, k].[i', k'].[l, j].[l', j'] \tag{2.3}$$

$$i' < i : [i', k'].[i, k].[l', j'].[l, j] \tag{2.4}$$

Definition 3 An *h-pseudoknot* $Pk_{i,j'}$ is obtained when we merge two h-type generating building blocks. We can distinguish an upstream $B^u_{i,k,l,j}$ and a downstream $B^d_{i',k',l',j'}$ building block:

$$PK_{i,j'} = B^u_{i,k,l,j} \cup B^d_{i',k',l',j'} \tag{2.5}$$

When $(k,l),(k',l')$ and $(i,j),(i',j')$ are base pairs we produce three unpaired regions:

$$L1 = \{n | k < n < i'\} \tag{2.6}$$

$$L3 = \{n | k' < n < l\} \tag{2.7}$$

$$L2 = \{n | j < n < l'\} \tag{2.8}$$

All vertices $L1$ are immediately interior to (k,l), vertices $L3$ are immediately interior to (k',l'). All vertices $L2$ are immediately interior to both (k,l) and (k',l'). These three regions are in the literature referred as loops which does not meet our definition of loops (**definition 3.2**) because it is not possible to uniquely assign the vertices $L2$ to just one loop.
With the definition given above we can easily extend the string notation $\mathbf{S}$ to fit our requirements:

1. If vertex i is unpaired then $\mathbf{S}_i =$"."
2. If $(p,q), p < q$ is a base pair and $(p,q) \in M'$ then $\mathbf{S}_p =$"(" and $\mathbf{S}_q =$")"
3. If $(p,q), p < q$ is a base pair and $(p,q) \in PK_{i_t,j_t}, 1 \leq t \leq h$ and

$$\begin{array}{llllll} (p,q) & \in & B^{u;t}_{i,k,l,j} & : & \mathbf{S}_p = & \text{"(" and } \quad \mathbf{S}_q = \text{")"} \\ (p,q) & \in & B^{d;t}_{i',k',l',j'} & : & \mathbf{S}_p = & \text{"[" and } \quad \mathbf{S}_q = \text{"]"} \end{array}$$

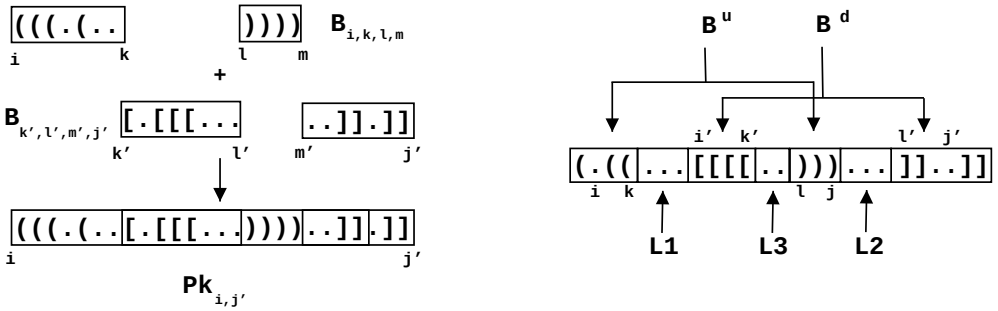


Figure 2.6: Two h-type generating building blocks merged to result a h-pseudoknot. The right picture denotes the three unpaired regions $L1, L2$ and $L3$.

Definition 4 A set of base pairs M is called an RNA secondary structure with h-pseudoknots if the following conditions are satisfied:

1. $M = M' \cup Pk_{i_1,j_1} \cup Pk_{i_2,j_2} ... \cup Pk_{i_t,j_t}$ where t is a non-negative integer and $1 \leq i_1 < j_1 < i_2 < j_2 < ... < i_t < j_t \leq n$.

2. Each Pk_{i_h,j_h} is a h-pseudoknot

3. M' is a secondary structure without pseudoknots for a sequence Σ', where Σ' is obtained by deleting all $\sigma_{i_h}, \sigma_{i_h+1} \cdots \sigma_{j_h}$ from Σ (i.e., $\Sigma' = \sigma_1\sigma_2 \cdots \sigma_{i_1-1}\sigma_{j_1+1} \cdots \sigma_{i_2-1}\sigma_{j_2+1} \cdots \sigma_{i_t-1}\sigma_{j_t+1} \cdots \sigma_n$).

4. For each $(i,j) \in Pk_{i_h,j_h}, 1 \leq h \leq t$ there is no $(i',j') \in M'$ which satisfies $i' < i < j < j'$, i.e. there is no base pair in Pk_{i_h,j_h} which is interior to a base pair in M'.

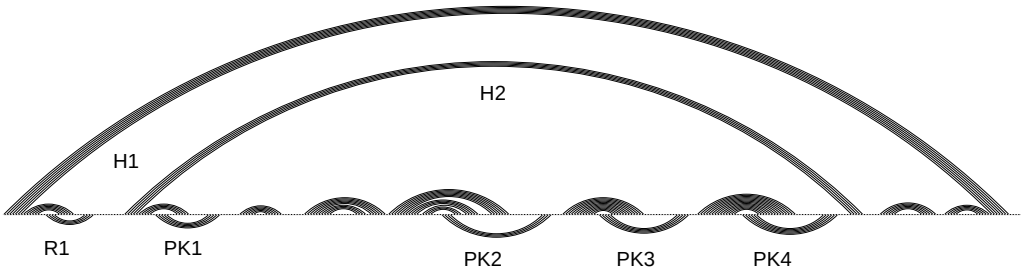


Figure 2.7: tmRNA in linked graph representation - a structure that violates **definition 4.4**. The substructures R1, PK1, PK2, PK3, and PK4 are h-pseudoknots; the stems H1 and H2 enclose pseudoknots, hence produce 'pseudoknot multi loops'.

Definition 5 A *simple h-pseudoknot* is an h-pseudoknot where $B^u_{i,k,l,j}$ and $B^d_{i',k',l',j'}$ are stacks. Because of the symmetry of this building blocks we can use a reduced notation where $ST_{i,j,S}$ denotes a stack with the terminal base pair (i,j) and stack size S.

$$PK_{i,j'} = ST^u_{i,j,S} \cup ST^d_{i',j',S} \tag{2.9}$$

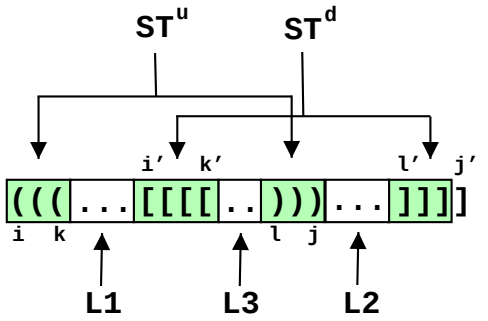


Figure 2.8: Simple h-pseudoknot

The following stereo-chemical considerations are illustrated on simple h-pseudoknots. The upstream stack is denoted $S1$, the downstream $S2$ to be consistent with the literature.

2.3 Stereo-chemical Considerations

One of the main reasons why secondary structures are popular among theoreticians is the fact that they are discrete and therefore computationally easy to handle. If we stick to secondary structures without pseudoknots we do not need to care about coordinates, space exclusion and things like that. We will always get base pair patterns which give stereo-chemically feasible molecules. The only requirement is a minimum loop size for hairpin loops which can be met easily.

The picture changes when we consider h-pseudoknots. The two stacked $(S1, S2)$ regions have to be bridged somehow and therefore the stereo-chemical situation gets more complicated. In addition thermodynamic considerations require the two stacks to be *coaxially stacked*. This imposes another stereo-chemical constrain.

Theoretically the two stacks $(S1, S2)$ can be coaxially stacked in four different ways:

(1) $L1$ bridges $S2$ and $L2$ bridges $S1$

(2) $L3$ bridges $S2$, $L2$ bridges $S2$ and $S1$

(3) $L3$ bridges $S1$, $L1$ bridges $S2$ and $S1$

(4) $L1$ bridges $S1$, $L3$ bridges $S2$ and $S1$, $L2$ bridges $S2$

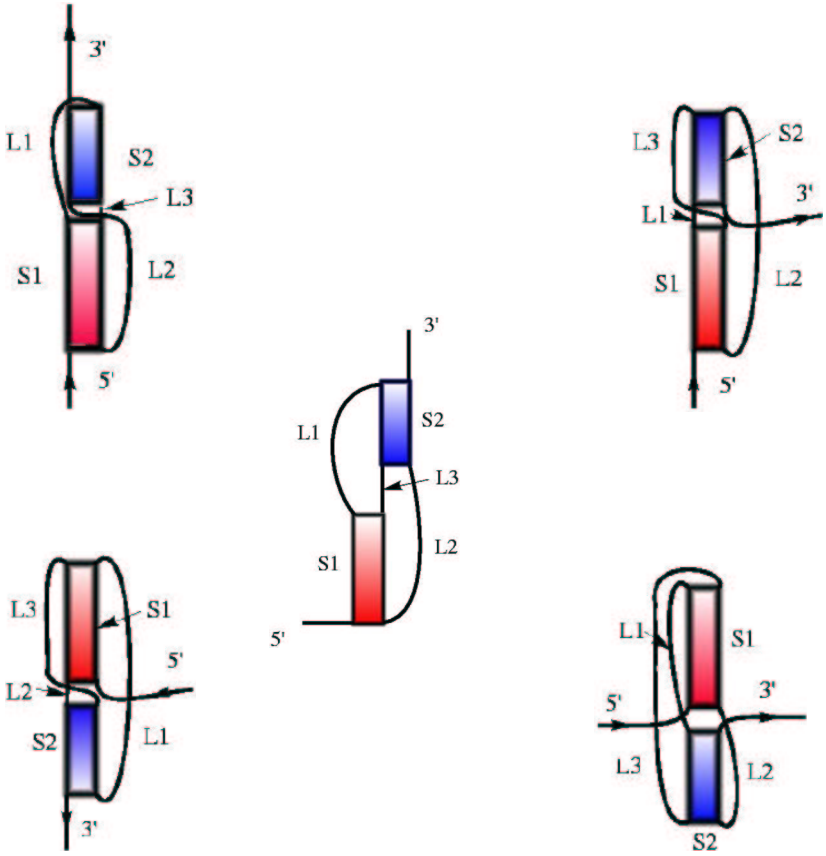


Figure 2.9: The four types of coaxial stacking. In the center of the figure no coaxial stacking is indicated. All described types of coaxial stacking are displayed.

In the first three cases the two stacking interfaces are directly connected by the RNA backbone. Efficient coaxial stacking is only assumed if the number of bases in the connecting region is 0 or 1. An intervening base between the to

stacks is supposed to from a base pair mismatch. The mismatch partner can be provided by one of the two remaining unpaired regions (**figure 2.10**).
Obviously the first stacking variant gives the shortest loops and therefore least destabilizing entropy contribution to the free energy of the pseudoknot. Additionally all other types of coaxial stacking interactions are disturbed by the entering and/or exiting parts of the RNA-chain.

Definition 1 The type of coaxial stacking in an h-pseudoknot can be denoted according to the location of the 5' and 3' end of the h-pseudoknot relative to the stacking interface.

(1) $h_{3'}$-pseudoknot - 3' end at the stacking interface

(2) h_0-pseudoknot - no end at the stacking interface

(3) $h_{5'}$-pseudoknot - 5' end at the stacking interface

(4) $h_{5',3'}$-pseudoknot - 3' and 5' end at the stacking interface

As pointed out above the h_0-pseudoknot is the most plausible type of coaxial stacking. This assumption is also confirmed when we take a look at experimentally proven base pair patterns of h-pseudoknots [63](see appendix A). Only 5 out of 148 h-pseudoknots show $|L3| > 1$.
The fact that $L1$ and $L2$ have to bridge $S2$ and $S1$ brings geometrical dimension in the concept of secondary structures. For a given stack-length we need a minimum loop-length to bridge the stack:

$L_{1min}(S_2)$ minimum loop size of L_1 to bridge S_2
$L_{2min}(S_1)$ minimum loop size of L_2 to bridge S_1

Definition 2 Two h_0-generating building blocks are *stereo-compatible* if the conditions:

$$L_{1min}(S_2) \leq L_1 \tag{2.10}$$

$$L_{2min}(S_1) \leq L_1 \tag{2.11}$$

are fulfilled.

This requirement can either be introduced in the definition of allowed h_0-pseudoknots or we can include it in the set of thermodynamic energy parameters (by assigning infinite free energy to loops which are to short). We prefer the first method because we want to give the recursions for the enumeration and the maximum matching of secondary structures with pseudoknots (both have nothing to do with thermodynamic energy parameters).

In **definition 4.4** we excluded the pseudoknots from being internal to any base pair. This restriction is also stereo-chemically motivated because otherwise we would have to bridge the whole pseudoknot with a base pair containing loop. While this case seems to be manageable, the situation get increasingly complicated when several pseudoknots and secondary structure components are enclosed.

It is our objective to find definitions for base pair patters that fit as close as possible to experimentally known pseudoknots. A closer look at the pseudoknot-database [63] (appendix) tells us, that on the one hand a general h-pseudoknot gives by far to much structural freedom and on the other hand the concept of simple h_0-pseudoknots is to narrow. Therefore we give another definition.

Definition 3 A *restricted h-pseudoknot* has a maximum number of bulges and interior loops per building block. The maximum size and asymmetry of each interior loop may also be restricted. The concept of restricted h-pseudoknots is elaborated in more detail in **chapter 3**, where we give recursive definitions.

2.4 Energy Model

The energy model is the most essential part of every energy derived structure prediction algorithm. In other words, any result produced depends at first on the quality of the underlying energy function. And the reliability of the energy model depends on the quality of the empirical thermodynamic data. Unfortunately empirical energy parameters of sufficient accuracy are not available for pseudoknots. Moreover the general type of the energy model determines the architecture of the dynamic programming recursions. In fact the energy function has to meet some properties to allow efficient dynamic programming. In the upcoming section we describe how an energy is assigned to an RNA secondary structure.

2.4.1 Thermodynamic Energy Model

The results of both quantum chemical calculations and thermodynamic measurements suggest that horizontal (base pairing) contributions to the total energy depend exclusively on the base pair composition, whereas vertical (base stacking) contributions depend on base pair composition *and* base sequence i.e. the upstream and downstream neighbors along the chain [54]. The *nearest neighbor model* introduces the assumption that the stability of a base pair, or any other structural element of an RNA, is dependent only on the identity of the adjacent bases and/or base pairs. The model is justified by the major contribution of short-range interactions (hydrogen bonding, base stacking) to the overall stabilizing energy of nucleic acid structures. In addition, it is natural to assign loop entropies to entire loops instead of individual bases. Treating stacks as special types of loops, one assumes therefore that the energy of an RNA secondary structure Φ is given by the sum of energy contributions ϵ of it's loops L.

$$E(\Phi) = \sum_{L \in \Phi} \epsilon(L) + \epsilon(L_{ext}), \tag{2.12}$$

where L_{ext} is the contribution of the "exterior" loop containing the free ends. Note that here stacked pairs are treated as minimal loops of degree 2 and size 0.

In the following we shall discuss the individual contributions in some detail. In particular, the energy model contains the following contributions [61]:

Stacked pairs and G-U mismatches contribute the major part of the energy stabilizing a structure. Surprisingly, in aqueous solution parallel stacking of base pairs is more important than hydrogen bonding of the complementary bases. By now all 21 possible combinations of A-U G-C and G-U pairs have been measured in several oligonucleotide sequences with an accuracy of a few percent. The parameters involving G-U mismatches were measured more recently in Douglas Turner's group and brought the first notable violation of the nearest-neighbor model: while all other combinations could be fitted reasonably well to the model, the energy of the $\begin{smallmatrix}5'\text{G-U}3'\\3'\text{U-G}5'\end{smallmatrix}$ stacked pair seems to vary from $+1.5\,\text{kcal/mol}$ to -1.0 kcal/mol depending on its context.

Unpaired terminal nucleotides and terminal mismatches: unpaired bases adjacent to a helix may also lower the energy of the structure through parallel stacking. In the case of free ends, the bases dangling on the $5'$ and $3'$ ends of the helix are evaluated separately, and unpaired nucleotides in multiloops are treated in the same way. For interior and hairpin loops, the so called *terminal mismatch* energy depends on the last pair of the helix and both neighboring unpaired bases. While stacking of an unpaired base at the $3'$ end can be as stabilizing as some stacked pairs, $5'$ dangling ends usually contribute little stability. Terminal mismatch energies are often similar to the sum of the two corresponding dangling ends. Typically, terminal mismatch energies are not assigned to hairpins of size three. Few measurements are available for the stacking of unpaired nucleotides on G-U pairs, and for this reason they have to be estimated from the data for G-C and A-U pairs.

Loop energies are destabilizing and modeled as purely entropic. Few experimental data are available for loops, most of these for hairpins. The parameters for loop energies are therefore particularly unreliable. Data in the newer compilation by Jaeger *et al.* [25] differ widely from the values given previously [15]. Energies depend only on the size and type (hairpin, interior or bulge) of the loop. Hairpins must have a minimal size of 3, and values for large loops

($k > 30$) are extrapolated logarithmically:

$$\mathcal{H}(k) = \mathcal{H}(30) + \text{const.} * \log(k/30) \tag{2.13}$$

Asymmetric interior loops are furthermore penalized [42], using an empirical formula depending on the difference $|u_1 - u_2|$ of unpaired bases on each side of the loop.

$$\Delta F_{\text{ninio}} = \min\left\{\Delta F_{\max}, |u_1 - u_2| * \Delta F_{\text{ninio}}\left[\min\{4.0, u_1, u_2\}\right]\right\} \tag{2.14}$$

For bulge loops of size 1, a stacking energy for the stacking of the closing and the interior pair is usually added, while larger loops are assumed to prohibit stacking. Finally, a set of eight hairpin loops of size 4 are given a bonus energy of 2 kcal/mol. These tetraloops have been found to be especially frequent in rRNA structures determined from phylogenetic analysis. Melting experiments on several tetraloops [2] show a strong sequence dependence that is not yet well reflected in the energy parameters. No measured parameters are available for multi-loops, their contribution (apart from dangling ends within the loop) being approximated by logarithmic extrapolation.

$$\Delta G = Epk + \text{const.} * \log(u + m), \tag{2.15}$$

where u is the size of the loop and m is the number of base pairs interior to the loop, i.e. its degree-1. Energy parameters for the contributions described above have been derived mostly from melting experiments on small oligonucleotides. The first compilation of such parameters was done by Salser [55]. The parameters most widely in use today are based on work of D. Turner and coworkers . The current work uses the compilation of [15, 23, 61], who performed measurements at 37°C in 1 M NaCl. More recently the differences between symmetric and asymmetric loops have been reported to be only half the magnitude suggested by Papanicolau *et al.* [42] and of higher sequence dependence Serra *et al.* found a dependence of hairpin loop energies on the closing base pair [58] and presented a model to predict the stability of hairpin loops [57]. Walter and coworkers suggested a model system for the coaxial

stacking of helices [64]. Wu and Walter studied the stability of tandem GA mismatches and found them to depend upon both sequence and adjacent base pairs [34, 65]. Ebel and coworkers measured the thermodynamic stability of RNA duplexes containing tandem G-A mismatches [53]. Morse and Draper presented thermodynamic parameters for RNA duplexes containing several mismatches flanked by C-G pairs. Mismatches are reported to have a wide range of effects on duplex stability; the nearest neighbor model is considered not to be valid for G-A mismatches [41]. These results are, however, not yet included into the parameter set used in this work.

H-pseudoknots: Unfortunately there are no measured thermodynamic parameters for pseudoknots available. Therefore we totally rely on approximations which are more complicated than those for non-pseudoknot loops. Gultyaev *et al* [20] conceived a model for h_0-pseudoknots which is based on the general theory of polymer loop thermodynamics and stereo-chemical considerations. Additionally he uses empirical support from experimentally and/or phylogenetically proven secondary structures of h-pseudoknots.
The free energy of an h-pseudoknot structure is mainly the sum of the free energies of stacking in the stems (stabilizing negative values), the coaxial stacking of the stems (if possible) and the positive destabilizing loop values. The free energy of the stems (including small bulges and interior loops within them) are calculated using the standard model described above. For the loops some assumption about h_0-pseudoknot topology is needed:

- The loop $L1$ spans the *deep grove* of RNA helix $S2$, whereas $L2$ crosses stem $S1$ in the *shallow groove*. Therefore, the loops are not equivalent stereo-chemically and their features depend on the lengths of the corresponding stems [45]. This should be taken into account by introducing two variables $A_{deep}(S2)$ and $A_{shallow}(S1)$ for the loops $L1$ and $L2$, respectively.

- The distances between phosphate atoms connected by the loops along the RNA grooves are minimal at $S2$ of 6-7 bp and at $S1$ of 3 bp [45].

This is also consistent with frequencies of natural occurrence of stem lengths [9]. Therefore, it is assumed that A_{deep} is minimal for $S2$ of 6 or 7 bp, and $A_{shallow}$ is minimal for $S1 = 3$ bp.

- Analysis of pseudoknot geometries also suggests the minimal sizes of loops possible for given stem lengths. In the deep groove, 7 bp can be bridged by a loop of 1 nt only. Bridging over the shallow groove requires at least 2 nt, and the distance to be crossed increases significantly with the length of the stem [45]. However, a bending and/or distortion of the RNA A-helical geometry is also possible, so the requirements for big stems may be less rigid. The model assumes a minimally allowed size of loops $L2$ (shallow grove) of 2 nt for an $S1$ of 3bp, 3 nt for 4 bp, 4 for an $S1$ of 5 or 6 bp, and a further increase of 1 nt for each increment of 2 bp. For the deep groove, a loop of 1 nt was allowed for stems of 4-7 bp, and loops of 2 nt for stems of 3 bp or more than 7 bp.

- Instead of just using a logarithmic increase of entropy with loop size, the dependence on the difference between the loop length and the minimally allowed length is introduced [46,60]. Such an approximation can partially reflect restrictions of conformational freedom imposed by the stem end-to-end distance.

Considering all these assumptions we get two dependences.

$$\begin{array}{rcl} \Delta G_{L1} & = & A_{deep}(S2) \quad + \quad 1.75\ RT\ ln(1 + L1 - L_{mindeep}(S2)) \\ \Delta G_{L2} & = & A_{shallow}(S1) \quad + \quad 1.75\ RT\ ln(1 + L2 - L_{minshallow}(S1)) \end{array} \tag{2.16}$$

where $L_{mindeep}$ and $L_{minshallow}$ define the shortest loops. The *coaxial stacking* is only considered for what we defined as h_0-pseudoknot. Coaxial stacking without an intervening base pair gives a free energy which is about 1 kcal/mol more stable than the corresponding stacking energy in a regular helix. In the case of an intervening base pair the mismatch values of interior loops are used. However, it remains unclear from which loop the second mismatch partner should come and if this base could still be used to bridge the corresponding stem. Moreover it is not known to what extend we can disturb the building blocks

with bulges and interior loops and still adhere to $L_{mindeep}$ and $L_{minshallow}$.

Basically the approximations for the h_0-pseudoknot should not overestimate

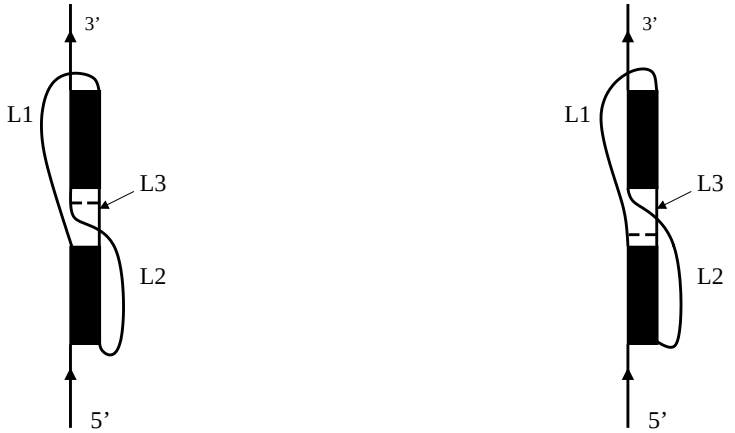


Figure 2.10: Mismatch types of intervening bases

its stability and therefore result in to high abundances of h_0-pseudoknots in an energy derived folding algorithm. At least we know from the rare thermodynamic experiments, that an h_0-pseudoknot is only marginally more stable than each of the two alternative stem-hairpin structures.

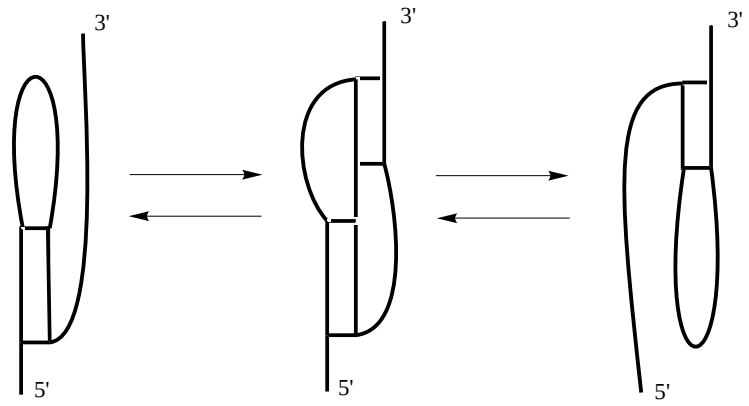


Figure 2.11: Equilibrium between an H-pseudoknot and alternative hairpins

3 Dynamic Programming

3.1 The Principle of Dynamic Programming

The method of dynamic programming is used to solve a broad class of different computational problems from compiler optimization to economics and molecular biology. The first application in molecular biology was an optimal alignment algorithm. In this work we focus our attention to dynamic programming algorithms that deal with RNA secondary structures, and how they change when we extend the set of allowed base pair patterns. Generally all dynamic programming algorithms have to pass three stages:

1 Initialization of the matrices

2 Matrix filling - scoring

3 Traceback/Backtracking

Depending on the scoring function in step two several problems with different time and space complexity can be solved for a given RNA sequence. Ordered by increasing resource demand:

- Enumeration - count the number of legal base pair patterns
  Matching - find the maximum matching i.e. maximum number of base pairs

- Free energy - find the minimum free energy according to a given energy model
  Partition function - calculate the sum of all Bolzman-weighted legal structures

- Density of states - calculates the density of secondary structures at any energy state

According to the scoring functions the corresponding backtracking yields various results:

- Enumeration - produce a random structure
- Matching
  - produce the structure with the maximum number of base pairs
  - produce all suboptimal-matching structures within a given base pair range above the maximum number
- Free energy
  - produce the mfe-structure
  - produce all suboptimal structures within a given energy range above mfe

The question arises if these algorithms are still feasible when we introduce pseudoknots to the structure space. And if they are realizable, how practicable are they in terms of time and memory demand. First of all the answer depends on the scale of additional base pair patterns, simplifying called pseudoknots, we allow. It is by no means obvious what the best formalism is to describe the set of all legal base pair pattern. A list of coupled recursions is not a very vivid representation but would be a convenient template for a computer program. A formal grammar which defines the set of all allowed strings, associated with a diagram representation called Feynman diagrams, is both more descriptive and gives reasonable access to the underlying recursions [51,52]. Graph theoretical definitions are also useful in extending the structure space. As indicated above the term "pseudoknot" gives not at all a sufficient definition to what extend the structure space should be enlarged. Simply calling all structures comprising an overlapping base pair a pseudoknot is not of great help.
For example we can we use the so called book embedding [22] to confine the complexity of structure space. A $p-$book is a set of p distinct half-planes (the *pages* of the book) that share a common boundary line l, called the spine

of the book which corresponds to the backbone of an RNA molecule. An embedding of a graph into a book B consists of an ordering of the vertices (bases) along the spine of the book together with an assignment of each edge to a page of the book, in which edges assigned to the same page do not cross. The *book-thickness* (sometimes also called the page-number) of a graph is the minimal number p of pages of a book into which it can be embedded. Thus ordinary secondary structure graphs need $p = 1$ pages, a pseudoknot at least $p = 2$ pages. An upper limit for the book-thickness consequently constricts the pseudoknot complexity. Structures with $p = 2$ pages correspond to the class of planar graphs which means that they can be drawn without crossing vertices. Almost all known structures fall into this class (one exception: α-mRNA). Unfortunately restricting structures to $p = 2$ gives nevertheless NP-hard algorithms [1] even with a very simple scoring function. Clearly we need stronger restrictions to get polynomial dependencies.

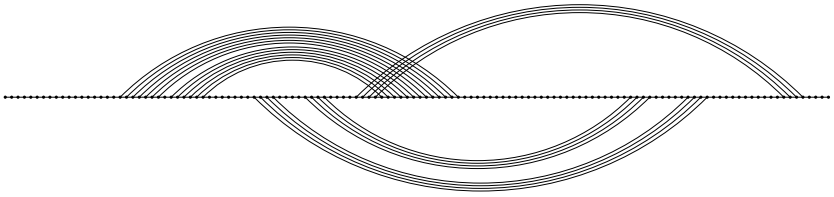


Figure 3.1: Secondary structure of αmRNA - with book-thickness $p = 3$

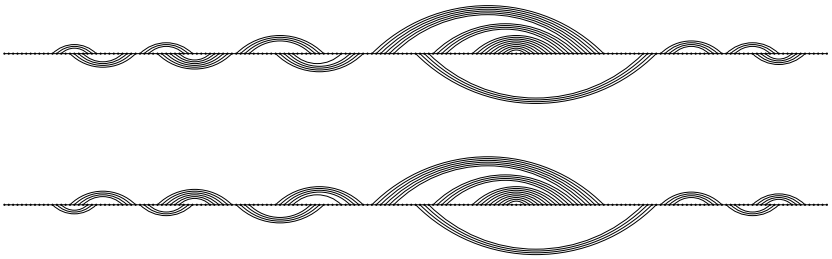


Figure 3.2: Two representations of the tmvRNA secondary structure - with book-thickness $p = 2$

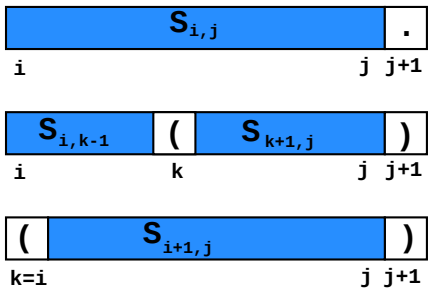
3.2 Enumerations

Enumerations are the fundament of RNA-dynamic programming. The recursions to enumerate all possible structures on a given sequence are an implicit definition of all allowed base pair patterns - they are in fact recursive definitions. Thus the design of the recursions delimits the structure space. Waterman gave the first recursion for counting secondary structures [66].

3.2.1 The Basic Recursion - Counting Without Pseudoknots

When we elongate an interval i, j by one base $i, j+1$, this base can adopt two states. Either it is paired or it is unpaired. If it is unpaired all structures $S_{i,j}$ with an additional unpaired base have to be counted. If it is paired, it may form a base pair with any upstream region within i, j (with the restriction of a minimal hairpin loop-size hp). This upstream position k divides the interval $i, j+1$ in two valid sub-structures.

$$S_{i,j+1} = S_{i,j} + \sum_{i \leq k \leq j-hp} S_{i,k-1} S_{k+1,j} \; \Pi_{\sigma_k, \sigma_{j+1}} \tag{3.1}$$



$$S_{i,j} = 1, \forall \; |i-j| < hp+2$$

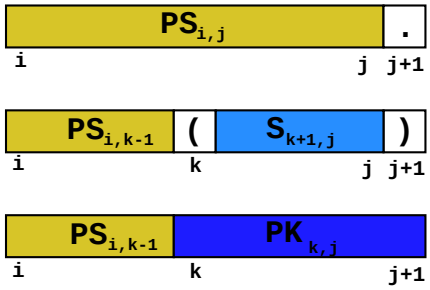
Only one matrix is needed to hold the numbers for all possible segments.

3.2.2 The General h-Pseudoknots

Definition 4.4 requires to forbid pseudoknots from being internal to any base pair, thus we need two matrices more. Let $PK_{i,j}$ be the number of h-pseudoknots and let $S_{i,j}$ denote the number of secondary structures on i, j. Then $PS_{i,j}$ gives the number of secondary structures including h-pseudoknots. pm is the minimum size of an h-pseudoknot.

$$PS_{i,j+1} = PS_{i,j} + \sum_{i \le k \le j-hp} PS_{i,k-1} S_{k+1,j} \Pi_{\sigma_k,\sigma_{j+1}} + \sum_{i \le k \le j-pm+1} PS_{i,k-1} PK_{k,j+1} \tag{3.2}$$

$$PS_{i,j} = 1, \forall\ |i-j| < hp+2$$

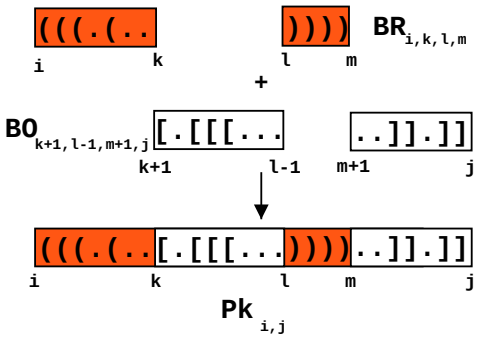


There are two recursions coupled with PS: the recursion for S is the same as above, the PK recursion is given later. It is no problem to allow pseudoknots inside a loop, actually the opposite is true because we can save one matrix.

$$PS_{i,j+1} = PS_{i,j} + \sum_{i \le k \le j-hp} PS_{i,k-1} PS_{k+1,j} \Pi_{\sigma_k,\sigma_{j+1}} + PK_{i,j+1} \tag{3.3}$$

In order to complete the set of recursions we give the general form for PK :

$$PK_{i,j} = \sum_{i<k<j} \sum_{k<l<j} \sum_{l<m<j} BR_{i,k,l,m} BO_{k+1,l-1,m+1,j} \tag{3.4}$$

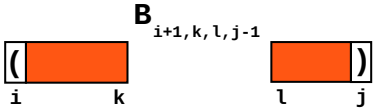


$BO_{i,k,l,m}$ is the number of possible building blocks on the intervals $[i,k]$ and $[l,m]$ with the restriction that (i,m) forms a base pair. For $BR_{i,k,l,m}$ we need the additional constrain that l is also part of a base pair.

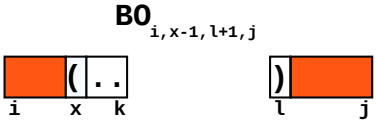
The use of $BO_{i,k,l,m}$ and $BR_{i,k,l,m}$ is a method to avoid multiple counting of the same structure. The auxiliary variable $B_{i,k,l,j}$ is the number of building blocks on the intervals $[i,k]$ and $[l,m]$.

$$B_{i,k,l,j} = B_{i,k,l,j-1} + \sum_{x=i}^{k} B_{x+1,k,l,j-1}\Pi_{\sigma_x,\sigma_j} \tag{3.5}$$

$$BO_{i,k,l,j} = B_{i+1,k,l,j-1}\Pi_{\sigma_i,\sigma_j} \tag{3.6}$$



$$BR_{i,k,l,j} = \sum_{x=i+1}^{k} BO_{i,x-1,l+1,j}\Pi_{\sigma_x,\sigma_l} \tag{3.7}$$



Clearly the coupled PK recursions determine the resource demand to calculate PS.

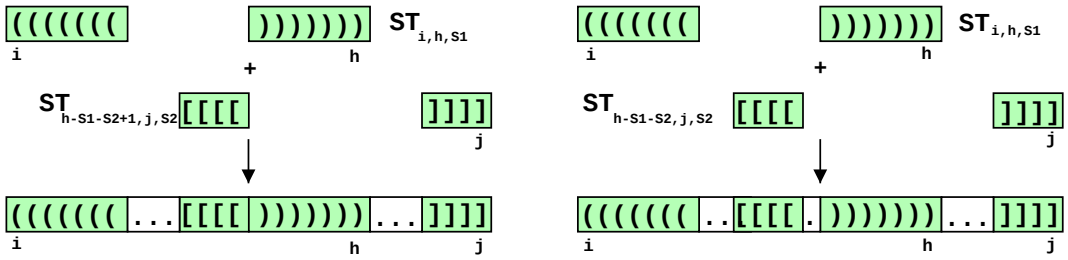
3.2.3 The Simple h-Pseudoknots

A simple h-pseudoknot consists of two overlapping stacks which are coaxially stacked. The fact that we only consider stacks enables us to reduce the number of indices. We give only the recursions for h_0-pseudoknots. $ST_{i,j,S}$ denotes a stack with the terminal base pair (i, j) and stack size S.

$$ST_{i,j,S} = \begin{cases} 1 & : \text{ stack compatible with the sequence} \\ 0 & : \text{ stack not compatible with the sequence} \end{cases}$$

$$ST_{i,j+1,S} = ST_{i+1,j,S-1}\Pi_{\sigma_i,\sigma_{j+1}}$$

$$PK_{i,j} = \sum_{h=h_{min}}^{h_{max}} \left[\sum_{\substack{S_1=S_{1min} \\ ST=1}}^{S_{1max}} ST_{i,h,S_1} \left[\sum_{\substack{S_2=S_{2min} \\ ST=1}}^{S_{2max}} ST_{h-S_1-S_2+1,j,S_2} + \sum_{\substack{S_2=S_{2min} \\ ST=1}}^{S_{2max'}} ST_{h-S_1-S_2,j,S_2} \right] \right] \tag{3.8}$$



There are two loops for S_2 because one intervening base is allowed. Each loop for the stack-sizes can be broken if $ST = 0$. The start and stop conditions for the loops enable us to include the stereo-chemical situation into the recursions.

If we neglect all geometrical constrains for the loops and choose minimum stack lengths ($S_{1\ min}, S_{2\ min}$) only:

$$h_{min} = 2\ S_{1\ min} + S_{2\ min} \tag{3.9}$$

$$h_{max} = j - S_{2\ min} \tag{3.10}$$

$$S_{1\ max} = \frac{(h - i + 1) - S_{2\ min}}{2} \tag{3.11}$$

$S_{2\ max}$ can be limited either by the upstream or the downstream part of the stack.

$$S_{2\ max} = \min \begin{cases} S_{2\ max\ up} = (h - i + 1) - 2\ S_1 \\ S_{2\ max\ down} = j - h \end{cases} \tag{3.12}$$

$$S_{2\ max} = \min \begin{cases} S_{2\ max\ up} = (h - i) - 2\ S_1 \\ S_{2\ max\ down} = j - h \end{cases} \tag{3.13}$$

As pointed out previously, stacks and loops are not independent. We need two introduce $L_{1\ min}(S_2)$ and $L_{2\ min}(S_1)$:

$$h_{min} = 2\ S_{1\ min} + S_{2\ min} + L_{1\ min}(S_{2\ min}) \tag{3.14}$$

$$h_{max} = j - (S_{2\ min} + L_{2\ min}(S_{1\ min})) \tag{3.15}$$

$$S_{1\ max} = \frac{h - i + 1 - S_{2\ min} - L_{1\ min}(S_{2\ min})}{2} \tag{3.16}$$

$$S_{2\ max} = \min \begin{cases} S_{2\ max\ up} = \max\ \{S_2 | L_{1\ min}(S_2) + S_2 \leq h - i + 1 - 2\ S_1\} \\ S_{2\ max\ down} = j - h - L_{2\ min}(S_1) \end{cases} \tag{3.17}$$

$$S_{2\ max'} = \min \begin{cases} S_{2\ max'\ up} = \max\ \{S_2 | L_{1\ min}(S_2) + S_2 \leq h - i - 2\ S_1\} \\ S_{2\ max'\ down} = j - h - L_{2\ min}(S_1) \end{cases} \tag{3.18}$$

3.2.4 The Restricted h-Pseudoknot

The introduction of interior loops and bulges into the stacks brings the concept of the restricted h-pseudoknot into the game. It is an intermediate between the general h-pseudoknot and the simple h-pseudoknot. For the beginning we impose strong restrictions: only one symmetric interior loop with size 2 or one bulge with size one per building block is allowed. Thus we need three matrices for the building blocks:

$$
\begin{array}{lcrcl}
B_{i,j,S} & : & |j-q|-|p-i| & = & 0 \quad \text{no bulges } S=j-q+1 \\
B^5_{i,j,S} & : & |j-q|-|p-i| & = & -1 \quad \text{bulge at 5' interval } [i,p], S=p-i+1 \\
B^3_{i,j,S} & : & |j-q|-|p-i| & = & 1 \quad \text{bulge at 3' interval } [q,j], S=j-q+1
\end{array}
$$

The numbers for the building blocks are also calculated recursively.

$$B_{i,j,S} = B_{i+1,j-1,S-1} * \Pi_{\sigma_i,\sigma_j} + ST_{i+2,j-2,S-2} * \Pi_{\sigma_i,\sigma_j} \tag{3.19}$$

$$B^5_{i,j,S} = B^5_{i+1,j-1,S-1} * \Pi_{\sigma_i,\sigma_j} + ST_{i+2,j-1,S-2} * \Pi_{\sigma_i,\sigma_j} \tag{3.20}$$

$$B^3_{i,j,S} = B^3_{i+1,j-1,S-1} * \Pi_{\sigma_i,\sigma_j} + ST_{i+1,j-2,S-2} * \Pi_{\sigma_i,\sigma_j} \tag{3.21}$$

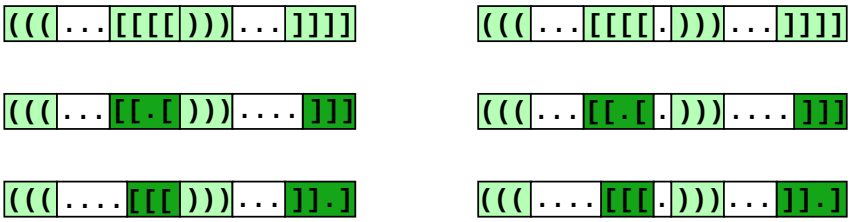
In order to get the number of pseudoknots we again have to count all allowed combinations of building blocks. For the sake of clarity we combine the three variables B, B^5, B^3 together by introducing another index A which denotes the difference in length between the upstream and downstream part of a building

block.

$$B_{i,j,S,A} = \begin{cases} A = -1 : B^3_{i,j,S} \\ A = 0 : B_{i,j,S} \\ A = +1 : B^5_{i,j,S} \end{cases} \tag{3.22}$$

$$PK_{i,j} = \sum_{h=h_{min}}^{h_{max}} \sum_{S_{1min}}^{S_{1max}} \sum_{-1 \le A_1 \le 1} B_{i,h,S_1,A_1} \left[\sum_{S_{2min}}^{S_{2max}} \sum_{-1 \le A_2 \le 1} B_{h-S_1-S_2+1,j,S_2,A_2} + \sum_{S_{2min}}^{S_{2max}} \sum_{-1 \le A_2 \le 1} B_{h-S_1-S_2,j,S_2,A_2} \right] \tag{3.23}$$

For the innermost sum $-1 \le A \le 1$ we give the following example:



All other sums are calculated in the same way. Like in the case of simple h_0-pseudoknots, the stereo-chemical constrains are considered in the start and stop conditions of the sums. However, because building blocks are more variable it gets more confusing to write down all cases.

When we do not impose any kind of coaxial stacking we get the following recursion:

$$PK_{i,j} = \sum_{h=h_{min}}^{h_{max}} \sum_{S_{1min}}^{S_{1max}} \sum_{-1 \le A_1 \le 1} B_{i,h,S_1,A_1} \left[\sum_{m=0}^{m_{max}} \sum_{S_{2min}}^{S_{2max}} \sum_{-1 \le A_2 \le 1} B_{h-S_1-S_2+1-m,j,S_2,A_2} \right] \tag{3.24}$$

where m_{max} determines the size of $L3$ and again is limited by stereo-chemistry. Further relaxations of the restrictions are made by changing the recursions for the building blocks to introduce more or larger loop, which will be discussed later (mfe-folding).

Compared to the general $h - pseudoknot$ the restricted version seems to be needlessly complicated. A lot more sums and indices have to be considered. But the striking difference is that we have at any stage of the recursion the information about the size of the loops $L1$,$L2$,$L3$ and the corresponding building blocks. This information is crucial to decide whether a given base pair pattern is stereo-chemically feasible and helps to avoid dispensable areas of the structure space.

3.3 The Maximum Matching

Since the enumerations are equivalent with an implicit definition of the structure space they do not need a complicated scoring function. The contribution of a potential structure to the total sum is either 1 or 0. To decide if a potential new structure is compatible with a sequence needs just one lookup in the Π matrix. A very similar algorithm solves the maximum matching problem. Maximum matching is concerned with finding the structure providing the maximum number of base pairs and is therefore an optimization problem. The dynamic programming solution works on the basis of the following recursion:

$$PS_{i,j+1} = \max \begin{cases} PS_{i,j} \\ \max\limits_{i \le k \le j-hm} \Big[PS_{i,k-1} + 1 + PS_{k+1,j} \Big] \Pi_{\sigma_k, \sigma_j+1} \end{cases} \tag{3.25}$$

which fills the PS matrix. The element $PS_{1,n}$ holds the maximum number of base pairs the sequence allows. The maximum base pair number of all possible subsequences are stored in the corresponding elements $PS_{i,j}$.

3.3.1 Maximum Matching with h-Pseudoknots

Like in the case of the enumerations we introduce two new variables, $PPK_{i,j}$ the max. number of pseudoknot base pairs on (i, j) and $PPS_{i,j}$ the max. number of base pairs including pseudoknots.

$$PPS_{i,j+1} = \max \begin{cases} PPS_{i,j} \\ \max\limits_{i \le k \le j-hm} \Big[PPS_{i,k-1} + 1 + PS_{k+1,j} \Big] \Pi_{\sigma_k, \sigma_j+1} \\ \max\limits_{i \le k \le j-pm} \Big[PPS_{i,k-1} + PPK_{k,j+1} \Big] \end{cases} \tag{3.26}$$

There are two recursions coupled with PPS: PS is given above and PPK corresponds to the h_0-pseudoknot. Again there is no problem to allow pseu-

doknots inside a loop:

$$PPS_{i,j+1} = \max \begin{cases} PPS_{i,j} \\ \max\limits_{i \le k \le j-hm} \Big[PPS_{i,k-1} + 1 + PPS_{k+1,j} \Big] \Pi_{\sigma_k, \sigma_j + 1} \\ PPK_{i,j+1} \end{cases} \tag{3.27}$$

Obviously there is a simple way to convert the recursions for enumerations (**equation 3.1**) to recursions for maximum matching: from multiplication to addition, and from n-additions to finding the maximum of n-alternatives.

Enumeration	Max.-Match
$S_a S_b$	$PS_a + PS_b + 1$
$\sum\limits_n S$	$\max\limits_n PS$

Proceeding to the pseudoknot recursions is now straight forward, thus we only give the details for the general and the simple h_0-pseudoknot.

The General h-Pseudoknot

$$PB_{i,k,l,j} = \max \begin{cases} PB_{i,k,l,j-1} \\ \max\limits_{i \le x \le k} \Big[PB_{x+1,k,l,j-1} + 1 \Big] \Pi_{\sigma_x, \sigma_j} \end{cases} \tag{3.28}$$

$$PBO_{i,k,l,j} = \Big[PB_{i+1,k,l,j-1} + 1 \Big] \Pi_{\sigma_i, \sigma_j} \tag{3.29}$$

$$PBR_{i,k,l,j} = \max_{i+1 \le x \le k} \Big[PBO_{i,x-1,l+1,j} + 1 \Big] \Pi_{\sigma_x, \sigma_l} \tag{3.30}$$

$$PPK_{i,j} = \max_{i<k<j} \left\{ \max_{k<l<j} \left\{ \max_{l<m<j} \Big[PBR_{i,k,l,m} + PBO_{k+1,l-1,m+1,j} \Big] \right\} \right\} \tag{3.31}$$

The Simple h-Pseudoknot

$$PST_{i,j+1,S} = \Big[PST_{i+1,j,S-1} + 1 \Big] \Pi_{\sigma_i, \sigma_{j+1}} \tag{3.32}$$

$$PPK_{i,j} = \max_{h} \left\{ \max_{S_1} \left\{ PST_{i,h,S_1} + \max \left\{ \begin{array}{l} \max_{S_2} PST_{h-S_1-S_2+1,j,S_2} \\ \max_{S_2'} PST_{h-S_1-S_2',j,S_2'} \end{array} \right\} \right\} \right\} \tag{3.33}$$

Of course we apply the same start and stop conditions for finding the maximum as we did to calculate the corresponding sum.

3.3.2 Backtracking

While the first two steps of dynamic programming fill all the matrices, the last step is concerned with constructing a structure (or many structures, in case of suboptimal folding). This task is accomplished by searching systematically through the stored matrices proceeding from larger to smaller fragments. We give a fairly detailed description of the backtracking procedure because it is the fundament for the suboptimal folding algorithm.

Secondary Structure Backtracking

Basically we ask which two smaller segments are sticked together to get one larger segment, thus we do not look for the maximum m of n alternatives, but for the first exact match m of n alternatives.

$$PS_{i,j+1} == \left\{ \begin{array}{l} PS_{i,j} \\ \underset{k}{loop} \Big[PS_{i,k-1} + 1 + PS_{k+1,j} \Big] \Pi_{\sigma_k,\sigma_j+1} \end{array} \right. \tag{3.34}$$

Consequently for each match we get two smaller segments which enter the same process until the breaking condition is reached. The breaking condition is determined by the minimal size of a hairpin loop. $P_{i,j+1} == P_{i,j}$ is just a special case where we already know that one of the subsegments already reached the breaking condition. The recursion indicates a *last in, first out* or *depth first* strategy. In other words the two shorter segments are pushed on a stack, then the last pushed segment is popped from the stack and again

divided into two smaller segments which are pushed again on the stack. The procedure continues until no segment is left to pop from the stack. Whenever two segments are produced the corresponding characters '(',')' or '.' are written to the structure string. Again there is a structural similarity between the expressions used for the max. matching and the backtracking.

Backtracking	Max.-Match
$PS == \underset{n}{loop} PS_n$	$PS = \max_n PS_n$

The pseudoknot conditions are straight forward, thus only the general and the simple h-pseudoknots are outlined.

H-Pseudoknot Backtracking

$$PPS_{i,j+1} == \begin{cases} PPS_{i,j} \\ \underset{i \le k \le j-hm}{loop} \Big[PPS_{i,k-1} + 1 + PS_{k+1,j} \Big] \Pi_{\sigma_k, \sigma_j+1} \\ \underset{i \le k \le j-pm}{loop} \Big[PPS_{i,k-1} + PPK_{k,j+1} \Big] \end{cases} \tag{3.35}$$

The General h-Pseudoknot

$$PPK_{i,j} == \underset{i<k<j}{loop} \left\{ \underset{k<l<j}{loop} \left\{ \underset{l<m<j}{loop} \Big[PBR_{i,k,l,m} + PBO_{k+1,l-1,m+1,j} \Big] \right\} \right\} \tag{3.36}$$

$$PB_{i,k,l,j} == \begin{cases} PB_{i,k,l,j-1} \\ \underset{i \le x \le k}{loop} \Big[PB_{x+1,k,l,j-1} + 1 \Big] \Pi_{\sigma_x, \sigma_j} \end{cases} \tag{3.37}$$

$$PBO_{i,k,l,j} == \Big[PB_{i+1,k,l,j-1} + 1 \Big] \Pi_{\sigma_i, \sigma_j} \tag{3.38}$$

$$PBR_{i,k,l,j} == \underset{i+1 \le x \le k}{loop} \Big[PBO_{i,x-1,l+1,j} + 1 \Big] \Pi_{\sigma_x, \sigma_l} \tag{3.39}$$

The Simple h-Pseudoknot

$$PPK_{i,j} == \underset{h}{loop}\left\{\underset{S_1}{loop}\left\{PST_{i,h,S_1} + \left\{\begin{array}{l}\underset{S_2}{loopPST}_{h-S_1-S_2+1,j,S_2} \\ \underset{S_2'}{loopPST}_{h-S_1-S_2',j,S_2'}\end{array}\right\}\right\}\right\} \tag{3.40}$$

3.4 RNA Secondary Structure Folding

Dynamic programming is not the only computational method to predict the secondary structure of a given sequence. There are several ways to deduce an RNA structure from a given sequence. In principle we can divide them into to broad classes: Folding by *phylogenetic comparison* and *energy directed* folding.

3.4.1 Phylogenetic Structure Analysis

Given a large enough number of sequences with identical secondary structure, that structure can be deduced by examining covariances of nucleotides in these sequences. This is the principle used for structure prediction through phylogenetic comparison of *homologous* (common ancestry) sequences. Basically these methods just look for compensatory mutations such as an A change to C in position i of the aligned sequences simultaneously with a change from U to G in position j, indicating a base pair (i, j). So the sequence alignment is the most complicated theoretical part (if the sequences in the set are to dissimilar). The basic assumption is that structure is more conserved during evolution than sequence, since it is the structure that determines function. The only experimental information needed is a large enough number of sequences. Fortunately nucleic acid sequences are nowadays one of the best accessible molecular biological informations. In fact the success of the method in the prediction of, for instance, the secondary structures of the 16S ribosomal RNAs,RNAseP RNA or the clover-leaf structure of tRNAs provides an excellent justification for this method.

The advantages: Since no assumptions about pairing rules are necessary, non-canonical pairs and tertiary interactions can be detected as well.

The disadvantages: A sufficiently large set of sequences which exhibit the proper amount of variation has to be provided. So the sequences should be dissimilar enough to show many covariations while still yielding a good alignment. If there are strongly conserved regions (i.e. the function is sequences dependent) or parts of the structure are highly variable (because non-functional)

our assumption holds not true. As a consequence, phylogenetically determined structures usually are incomplete, that means, they do not show all base pairs of the actual structures.
Nevertheless phylogenetic comparison can generate the most reliable structure models to date and are therefore frequently used for comparison of other folding algorithms.

3.4.2 Energy Directed Folding

Most methods for prediction of RNA secondary structure work on the basis of energy model presented in chapter 1. They can be further divided into methods that try to find the structure of minimal free energy (or equilibrium ensemble) and "kinetic" algorithms.

Minimal free energy (mfe) Zucker *et al.* published the first dynamic programming mfe-algorithm with an energy model which is still used without principle changes. It is the object of this work to extend this algorithm by introducing pseudoknots. Mfe-folding can be extended to calculate suboptimal structures within a given energy band above the minimum free energy. This method provides us with a lot of useful information about the RNA folding landscape.

Kinetic folding algorithms try to mimic the folding process in order to derive the biologically active structure. They can simulate the process of RNA 5'-3' synthesis, pseudoknots and other tertiary interactions can be included. It is by no means clear that the biological relevant structure of RNA is the structure of minimal energy, instead the structure might be trapped in some local minimum during the folding process. Therefore kinetic folding algorithms are not the fist choice to find the mfe-structure. We manly use them to simulate the time dependence of folding and refolding and to study the characteristics of folding paths. In contrast to mfe- algorithms, kinetic-algorithms are more flexible regarding the energy function and structural diversity.

3.5 Minimum Free Energy Folding with H-Pseudoknots

Proceeding from the maximum matching algorithm to the mfe-folding algorithm means a significant increase in complexity regarding the scoring i.e. energy function. As a consequence we need more different matrices and a few approximations.

Let $C_{i,j}$ be the minimum energy possible on the substructure $I_{i,j}$ provided that i and j pair and $I_{i,j}$ is not part of or contains a pseudoknot. Since the energy of some substructure $S_{i,j}$ with i and j paired is given by the energy of the loop closed by (i, j) plus the energy of any loops directly interior to it,

$$C_{i,j} = \min\{\mathcal{E}(L) + \sum C_{p,q}\} \tag{3.41}$$

and $C_{i,i} = \infty$.
Three subsets are contributing to this set of structures, depending on the number of base pairs immediately interior to (i, j). The minimum energies of these three subsets are (recursively) obtained from smaller fragments:

$$C_{i,j} = \min \begin{cases} \mathcal{H}(i,j) \\ \text{(.....)} \\ \min\{C_{p,q} + \mathcal{I}(i,j,p,q)\} \\ \text{(....[ ]......)} \\ \min\{F^M_{i+1,k-1} + F^M_{k,j-1} + M_C\} \\ \text{([ ][ ])} \end{cases} \tag{3.42}$$

$H(i, j)$ denotes the free energy of a hairpin loop closed by (i, j). The second subset is the minimum energy of all structures, where (i, j) closes a loop of degree 2 (stacks, bulges and interior loops); their minimum energy equals the sum of the minimum energy of the smaller fragment, $C_{p,q}$ and the energy of the closing loop, $\mathcal{I}(i, j, p, q)$. In the third element multi loop structures enclosed by (i, j) are obtained by constructing the multi loop from two parts, $F^M_{i+1,k-1}$ and $F^M_{k,j-1}$, plus the multi loop closing energy M_C. If we consider dangling

ends for multi loops we have to consider four cases. $d^{5,(3)}_{i,j,j-1,(i+1)}$ denotes the energy contribution of the 5'(3') dangling end indicating the (i,k)-pair and the $i-1(j+1)$ unpaired end.

$$C_{\substack{i,j \\ Multi}} = \min \begin{cases} F^M_{i+1,k-1} + F^M_{k,j-1} + \mathcal{M}_\mathcal{C} \\ \text{(} \quad \text{)} \\ F^M_{i+2,k-1} + F^M_{k,j-1} + d^3_{i,j,i+1} + \mathcal{M}_\mathcal{C} \\ \text{(<} \quad \text{)} \\ F^M_{i+1,k-1} + F^M_{k,j-2} + d^5_{i,j,j-1} + \mathcal{M}_\mathcal{C} \\ \text{(} \quad \text{>)} \\ F^M_{i+2,k-1} + F^M_{k,j-2} + d^5_{i,j,j-1} + d^3_{i,j,i+1} + \mathcal{M}_\mathcal{C} \\ \text{(<} \quad \text{>)} \end{cases} \tag{3.43}$$

The $F^M_{i,j}$ are calculated recursively with the initial condition $F^M_{i,i} = \infty$.

$$F^M_{i,j} = \min \begin{cases} C_{i,j} + \mathcal{M}_\mathcal{I} \\ \\ C_{i+1,j} + d^5_{i+1,j,i} + \mathcal{M}_\mathcal{I} \\ \text{>} \\ C_{i,j-1} + d^3_{i,j-1,j} + \mathcal{M}_\mathcal{I} \\ \text{<} \\ C_{i+1,j-1} + d^5_{i+1,j-1,i} + d^3_{i+1,j-1,j} + \mathcal{M}_\mathcal{I} \\ \text{>} \quad \text{<} \\ F^M_{i+1,j} + \mathcal{M}_\mathcal{B} \\ \text{.} \\ F^M_{i,j-1} + \mathcal{M}_\mathcal{B} \\ \text{.} \\ \min(F^M_{i,k-1} + F^M_{k,j}) \\ \\ \end{cases} \tag{3.44}$$

If we do not require (i,j) to be paired and again include the dangles we get the minimum fee energy on (i,j) analog to the multi loop case but without

assigning energies to non-dangling bases and base pairs ($\mathcal{M}_{\mathcal{B}} = \mathcal{M}_{\mathcal{I}} = 0$).

$$F_{i,j} = \min \begin{cases} C_{i,j} \\ C_{i+1,j} + d^5_{i+1,j,i} \\ C_{i,j-1} + d^3_{i,j-1,j} \\ C_{i+1,j-1} + d^5_{i+1,j-1,i} + d^3_{i+1,j-1,j} \\ PK_{i,j} \\ PK_{i+1,j} + pkd^5_{i+1,i} \\ PK_{i,j-1} + pkd^3_{j-1,j} \\ PK_{i+1,j-1} + pkd^5_{i+1,i} + pkd^3_{j-1,j} \\ \min_{i<k\leq j}(F_{i,k-1} + F_{k,j}) \end{cases} \tag{3.45}$$

Let $PK_{i,j}$ be the minimum energy possible on the substructure $I_{i,j}$ provided that this substructure is a restricted h-pseudoknot. The best $PK_{i,j}$ is obtained by producing all possible combinations of best building blocks.

$$PK_{i,j} = \min\{B^u + B^d + \mathcal{E}_{L1} + \mathcal{E}_{L2} + \mathcal{E}_{Coax}\} \tag{3.46}$$

B^u denotes the best upstream and B^d the best downstream building block. The superscripts u and d do not denote two different matrices but two different sets of indices. Of course B^u and B^d have to be h-type generating and stereo-compatible. A building block $B_{i,p,q,j}$ can be viewed as a substructure with a gap, where (i, j) and (p, q) are base pairs. The degree of each loop (except (p, q)) comprising this substructure is 2. Since we impose strong restrictions

on the building blocks we use the same notation as for the enumerations.

The Simple h_0-Pseudoknot

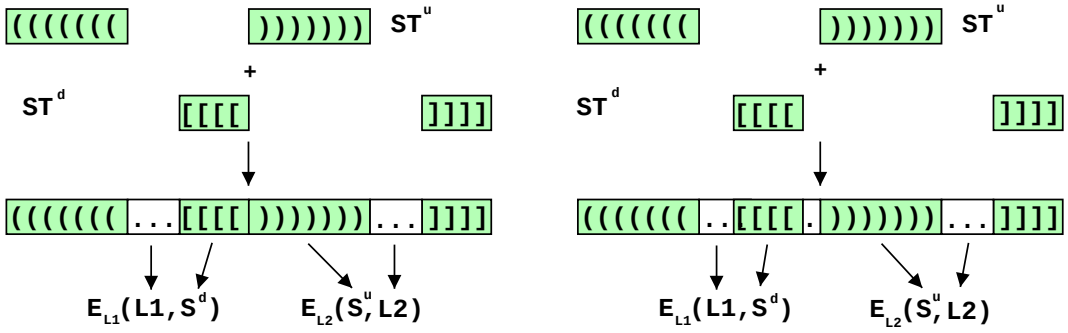
Of course all recursions are very similar to the enumerations. The best building blocks are calculated recursively. Let $ST_{i,j,S}$ be the energy of a stack with the closing pair (i, j) and stack-size S.

$$ST_{i,j,S} = ST_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \tag{3.47}$$

where $ST_{i,i,S} = \infty$ and $ST_{i,j,1} = \infty$ which means that we only consider pseudoknots with minimal stack-size 2. Equation 3.46 simplifies to

$$PK_{i,j'} = \min\{ST^u_{i,j,S^u} + ST^d_{i',j',S^d} + \mathcal{E}_{L1}(L1, S^d) + \mathcal{E}_{L2}(L2, S^u) + \mathcal{E}_{Coax}\} \tag{3.48}$$

where the destabilizing contributions of the loops $(L1, L2)$ depend on the loop-size and the stack-size to bridge.



In case of an intervening base we have to check from which loop the mismatch partner should come to yield the best possible coaxial stacking energy. Unfortunately the energy model does not give us any hint how to deal with this situation. There might be some stereo-chemical preference for one loop to donate the mismatch partner, or the excess of bases (over the minimal number of bases in a loop to bridge the stem) determines the stacking behavior. However,

in our implementation we compare the two alternatives and take the one with the lower free energy.

The Strongly Restricted h_0-Pseudoknot

In the strongly restricted h_0-pseudoknot model we allow one interior loop or bulge per building block. The bulge may be formed by one unpaired base, the interior loop has to be symmetric and contains two unpaired bases. This definition implies that three kinds of building blocks can be produced:

$$B_{i,j,S} = \min \begin{cases} B_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ (|(((\qquad))))| \\ ST_{i+2,j-2,S-2} + \mathcal{I}(i, i+2, j-2, j) \\ (|.(((\qquad)))|.| \end{cases} \tag{3.49}$$

$$B^5_{i,j,S} = \min \begin{cases} B^5_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ (|((.(\qquad)))| \\ ST_{i+2,j-1,S-2} + \mathcal{I}(i, i+2, j-1, j) \\ (|.(((\qquad)))| \end{cases} \tag{3.50}$$

$$B^3_{i,j,S} = \min \begin{cases} B^3_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ (|(((\qquad)).)| \\ ST_{i+1,j-2,S-2} + \mathcal{I}(i, i+1, j-2, j) \\ (|(((\qquad)))|.)| \end{cases} \tag{3.51}$$

Relaxed Restrictions for h_0-Pseudoknots

There is no principle problem to relax the strong restriction imposed on the building blocks. We just have to consider that more variable building blocks imply increasing memory and time demand. It is also questionable if the energy model for h_0-pseudoknots is still applicable if we allow the stacks to be disturbed by larger and more interior loops and bulges. Nevertheless we give examples in which way our system of recursions has to be extended.

In the first two examples we use the same range of building block-asymmetry as before (the difference between the length of the upstream and the downstream

part of a building block is $-1 \leq A \leq 1$), which means no increase in memory demand.

In the first example we do not impose a limit on the number of bulges and interior loops per building block but stick to the previous maximum sizes:

$$B_{i,j,S} = \min \begin{cases} B_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ B_{i+2,j-2,S-2} + \mathcal{I}(i, i+2, j-2, j) \\ B^3_{i+2,j-1,S-2} + \mathcal{I}(i, i+2, j-1, j) \\ B^5_{i+1,j-2,S-2} + \mathcal{I}(i, i+1, j-2, j) \end{cases} \tag{3.52}$$

$$B^5_{i,j,S} = \min \begin{cases} B^5_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ B_{i+2,j-1,S-2} + \mathcal{I}(i, i+2, j-1, j) \end{cases} \tag{3.53}$$

$$B^3_{i,j,S} = \min \begin{cases} B^3_{i+1,j-1,S-1} + \mathcal{I}(i, i+1, j-1, j) \\ B_{i+1,j-2,S-2} + \mathcal{I}(i, i+1, j-2, j) \end{cases} \tag{3.54}$$

As we can see there is no significant increase in time dependence. In the second example we introduce larger interior loops and bulges.

$$B_{i,j,S1} = \min \begin{cases} \min\{B_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i| = |j-q| \\ \min\{B^5_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i+1| = |j-q| \\ \min\{B^3_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i-1| = |j-q| \end{cases} \tag{3.55}$$

$$B^5_{i,j,S1} = \min \begin{cases} \min\{B_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i-1| = |j-q| \\ \min\{B^5_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i| = |j-q| \\ \min\{B^3_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i-2| = |j-q| \end{cases} \tag{3.56}$$

$$B^3_{i,j,S1} = \min \begin{cases} \min\{B_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i| = |j-q-1| \\ \min\{B^5_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i| = |j-q-2| \\ \min\{B^3_{p,q,S2} + \mathcal{I}(i, j, p, q)\}, |p-i| = |j-q| \end{cases} \tag{3.57}$$

where $|j-q-1| + |p-i-1| \leq I_{max}$, which is the same maximum interior loop size as used for ordinary secondary structures. Because of the restricted

building block asymmetry the interior loops also have to meet certain conditions beside a maximum size which are specified for each case. The time scale is still not altered dramatically as long as we require a maximum interior loop size.

The next step in relaxing the conditions of our h_0-pseudoknot is to allow a larger asymmetry between the upstream and the downstream part of a building block. It is convenient so use the same notation as we used for the enumerations:

$$B_{i,j,S,A} = \begin{cases} A = -1 : B^3_{i,j,S} \\ A = 0 : B_{i,j,S} \\ A = +1 : B^5_{i,j,S} \end{cases} \tag{3.58}$$

In the previous example we had three entries for each $B_{i,j,S}$ and for each $B_{i,j,S,A}$ three types of combinations of building blocks and interior loops. When we only focus on the asymmetries we can write a reduced form:

$$0 = \min \begin{cases} \min \{ 0 \quad 0\} \\ \min \{+1 \quad -1\} \\ \min \{-1 \quad +1\} \end{cases}$$

$$+1 = \min \begin{cases} \min \{ 0 \quad +1\} \\ \min \{+1 \quad 0\} \\ \min \{-1 \quad +2\} \end{cases} \tag{3.59}$$

$$-1 = \min \begin{cases} \min \{ 0 \quad -1\} \\ \min \{+1 \quad -2\} \\ \min \{-1 \quad 0\} \end{cases}$$

When the asymmetry range is changed to $-2 \leq A \leq +2$ the following combinations have to be checked:

$$
\begin{aligned}
0 = \min \begin{cases} \min & \{-2 & +2\} \\ \min & \{-1 & +1\} \\ \min & \{\ 0 & \ 0\} \\ \min & \{+1 & -1\} \\ \min & \{+2 & -2\} \end{cases}
\qquad
+1 = \min \begin{cases} \min & \{-2 & +3\} \\ \min & \{-1 & +2\} \\ \min & \{\ 0 & +1\} \\ \min & \{+1 & \ 0\} \\ \min & \{-2 & -1\} \end{cases}
\\
-1 = \min \begin{cases} \min & \{+2 & -3\} \\ \min & \{+1 & -2\} \\ \min & \{\ 0 & -1\} \\ \min & \{-1 & \ 0\} \\ \min & \{+2 & +1\} \end{cases}
\qquad
+2 = \min \begin{cases} \min & \{-2 & +4\} \\ \min & \{-1 & +3\} \\ \min & \{\ 0 & +2\} \\ \min & \{+1 & +1\} \\ \min & \{+2 & \ 0\} \end{cases}
\\
-2 = \min \begin{cases} \min & \{+2 & -4\} \\ \min & \{+1 & -3\} \\ \min & \{\ 0 & -2\} \\ \min & \{-1 & -1\} \\ \min & \{-2 & \ 0\} \end{cases}
\end{aligned}
\tag{3.60}
$$

Clearly this approach can be extended to larger ranges of asymmetry in an analogous way. At the end, when we do not impose any restrictions to the gaps of the building block, each interval (i, j) needs $\mathcal{O}(n^2)$ memory (the number of possible gaps) and $\mathcal{O}(n^2)$ time. Recursion 3.46 for the best h_0-pseudoknot on the segment $[i, j]$ generalizes to:

$$
PK_{i,j} = \min_{h,S_u,A_u} \left\{ B_{\upsilon}^{u} + \min \left\{ \begin{array}{l} \min\limits_{S_d,A_d} B_{\delta}^{d} + \mathcal{E}_{Loops} + \mathcal{E}_{Coax} \\ \min\limits_{S'_d,A'_d} B_{\delta'}^{d'} + \mathcal{E}_{Loops} + \mathcal{E}_{Coax} \end{array} \right\} \right\} \tag{3.61}
$$

$$
\begin{aligned}
\upsilon &= i, h, S_u, A_u \\
\delta &= h - S_u - S_d + 1, j, S_d, A_d \\
\delta' &= h - S_u - S'_d, j, S'_d, A'_d
\end{aligned}
$$

Time and Space complexity

In the case of simple secondary structures two tricks facilitate a time dependence of $\mathcal{O}(len^3)$ and a memory requirement of $\mathcal{O}(len^2)$:

The Maximum Size for Interior Loops

With the help of a maximum size for interior loops, I_{max}, it takes a constant number of checks to find the best interior loop for a given closing base pair $C_{i,j}$. Without an I_{max} it would take $\mathcal{O}(n^2)$ steps for each $C_{i,j}$,$(n = j - i)$. This approach is applicable because it is unlikely that very large interior loops are more stable than any alternative secondary structure on a given subsequence (if the temperature is not to high). Obviously with the same argument we could introduce a maximum size for hairpin loops. The time saving effect would be very small because for each (i, j) we need only one lookup ($\mathcal{O}(0)$). It has to be mentioned that it is also possible to avoid a cutoff in loop-size and still retain a $\mathcal{O}(n^3)$ time dependence for the sequence ([33], [67]). However, the method requires $\mathcal{O}(n^3)$ memory and a special type of energy function for the interior loops.

The Linear Ansatz for Multi Loops

If we consider multi loops with a maximum degree k it takes time proportional to n^{2k} for every $C_{i,j}$ and k itself grows linear with n. This would definitely give a prohibitive time dependence. In contrast, the linear ansatz enables us to get the best multi loop closed by (i, j) in $\mathcal{O}(n)$ steps, by trying all possibilities to build the multi loop from two smaller subsequences. Of course the consequence is some loss in accuracy and an additional memory demand for the auxiliary matrix F^M.

The h-Pseudoknots

It is plausible that an enlarged structure space implies higher memory and time requirements. In our approach we impose reasonable restrictions on the building blocks and the way we combine them. The considerations which led

to the restrictions where guided by the energy model and the stereo-chemical characteristic of h-pseudoknots.

For a given building block $B_{i,k,l,j} = B_{i,j,S,A}$ we actually restrict the size and position of the gap $[k+1, l-1]$. Without this method we would need $\mathcal{O}(len^4)$ memory space for all possible building blocks. One the other hand, as long as we use a maximum value for S and a maximum range for A, the memory requirement scales with $\mathcal{O}(len^2)$, which is the same as for ordinary secondary structures except a constant factor which depends on the actual numbers for S_{max} and A.

The number and size of internal loops and bulges within a building block is neither a problem as long a maximum size for internal loops is used. The only restriction comes from the fact, that in the recursions larger building blocks are constructed from smaller ones and therefore the size and asymmetry of the interior loops is limited by S_{max} and the range of A as well.

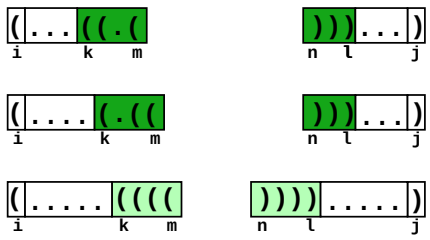


Figure 3.3: Implicit restrictions for interior loops beside I_{max}.
A base pair is wrapped around an interior loop and a building block $B_{k,m,n,l}$ to produce a larger building block $B_{i,m,n,j}$. $S_{max} = 9, \quad -1 \leq A \leq 1$;
A: legal combination, B: the asymmetry of the new building block is to big; C: the new building block is to big

In the next step we level the maximum size for the loops produced when we merge two building blocks. Again we use the same argument as for interior loops. The last restriction comes from the fact that only one type of coaxial stacking is considered which reduces the number of possible building block combinations.

All considerations given above enable the algorithm to

- produce all variations of pseudoknots which can be evaluated with the energy model
- avoid a huge number of pseudoknots for which we do not even have an approximate energy model
- do the calculation with the same order of space and time dependence, except a constant factor.

3.5.1 Backtracking

The backtracking for mfe-folding can be directly derived from the matrix filling recursions in the same way pointed out for the maximum matching case. The flow chart in **figure 3.5.1** gives a compact overview how the algorithm traces back without dangles (dangles would require additional case discriminations just spoiling clearness).
When we start with any given interval $[i, j]$ (for example $[1, len]$) and do not consider dangles $F_{i,j}$ is the result out of three alternatives:

- (i, j) is a base pair:

$$F_{i,j} = C_{i,j} \tag{3.62}$$

therefore we write a base pair in the structure string: $S_i =' ('$, $S_j =')'$. (i, j) can be the closing pair of a hairpin, an internal loop (stack and bulge included) or a multi loop.
If it closes a hairpin no further backtracking is necessary. In the case of an internal loop the condition

$$C_{i,j} = C_{p,q} + \mathcal{I}(i, j, p, q) \tag{3.63}$$

with $i < p < q < j$ finds (p, q) which is another base pair and therefore treated like (i, j). At a last possibility (i, j) closes a multi-loop, therefore we have to find a k, $i < k < j$ accomplishing

$$C_{i,j} = F^M_{i+1,k-1} + F^M_{k,j-1} \tag{3.64}$$

This condition yields to segments $[i+1, k]$ and $[k+1, j-1]$ which need to be traced back in the F^M array. Without dangles we have to distinguish 4 cases for each segment:

$$F^M_{i,j} = \begin{cases} C_{i,j} + \mathcal{M}_{\mathcal{I}} \\ F^M_{i,j-1} + \mathcal{M}_{\mathcal{B}} \\ F^M_{i+1,j} + \mathcal{M}_{\mathcal{B}} \\ F^M_{i,k-1} + F^M_{k,j} \end{cases} \tag{3.65}$$

in the first case we can write a base pair in the structure string and proceed the backtracking in the C array. In the second and third case we found an unpaired base at i (j) which is written in the structure string. In the fourth case a k has to be found fulfilling the given condition. All of the three last cases need further backtracking in the F^M array.

- (i, j) is not a base pair:
  first we check if there is a pseudoknot:

$$F_{i,j} = PK_{i,j} \tag{3.66}$$

  now we need to find the two appropriate building blocks that comprise the pseudoknot. Once we have them, we know in which building block arrays further backtracking has to be performed to elucidate their 'inner' structure (interior loop, bulges). Let's take for example a strongly restricted h_0-pseudoknot which is a combination of two building blocks accomplishing:

$$PK_{i,j} = B^u + B^d + \mathcal{E}_{L1} + \mathcal{E}_{L2} + \mathcal{E}_{Coax} \tag{3.67}$$

  The backtracking in B^u and B^d is done in the familiar way, by asking which smaller building block plus interior loop gives the larger one. Then we proceed with the smaller building block in the same way until there is nothing left to trace back.

- If (i, j) is not a base pair and $F_{i,j}$ is not a pseudoknot, we have to look for a k to accomplish.

$$F_{i,j} = F_{i,k-1} + F_{k,j} \tag{3.68}$$

  Again we obtain two segments $[i, k-1]$ and $[k, j]$, which both enter the backtracking process in the F array described at the top of this explanation.

It has to be mentioned that this kind of backtracking is not the most effective one and cannot be used for the generation of suboptimals. Nevertheless we think that a descriptive explanation facilitates the general understanding of backtracking.

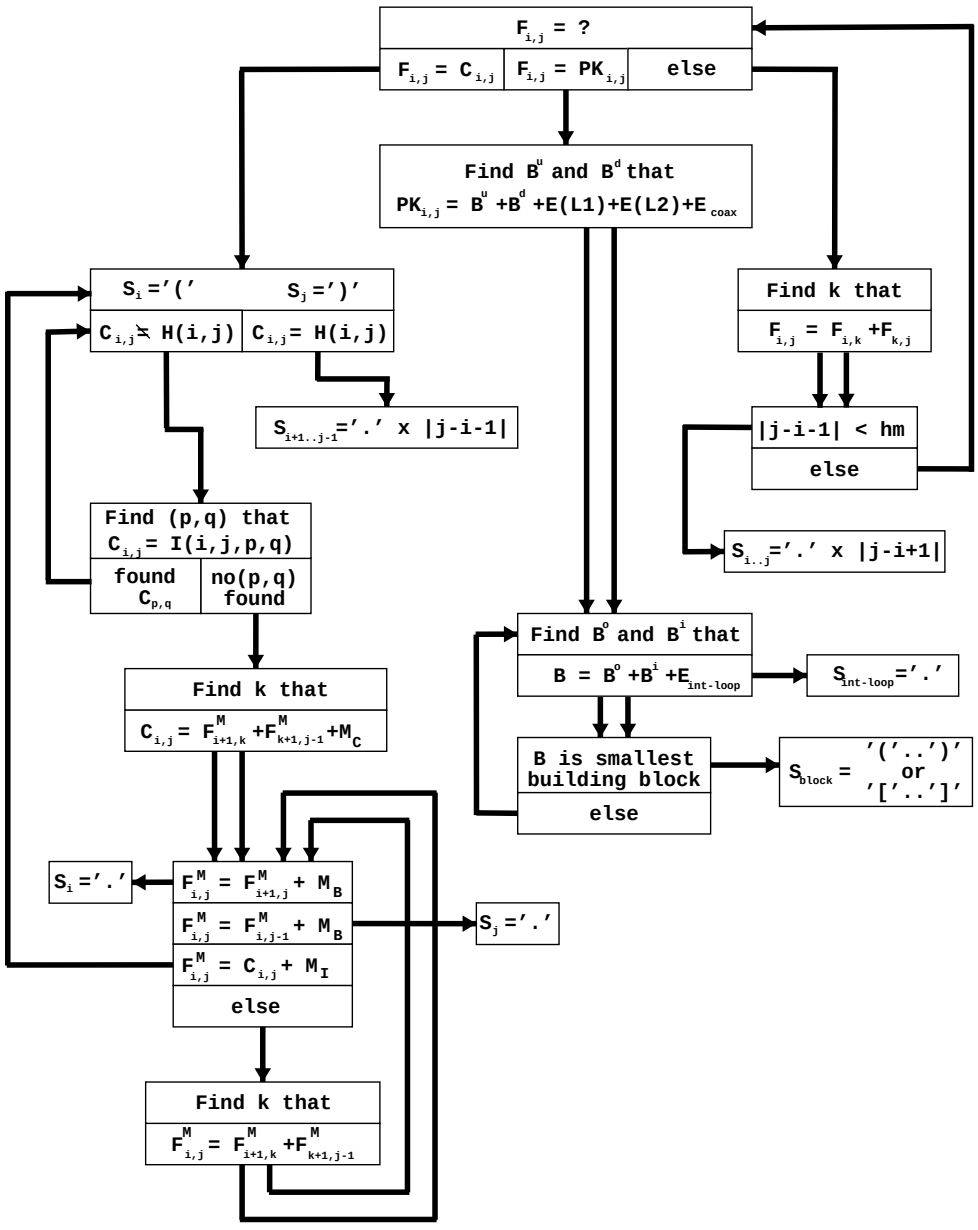


Figure 3.4: Backtracking the mfe-structure without considering dangles. B^u, B^d: upstream, downstream building block; B^o, B^i: outside, inside part of a building block;

3.6 Beyond h-pseudoknots

As outlined in previous sections, our notion of thermodynamic parameters for pseudoknots is mainly supported by theoretical considerations and experimentally proven secondary structures, leading to an approximation for the most simple kind of pseudoknot (the h-type). As a consequence we focused on the development of an algorithm that explores only the structure space which can be evaluated with this approximate energy model.
Indeed the most abundant entries in the database for pseudoknot secondary structures belong to the category of h-pseudoknots. Nevertheless we want to give an outline how to extend the algorithm to include less abundant types of pseudoknots.

3.6.1 The i-Pseudoknot

Analogous to the h-pseudoknot, this category might me called i-pseudoknot (or b-pseudoknot), because in this case the overlapping base pairs are formed by nucleotides form an interior loop (or bulge) and an exterior region. On the other hand we can view it as an h-type pseudoknot where each of the three unpaired regions $(L1, L2, L3)$ is allowed to form a one-component secondary structure. Actually the last description does more reflect the way our algorithm is extended.
Obviously the design of the algorithm strongly depends on the not yet developed energy model. With the h-type energy model in our mind, we want to anticipate some characteristics of i-pseudoknots, leading to two different algorithm-versions. However, we want to stress the fact that it is not the purpose of this work to develop an extended energy model.

Definition 1 An *i-pseudoknot* is an *h-pseudoknot* where each unpaired region $L1, L2$ and $L3$ is allowed to form a one-component secondary structure.

Like in the case of h-pseudoknots, we assume, that coaxial stacking is required to produce a stable i-pseudoknot. Additionally we postulate a stabilizing three-way junction. Consequently, if we start with an h-pseudoknot, five different positions of the additional component are possible:

- in $L1$ at the 5'-end - produces a 3w junction
  in $L1$ at the 3'-end - produces three coaxially stacked stems
- in $L2$ we produce a 3w junction
- in $L1$ at the 5'-end - produces three coaxially stacked stems
  in $L1$ at the 3'-end - produces a 3w junction

In all cases $L1$ and $L2$ still have to bridge the corresponding stems, hence it seems plausible to use the same approximations for their free energies as before. For a given combination of two building blocks $B^u_{i,k,l,j}$ and $B^d_{i',k',l',j'}$ the five potential positions are checked in the following way:

$$\mathcal{E}^1 = \min \begin{cases} \min\limits_{k<m<i'-hm} \{E_{L1}(L1, S2) + C_{m,i'-1} & , L1 = m-k-1 \\ \min\limits_{k+hm<m<i'} \{E_{L1}(L1, S2) + C_{k+1,m} & , L1 = i'-m-1 \\ E_{L1}(L1, S2), L1 = i'-k-1 \end{cases} \tag{3.69}$$

$$\mathcal{E}^2 = \min \begin{cases} \min\limits_{j<m<l'-hm} \{E_{L2}(L2, S1) + C_{m,l'-1} & , L2 = m-j-1 \\ \min\limits_{j+hm<m<l'} \{E_{L2}(L2, S1) + C_{j+1,m} & , L2 = l'-m-1 \\ E_{L2}(L2, S1), L2 = l'-j-1 \end{cases} \tag{3.70}$$

$$\mathcal{E}^3 = C_{k'+1,l-1} \tag{3.71}$$

and $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ correspond to the previous unpaired regions $L1, L2$ and $L2$.

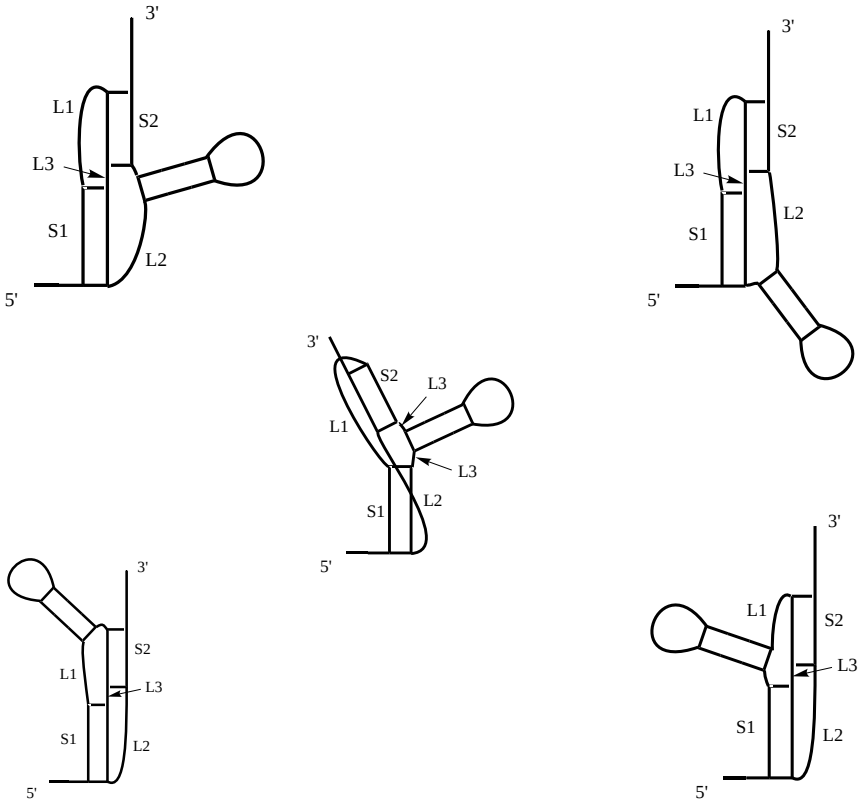


Figure 3.5: Five possibilities to place a component in one of the h-pseudoknot loops - resulting an i-pseudoknot.

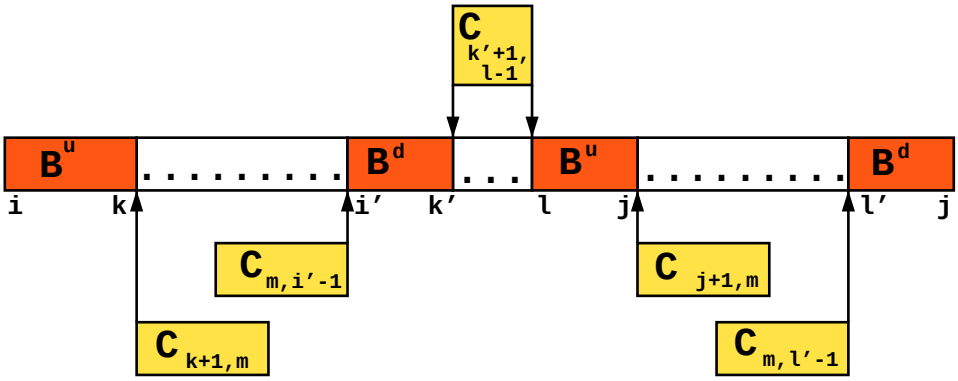


Figure 3.6: Representation of the i-pseudoknot generating recursions

Equation 3.46 generalizes to:

$$PK_{i,j} = \min\{B^u + B^d + \mathcal{E}^1 + \mathcal{E}^2 + \mathcal{E}^3 + \mathcal{E}^{SI}\} \tag{3.72}$$

$\mathcal{E}^{SI}$ is the free energy contribution from the stem interaction, either from two coaxial stackings or from a 3w junction. Of course we can store previously calculated values for $\mathcal{E}^1$ ($\mathcal{E}^2$) but we have to consider that for every $\mathcal{E}^1_{i,j}$ ($\mathcal{E}^2_{i,j}$) a third parameter for the bridged stack-size is needed. As we apply restrictions for the pseudoknots (hence for the third parameter) the time demand to fill the new arrays, scales with $\mathcal{O}(len)$, while the memory scales with $\mathcal{O}(len^2)$.
In the next approach we discuss the implementation of a linear ansatz and the additional structural freedom the algorithm consequently gains.

3.6.2 Linear Ansatz for Pseudoknot Loops

The introduction of a linear ansatz facilitates the handling of pseudoknot in a similar way it does with the multi loops in simple secondary structures. Nevertheless we have to take into account that we sacrifice the control over the stereo-chemical constraints.
We replace the formerly unpaired regions $L1, L2$ and $L3$ by an arbitrary secondary structure. The best energy for $L1$ is obtain by finding the best combination of two segments:

$$\mathcal{E}^1_{i,j} = \min \begin{cases} F^{L1}_{i,k-1} + F^{L1}_{k,j} + \mathcal{L}1_C \\ F^{L1}_{i+1,k-1} + F^{L1}_{k,j} + pkd^3_{i,i+1} + \mathcal{L}1_C \\ F^{L1}_{i,k-1} + F^{L1}_{k,j-1} + pkd^5_{j,j-1} + \mathcal{L}1_C \\ F^{L1}_{i+1,k-1} + F^{L1}_{k,j-1} + pkd^5_{j,j-1} + pkd^3_{i,i+1} + \mathcal{L}1_C \end{cases} \tag{3.73}$$

$$F^{L1}_{i,j} = \min \begin{cases} C_{i,j} + \mathcal{L}1_I \\ C_{i+1,j} + d^5_{i+1,j,i} + \mathcal{L}1_I \\ C_{i,j-1} + d^3_{i,j-1,j} + \mathcal{L}1_I \\ C_{i+1,j-1} + d^5_{i+1,j-1,i} + d^3_{i+1,j-1,j} + \mathcal{L}1_I \\ F^M_{i+1,j} + \mathcal{L}1_B \\ F^M_{i,j-1} + \mathcal{L}1_B \\ \min(F^M_{i,k-1} + F^M_{k,j}) \end{cases} \tag{3.74}$$

The constants $\mathcal{L}1_C, \mathcal{L}1_I$ and $\mathcal{L}1_B$ have the same meaning as in the multi loop recursions. $L2$ and $L3$ can be treated alike, but with different constants. Coaxial stacking is accomplished when we replace the functions E_{L1} and E_{L2} in **equation 3.69, 3.70** with the entries from the new recursions. The linear ansatz is an elegant method to calculate the energy of the best structure, but we do not know which distance this structure is able to bridge. Therefore the information about the stack-sizes $S1$ and $S2$ becomes useless. Subsequently building blocks can be constructed in a more flexible way, utilizing the linear ansatz. No further structural restrictions are necessary, building blocks are allowed to contain multi-loops and even other pseudoknots. The required recursions are outlined in detail in [51].

3.7 Suboptimal Folding

Suboptimal structures are often of biological interest and in a more general way, we need the suboptimals to understand the folding landscape of a given sequence. The first methods to calculate suboptimals where the so called combinatorial algorithms. Their main advantage is, that they are not restricted to a nearest neighbor energy model and are able to deal with tertiary interactions like pseudoknots. However, combinatorial algorithms easily get combinatorial problems when sequences become longer. A problem that can only be attenuated with more or less rough approximations which consequently lead us even farther away from a complete solution.

Fortunately the dynamic programming approach gives us access not only to the most stable structure. When we exploit the backtracking process, we are able to calculate all suboptimal structures within a given energy band. The first complete suboptimal folding algorithm was introduced by Wuchty *et al.* [72]. In contrast to previous dynamic programming approaches (Zucker *et al.*), this algorithm is able to find the complete set of structures within an energy band, not only a subset. In this section we give a detailed description how the algorithm works and how it is extended to find h_0-pseudoknots.

3.7.1 Waterman's Algorithm

Waterman and Byers suggested a dynamic programming algorithm to find all solutions in the neighborhood of an optimum . The purpose of the "shortest path problem" is to locate the shortest path from node 1 to node N in an acyclic network of N nodes and A arcs (e.g. RNA secondary structure in linked diagram representation). Each arc (i, j) has an associated weight $t(i, j)$. Nodes i are labeled with $f(i)$, the length of the shortest path from node i to node N. Provided Bellman's insight of optimality "sub paths of optimal paths

are themselves optimal" the recursion

$$f(i) = \min\{t(i,j) + f(j) : (i,j) \text{ an arc}\} \tag{3.75}$$

follows. The idea is, that to reach i from N, the last step is from some node j. The node j must be reached in an optimal manner, if j is an optimal path from N to i. Note, that $f(N) = 0$ is required to start the recursion. So far nothing else than the procedure of dynamic programming and backtracking was described. The new algorithm requires an interval ϵ above the optimal length $f(1)$ from the user. All paths less than or equal to the quantity $f(1)+\epsilon$ should then be found by the algorithm. The node labels $f(j)$ are found by working backwards from node N until node 1 is labeled. The new algorithm then performs a depth-first search with stacking, starting at node 1 and continuing until all near-optimal paths are found. Consider a node x not equal to the destination. Some path P with cumulative distance d led to node x from node 1. The test for entry of the arc (x, y) and distance d onto the stack now takes the general form for all $(x, y) \in A$

$$d + t(x,y) + f(y) \leq f(1) + e, \tag{3.76}$$

where d is the cumulative distance to node x from node 1 by path P (not necessarily the shortest path!), $t(x, y)$ is the distance from node x to node y, and $f(y)$ is the optimal remaining distance to node N from node y. The algorithm constructs a path P of length d from node 1 to node N. Then P and d are output and the stack is examined to see, if other near-optimal paths exist. Hence the algorithm performs a last in, first out or depth first search.

3.7.2 Suboptimal Maximum Matching

In order to show the general design of the *subopt*-algorithm we first apply Waterman's concept to the maximum matching problem. Maximum matching is well suited to sever as a model to exemplify the data-structure and search strategy, thus to keep it simple we left out the pseudoknots.

The suboptimal maximum matching backtracking finds all structures whose number of base pairs lies within $PS_{1,n}$ and $PS_{1,n}-\delta$. Starting with the segment $[1,n]$ the combination of segments $PS_{i,k-1}$ and $PS_{k+1,n-1}$, which fulfill this condition is found. The found combination of intervals $[1,k-1]$ and $[k+1,n-1]$ is written to an interval stack. Also the found base pair (l,n) is written to a separate base pair stack. Both stacks are contents of a so called state which is written to the state stack. In the next round of the algorithm the last state is taken from the state stack and the last interval, in general $[i,j]$, from the interval stack within that state (again last-in, first-out). Again within the interval $[i,j]$ those combinations of $PS_{i,k-1}$ and $PS_{k+1,n-1}$ are traced, whose number of base pairs lies within $PS_{1,n}$ and $PS_{1,n}-\delta$. However, this time also the already found base pair and the best possible number of base pairs of the intervals remaining on the stack denoted by $PS_{p,q}$ must be taken into account, so that the condition reads as

$$N_{bp} + [PS_{i,l-1} + 1 + PS_{l+1,j-1}] + \sum_{p,q} PS_{p,q} \leq PS_{1,n} - \delta \tag{3.77}$$

in analogy to

$$d + t(x,y) + f(y) \leq f(1) + \epsilon \tag{3.78}$$

N_{bp} denotes the number of all already found base pairs and p,q the several yet unconsidered intervals remaining on the interval stack. The state containing the interval stack the interval $[i,j]$ was taken from is copied and the new found base pair and intervals are written to the related stack within the state. The state is pushed back to the state stack. That happens to every new found combination of segmentations, which accomplish the condition outlined above. If no base pair, which accomplish the condition, can be found, the remaining state is pushed back to the state stack. The iteration goes on by taking out the first state of the state stack and following the first interval of the interval stack. If the interval stack is empty, a solution i.e. a structure, is found, and the state is skipped. The iteration continues until no state remains on the stack.

3.7.3 Suboptimal mfe-Folding

Modifications for the mfe-Matrix Filling

For the sake of an efficient implementation we introduce useful modifications to the matrix filling part of the algorithm, affecting the energy model for the dangles as well as the multi loop handling. In the *mfe* algorithm only unpaired bases are allowed to dangle and in case of two alternative dangle positions the one yielding the lower free energy contribution is chosen. In the *subopt*-dangle model every potential dangling base in a multi loop or exterior region contributes to the free energy, regardless if it is paired or not. Therefore we can simplify the recursions for the F array to something like:

$$F_{i,j} = \min \begin{cases} C_{i,j} + d^5_{i,j,i-1} + d^3_{i,j,j+1} \\ PK_{i,j} + pkd^5_{i,i-1} + pkd^3_{j,j+1} \\ F_{i,j-1} \\ \min_{i<k\leq j}(F_{i,k-1} + C_{k,j} + d^5_{k,j,k-1} + d^3_{k,j,j+1}) \\ \min_{i<k\leq j}(F_{i,k-1} + PK_{k,j} + pkd^5_{k,k-1} + pkd^3_{j,j+1}) \end{cases} \tag{3.79}$$

For the extended backtracking process it is necessary to introduce another auxiliary array F^{M1} for the multi loop decompositions. In the case of *mfe*-multi loop backtracking we tried to find the decomposition that yielded the best energy. A closer look at the recursions for the F^M array reveals that there might be a lot of possible decompositions giving the same energy and the same structure. This is not a problem, because in *mfe*-backtracking we take the first of the best decompositions. The *subopt*-backtracking has to avoid identical multi loop decompositions by introducing an additional array F^{M1}:

$$F^{M1}_{i,j} = \min\{C_{i,l} + \mathcal{M}_{\mathcal{B}}(j-l) + d^5_{i,l,i-1} + d^3_{i,l,l+1} + \mathcal{M}_{\mathcal{I}}\} \tag{3.80}$$

$F^{M1}_{i,j}$ denotes the minimum free energy of a multi loop segment containing only one stem situated at the upstream end of $[i,j]$ plus an arbitrary number of bases on the downstream side. Consequently we construct a multi loop with two different multi loop arrays:

$$C_{\substack{i,j \\ Multi}} = \min\{F^M_{i+1,k-1} + F^{M1}_{k,j-1} + d^5_{i,j,j-1} + d^3_{i,j,i+1} + \mathcal{M}_{\mathcal{C}}\} \tag{3.81}$$

The recursion for F^M simplifies to:

$$F_{i,j}^M = \min \begin{cases} \min\{F_{i,k-1}^M + F_{k,j}^{M1}\} \\ \min\{F_{k,j}^{M1} + \mathcal{M}_{\mathcal{B}}(k-i)\} \end{cases} \tag{3.82}$$

The first element gives a multi loop segment with at least two stems, the second one yields a segment with only one stem at an arbitrary position.

Extended Backtracking

Waterman's concept can be applied to thermodynamic RNA folding in order to find all suboptimal structures within a given energy range above the *mfe*-energy. To that end, we modify condition 3.77 to something like

$$E_f + E_{i,j} + \sum_{k,l} E_{k,l}^{min} \leq E_{1,n}^{min} + \delta \tag{3.83}$$

where E_f is the summed energy of all already found substructures. $E_{i,j}$ denotes the energy of the considered segment $[i,j]$ and $\sum E_{k,l}^{min}$ is the best possible energy of all remaining not investigated segments. $E_{1,n}^{min}$ is the optimal energy, whereas δ is the given energy range.The algorithm uses the same data structure like in the maximum matching backtracking case. All found base pairs and segments are written to separate base pair and interval stacks, which belong to a particular state. The states are written to a state stack.

Starting with a given segment $[i,j]$ we find all possible ways to produce a $F_{i,j}$ fulfilling the condition:

$$E_f + F_{i,j} + \sum_{k,l} E_{k,l}^{min} \leq F_{1,n}^{min} + \delta \tag{3.84}$$

according to **equation 3.79**:

$$E_f + F_{i,j-1} + \sum_{k,l} E_{k,l}^{min} \leq F_{1,n}^{min} + \delta \tag{3.85}$$

gives a new interval $[i, j-1]$,

$$E_f + F_{i,k-1} + C_{k,j} + d_{k,j,k-1}^5 + d_{k,j,j+1}^3 + \sum_{k,l} E_{k,l}^{min} \leq F_{1,n}^{min} + \delta \tag{3.86}$$

gives a base pair (k, j) and two new segments $[i, k-1]$ and $[k, j]$ for each k. The pseudoknots have to meet a similar condition:

$$E_f + F_{i,k-1} + PK_{k,j} + pkd^5_{k,k-1} + pkd^3_{j,j+1} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.87}$$

yielding also two segments $[i, k-1]$ and $[k, j]$ for each k. A segment $[i, j]$ that is formed by a base pair needs further backtracking according to **equation 3.42**. First we check which stacks, bulges and interior loops fulfill

$$E_f + C_{p,q} + \mathcal{I}(i, j, p, q) + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta, \tag{3.88}$$

then we look if a hairpin loop meets the condition

$$E_f + \mathcal{H}(i, j) + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta, \tag{3.89}$$

and at last we find out which unique multi loop decompositions accomplish

$$E_f + F^M_{i+1,k-1} + F^{M1}_{k,j-1} + d^5_{i,j,j-1} + d^3_{i,j,i+1} + \mathcal{M}_\mathcal{C} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.90}$$

The backtracking in the F^M and F^{M1} on arbitrary interval $[i, j]$ starts with the conditions

$$E_f + F^M_{i,j-1} + \mathcal{M}_\mathcal{B} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.91}$$

and

$$E_f + F^{M1}_{i,j-1} + \mathcal{M}_\mathcal{B} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.92}$$

If these conditions are fulfilled, no base pair is found and the $3'$ end is nibbled creating the segment $[i, j-1]$. The condition

$$E_f + C_{i,j} + d^5_{i,j,i-1} + d^3_{i,j,j+1} + \mathcal{M}_\mathcal{I} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.93}$$

gives the first base pair (i, j) of an internal branch of a multi loop, the segment $[i, j]$ proceeds backtracking at **equation 3.88**. Using the F^M array multi loop decompositions can be performed accomplishing the conditions

$$E_f + F^M_{i,k} + C_{k+1,j} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_{\mathcal{I}} + \sum_{k,l} E^{min}_{k,l} \tag{3.94}$$

$$\leq F^{min}_{1,n} + \delta \tag{3.95}$$

if the segment $[i, j]$ contains more than one stack. This condition finds a base pair $(k+1, j)$ and generates the segments $[i, k]$ (enters at **equation 3.84**) and $[k + 1, j]$ (enters at **equation 3.88**). If the considered segment contains only one stack, the condition turns to

$$E_f + C_{k+1,j} + d^5_{k+1,j,k} + d^3_{k+1,j,j+1} + \mathcal{M}_{\mathcal{I}} + \mathcal{M}_{\mathcal{B}}(k - i + 1) \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.96}$$

finding the base pair $(k+1, j)$ and the segment $[k+1, j]$. Both conditions hold for $i < k < l$.

The pseudoknot segment encountered in **equation 3.87** requires backtracking in the building block arrays.

$$E_f + F_{i,k-1} + B^u + B^d + \mathcal{E}_{L1} + \mathcal{E}_{L2} + Coax + pkd^5_{k,k-1} + pkd^3_{j,j+1} + \sum_{k,l} E^{min}_{k,l} \leq F^{min}_{1,n} + \delta \tag{3.97}$$

resulting in a set of building block pairs. Subsequently we produce suboptimal building blocks still fulfilling **equation 3.97**. Depending on the building block model in use, we use to that end **equations 3.49-3.51**, **equations 3.52-3.54** or **equations 3.55-3.57**.

4 Kinetic Folding Algorithm

It has been already mentioned in chapter 3, that beside the dynamic programming algorithms, the kinetic folding approach is an essential tool to investigate the nature of RNA folding. During the cause of a folding process, an RNA molecule may get trapped in a misfolded state or there may be two meta-stable states. There is no dynamic programming algorithm that gives direct access to this information.

Most kinetic folding algorithms start from some initial structure (e.g. the open chain) and progress, by incorporation of whole helixes, through a series of nearly optimal structures to the most probable one at the end of the folding process. The various methods differ mainly in how the next structure is selected. However, they apply more or less strong heuristic assumptions about the transition rates and are therefore not suitable to reconstruct folding pathways.
In this work, we follow a much more elementary approach introduced by Christoph Flamm *et al.* [12, 13]. To get a better understanding why this approach is the most elementary possible, with actually no heuristic involved, we need to give some explanations.

4.1 Conformation Space, Move Set and Energy Landscape

The *conformation* or *structure space* C comprises all secondary structures that are compatible with a given sequence. A *move set* is a set of rules, which defines how to convert one structure into another, i.e. how to move within the conformation space C. To be more precise, the *move set* is an order

relation on C, defining *adjacency* between the elements of C. It fixes the possible conformational changes that can take place in a single step during the simulation of folding and thus defines the topology of the conformational space. The following properties are important for move sets:

1 Each move has an inverse counterpart. At thermodynamic equilibrium the quotient of forward and backward reaction rates must give the microscopic equilibrium constant.

2 The outcome of an operation always leads to an element of the underlying state space.

3 The move set has to be ergodic. In other words starting from an arbitrary point of the state space every other point must be reachable by a sequence of legal operations.

4 Every move set defines a metric an the underlying state space.

The most elementary move set on the RNA secondary structure space, which fulfills all properties, consists of closing and opening of a single base pair (i, j). It is convenient to represent the structure space and its topology as a graph. Each vertex corresponds to a secondary structure, and the edges connect structures which are inter-convertible with one move. Thus secondary structures connected with an edge are called neighbors.

A *value landscape* is obtained by taking the graph as support of a function that assigns a value to every conformation. If this value is the free energy of the corresponding structure, the landscape is called *energy landscape* or *folding landscape*.

The Shift Move

Experimental data on a mechanism called *defect-diffusion* suggest a minor extension to the elementary move set. For instance the base of a bulge loop present i a helix, will be subject to a rapid base pair formation and dissociation

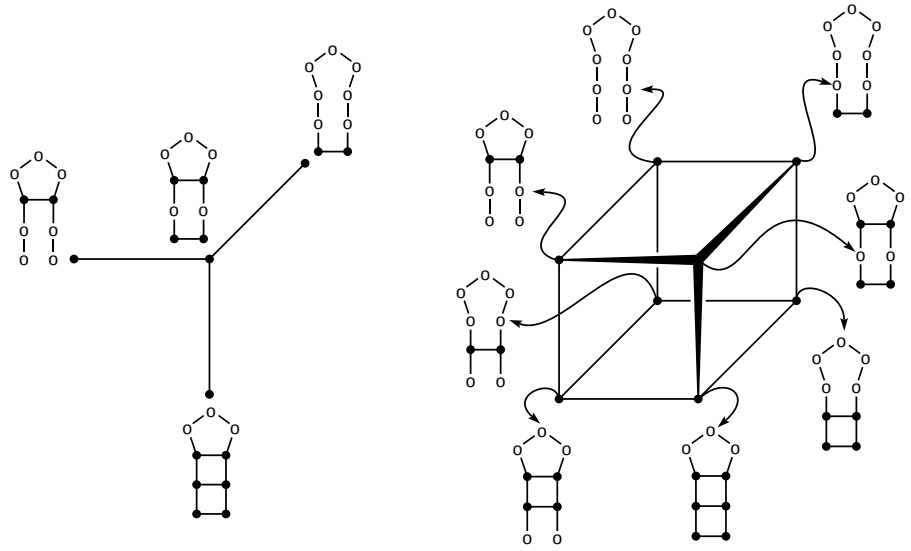


Figure 4.1: One move neighborhood of a vertex in the conformation space (l.h.s) and its embedding in the graph representing the conformation space (r.h.s.) for a short RNA molecule which can exhibit 3 base pairs.

process. According to experiments, this kind of chain sliding mechanism occurs very fast. In the framework of the elementary move set, the defect-diffusion is in most cases not a favorable process. To facilitate chain sliding, we extend the elementary move set by a further move called *shift-move*. The shift converts an existing base pair (i, j) into a new base pair (i, k) or (l, j) in one step.

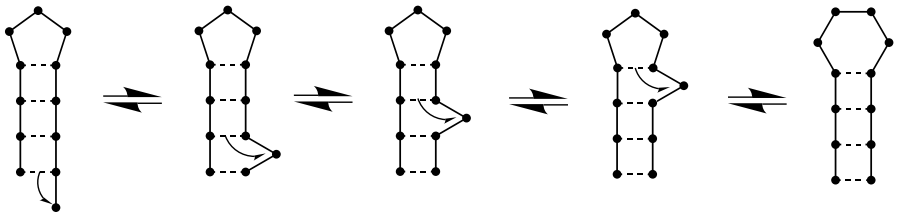


Figure 4.2: Defect diffusion: The bulge can easily migrate along the helix. Shift moves are indicated by arrows.

The Canonic Move Set

The *canonic* move set avoids per definition moves that produce a structure

with an isolated base pair. A base pair is isolated if it is not part of a stack, or in other words, an isolated base pair is a stack with stack-size 1. Thus the canonic move set works on a much smaller structure space, namely the space of all structures with minimal stack-size 2. Consequently the number of neighbors of a given structure is reduced significantly. The algorithm gains efficiency not only because of this combinatorial effect. Additionally, the canonic move set takes advantage of energy-model properties, because stem nucleation moves are now energetically more favorable. Moreover assumptions about the transition rates of nucleation moves can be made easily.

4.2 The Algorithm and its Implementation

The simulation is based on a continuous time Monte Carlo method proposed by Daniel Gillespie [18]. The rate constant which characterize the transition from one structure S_i to a neighboring structure S_j is calculated by a symmetric rule introduced by Kyori Kawasaki [30]. It is symmetric because it takes gradients for both uphill and downhill steps into account (in contrast to the Metropolis rule [38]).

$$k_{ij} := e^{-\frac{\Delta G}{2kT}} \tag{4.1}$$

Note that the free energy difference ΔG between the two states i and j must be divided by $2kT$ to get the detailed balance right. The Kawasaki dynamics approaches the Boltzmann distribution at equilibrium because it satisfies microscopic reversibility.

The method provides an internal clock to measure time. The time spent in a certain state is inversely proportional to the total flux Φ leading away from this state. If Φ is small, as for example at the bottom of a deep local minimum, the internal clock is advanced by a big time increment. For each step, the rate constants from the current state to all its neighbors are computed. Then, the time is advanced by an appropriate time increment adjusted to the sum of the rate constants. Finally, the current state is replaced by a state chosen from

the set of neighbors. The algorithm works in the following way:

0. **Initialization.**
   (a) Set the time variable $t = 0$ and the "stopping time" t_{stop}.
   (b) Specify the start structure and initialize the current structure S_{cur} with the start structure.
   (c) Specify and store the stop structure S_{stop}.
1. Generate the set of legal neighbor structures $\{S_n\}$ from S_{cur}.
2. Calculate all the reaction channel weights $R^{(\alpha)}_{curr}$ and the total flux $\Phi_{cur} = \sum_\alpha R^{(\alpha)}_{cur}$. Afterwards normalize the $R^{(\alpha)}_{cur}$'s.
3. Draw two random numbers $r_1,\ r_2 \in [0, 1]$ from a uniform number generator.
4. Cumulatively adding the successive values $R^{(1)}_{cur},\ R^{(2)}_{cur},\ ...$ until their sum is observed to equal or exceed r_1. Choose the structure with the index of the last term added to the sum as the new S_{cur}.
5. Calculate the time increment $t_{inc} = \frac{1}{\Phi_{cur}} \cdot \ln\left(\frac{1}{r_2}\right)$ and advance the clock $t = t + t_{inc}$.
6. If $t > t_{stop}$ or if S_{cur} equals S_{stop}, terminate the calculation; otherwise, return to Step 1.

By following the above procedure from time 0 to time t, only one possible realization of the stochastic process is obtained. In order to get a statistically complete picture of the temporal evolution of the folding of a RNA molecule, several independent realizations or "runs" have to be carried out. Each run must start with the same initial conditions and should proceed to the same time t.

4.3 How to Generate the Neighbors

The second requirement for the move set, that every move has to yield an element of the state space, is always fulfilled if we open a base pair. If we close a base pair, it depends on the current state if the move is legal. In the case of secondary structures without pseudoknots, every move which yields two overlapping base pairs is illegal. Consequently closing a base pair is only allowed within a loop or on the external regions (which can be viewed as an open loop). This implies a data structure derived from the rooted ordered tree representation of secondary structures. All loops are linked together in a tree-like fashion. In **figure 4.3** this links are indicated with dashed lines, bases which belong to the same loop are connected with solid lines, unpaired bases are drawn in red. Additionally a virtual root (blue), which is the closing base pair of the external region is introduced. It is obvious to which loop an unpaired base belongs. In the case of a base pair we assign the 5' base to the loop which engulfs the base pair, whereas the 3' base is assigned to the loop closed by the base pair. Producing all neighbors is now straight forward, we just have to visit every loop and produce all possible base pairs with the unpaired bases that belong to the loop. Then the loops closing base pairs are opened to complete the set of neighbors.

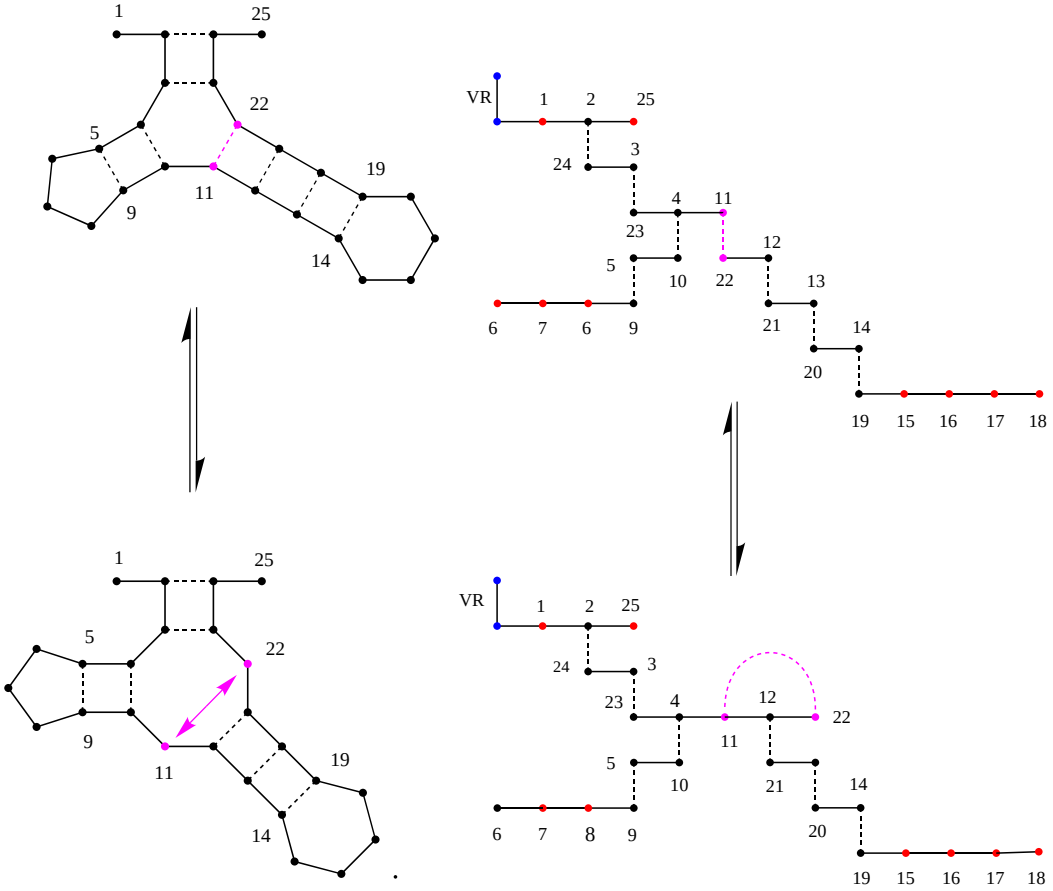


Figure 4.3: Data-structure for Simple Secondary Structures showing the changes introduced by a elementary move.

If L is the set of all unpaired bases of a loop, then

$$N_c(L) = \{(i,j)|i,j \in L, i + hm < j, \Pi_{\sigma_i,\sigma_j} = 1\} \tag{4.2}$$

is the set of all possible closing neighbors within L.

4.3.1 The Pseudoknot Neighbors

However, a tree representation of a secondary structure with h-pseudoknots is not possible. Instead we have to implement a more complex data structure. In order to explain the extended algorithm, it is convenient to distinguish two stages:

1. The "first contact" - produces the first base pair between the hairpin loop and the external region
2. Generate the neighbors of an already existing h-pseudoknot

Let E_u (E_d) be the set of all upstream (downstream) unpaired bases, which are adjacent to a multi-loop free component and let L_H be the set of all unpaired bases of the components hairpin loop. Then

$$N_c(PK) = \{(i,j)|PKR(i,j) = 1, \Pi_{\sigma_i,\sigma_j} = 1, \left\{ \begin{array}{l} i \in E_u, j \in L_H \\ i \in L_H, j \in E_d \end{array} \right\}\} \tag{4.3}$$

is the set of all "first-contact" base-pairs. PKR is a function which indicates if a base pair is in accordance with a given h-pseudoknot restriction. Thus the loop closed by the virtual root produces not only neighbors among its unpaired bases, but checks also multi-loop free components for potential pseudoknot base pairs. To that end, all components which are consistent with the h-pseudoknot restrictions are specially labeled. If such a labeled base pair is encountered the algorithm finds the corresponding hairpin-list and generates the first contacts.

When the first contact is established, for example with E_d, then both sets are split into two sets,

$$L_H \rightarrow L1, L2 \tag{4.4}$$

$$E_d \rightarrow L2, E'_d \tag{4.5}$$

resulting the following conditions for closing neighbors:

$$N_c(PK) = \{(i,j)|PKR(i,j) = 1, \Pi_{\sigma_i,\sigma_j} = 1, \left\{ \begin{array}{l} i \in L1, j \in E'_d \\ i \in L3, j \in L2 \\ i \in L1, j \in L3 \\ i \in E_u, j \in L2 \end{array} \right\}\} \tag{4.6}$$

An upstream first contact splits L_H and E_d sets into

$$L_H \rightarrow L3, L2 \tag{4.7}$$

$$E_d \rightarrow E'_u, L1 \tag{4.8}$$

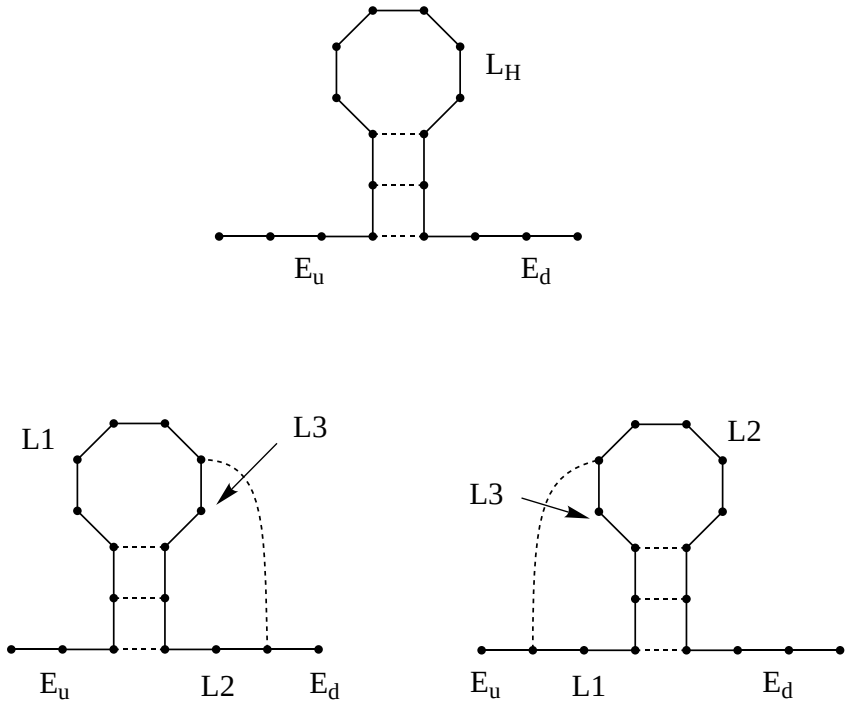


Figure 4.4: The "first contact" splits L_H and E_u (E_d) into two intervals

yielding the conditions

$$N_c(PK) = \{(i,j)|PKR(i,j) = 1, \Pi_{\sigma_i,\sigma_j} = 1, \left\{ \begin{array}{l} i \in E'_u, j \in L2 \\ i \in L1, j \in L3 \\ i \in L3, j \in L2 \\ i \in L1, j \in E_d \end{array} \right\}\} \quad (4.9)$$

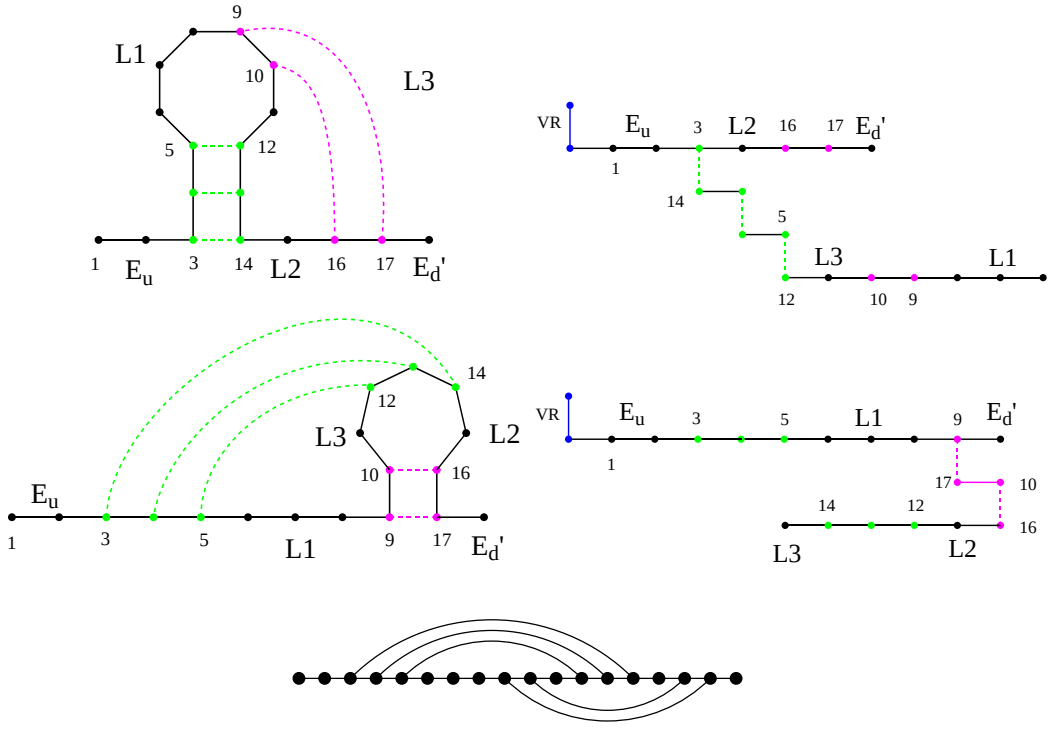


Figure 4.5: Extended data-structure for h-pseudoknots.

All other base pairs that involve $L1, L2$ or $L3$ are not allowed, as well as base pairs between E'_u and E_d or E_u and E'_d. In contrast to simple secondary structures, base pairs are produced between to unpaired regions and not within a loop. The data structure which provides the appropriate pairs of unpaired regions is not very different from the previously discussed. For a better understanding we draw the pseudoknot in liked diagram representation.

First we traverse the data-structure on page 1 (upper half-plain) and find that the hairpin-loop and E_d are connected by base pairs. Subsequently the algorithm finds the corresponding pairs of segments $(L1, L2)$, $(E_u, l2)$ and produces all allowed base pairs. Consequently page 2 (lower half-plain) is handled in the same way finding base pairs between segments $(L1, E'_d)$ and $(L3, L2)$.

5 Computational Results

In this chapter we apply the whole set of previously discussed algorithms to give some examples how the RNA-energy landscape is influenced by h_0-pseudoknots. To that end, we utilize the *suboptimal* and the *neighbor-generating* algorithm and make them part of the *barrier tree*-algorithm. The barrier tree is a strongly reduced but very practical representation of the energy landscape. It is obtained by flooding the energy landscape with suboptimal structures, beginning with the mfe-structure (see **figure 5.1**). Whenever the lowest saddle point between two valleys is reached, the two vertices representing the two local minima are connected with the saddle point vertex. The two edges are weighted according to the energy difference between the corresponding vertices. Thus the barrier tree provides information about the energetically lowest barrier between two connected minima.

5.1 RNA Conformational Switches

RNA conformational switching it thought to be fundamental to a number of biological processes, including translational regulation [3, 11, 49], protein synthesis [71], and mRNA splicing [35]. Typically the two competing secondary structures show mutually exclusive base pair patterns but nearly equal free energy. Thus the energy barrier between the to structures is very high and consequently the transition rates are very low. A transition process that may take hours to accomplish is not well suited as a biological regulatory switch. Indeed, experiments suggest the existence of an alternative folding path that is faster and does not involve high energetic intermediates [32]. With the help of a short model-RNA sequence, we can show that h_0-pseudoknots may facilitate the transition between the conformers.

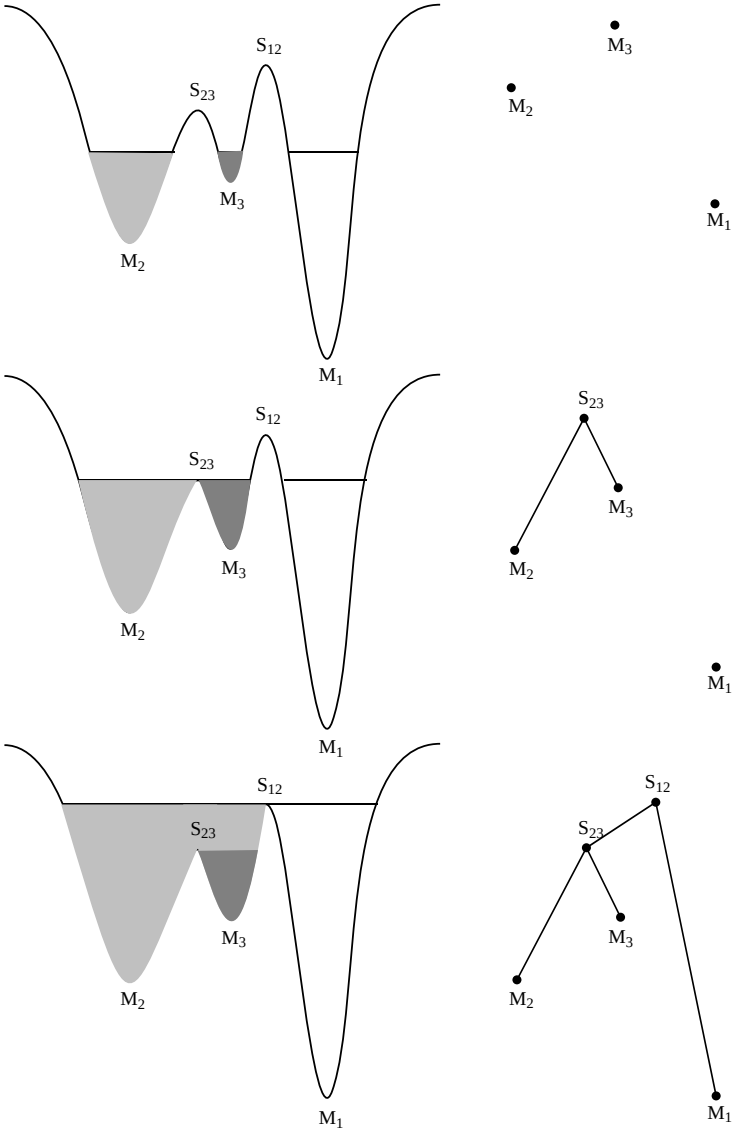


Figure 5.1: Flooding the Energy Landscape

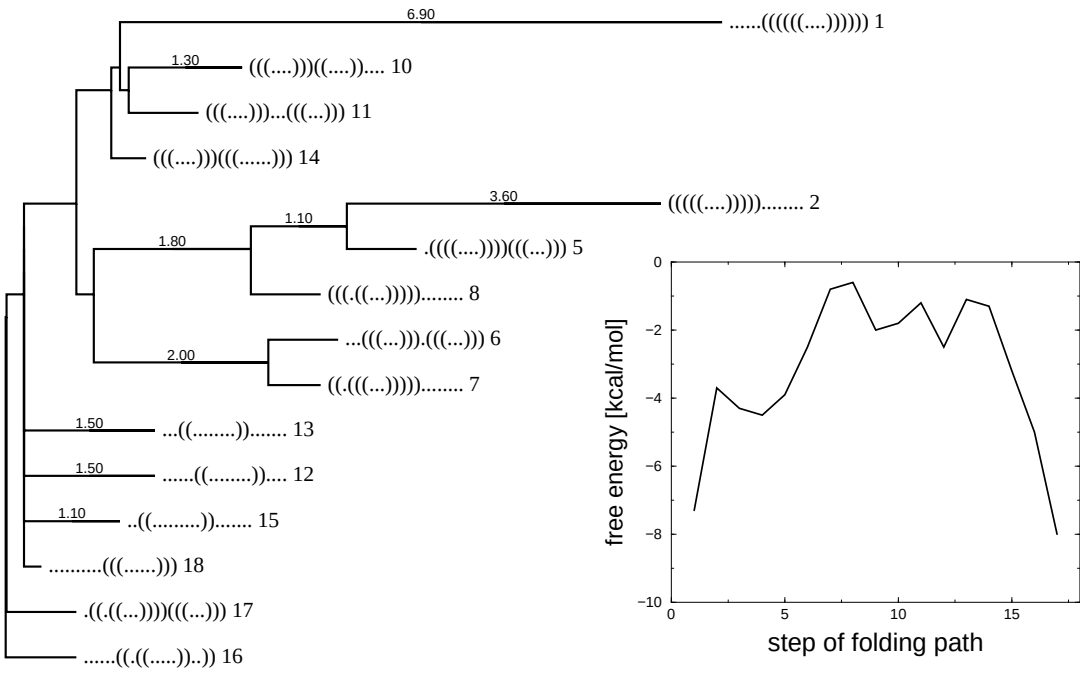


Figure 5.2: Barrier tree without h-pseudoknots

Step	Structure	Energy / kcal/mol	Type
1	(((((....))))).....	-7.30	L 2
2	.((((....)))).....	-3.70	S
3	.((((....))))((.....))	-4.30	I
4	.((((....))))(((...)))	-4.50	L 5
5	..(((....))).(((...)))	-3.90	I
6	..((......)).(((...)))	-2.50	I
7	............(((...)))	-0.80	I
8	((......))...(((...)))	-0.60	S
9	(((....)))...(((...)))	-2.00	L 11
10	(((....)))...((.....))	-1.80	I
11	(((....)))..........	-1.20	S
12	(((....)))((....))....	-2.50	L 10
13	((......))((....))....	-1.10	S
14	..........((....))....	-1.30	I
15	.........(((....)))...	-3.20	I
16	.......(((((....))))).	-6.20	I
17	......((((((....))))))	-8.00	L 1

Table 5.1: Folding path from the metastable structure 2 into the mfe-structure 1. The "Type" column indicates if a structure is a local minimum L, an intermediate I or a saddle point S.

Figure 5.2 shows the barrier tree without pseudoknots. A big barrier of 6.5 kcal/mol separates structure 1 and 2. **Table 5.1** depicts the circumstantial folding path from the metastable structure 2 to the mfe-structure 1. The melting of the stem in 1 is facilitated by a stabilizing stem which is nucleated in step 3. At step 7 the original stem is melted and in step 8 a second intermediate stem is introduced to melt the previous auxiliary stem which is incompatible with the target structure. At step 12 the target stem is nucleated and subsequently the auxiliary stem is melted while the mfe-structure forms. The peak of the whole path is reached at step 8, 7.4 kcal above the target structure. Intermediates from step 6 to 11 do not share a single base pair with the start- or target structure.

The situation changes totally when we consider h_0-pseudoknots.

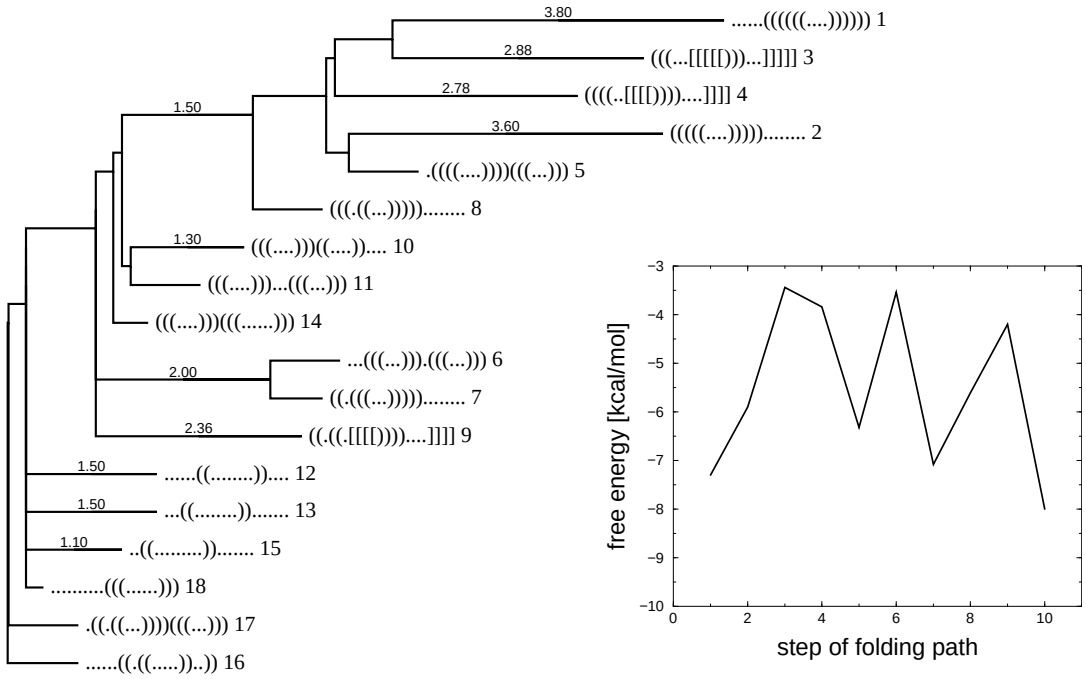


Figure 5.3: Barrier tree with h-pseudoknots

Step	Structure	Energy / kcal/mol	Type
1	`(((((....)))))........`	- 7.30	L 2
2	`((((......))))........`	-5.90	I
3	`((((...[[.))))....]].`	-3.44	S
4	`((((...[[[))))....]]].`	-3.84	I
5	`((((..[[[[))))....]]]]`	-6.32	L 4
6	`(((...[[[[.)))....]]]]`	-3.54	S
7	`(((...[[[[[)))...]]]]]`	-7.08	L 3
8	`((....[[[[[.))...]]]]]`	-5.61	I
9	`......((((.....))))).`	-4.20	S
10	`......((((((....))))))`	-8.00	L 1

Table 5.2: Folding path from the metastable structure 2 into the mfe-structure 1. The "Type" column indicates if a structure is a local minimum L, an intermediate I or a saddle point S.

The folding path in **table 5.2** has its peak at step 3 only 3.86 kcal above structure 2. It occurs when the pseudoknot, and at the same time the target stem is nucleated. After the nucleation the upstream stem is melted while the downstream stem is formed. All intermediate structures share base pairs with the start and/or the target structure. The more rugged profile is due to the pseudoknot energy model.

For a direct comparison of both barrier trees we omit the local minima 12, 13, 15,-16, 17 and 18 because they remain unchanged.

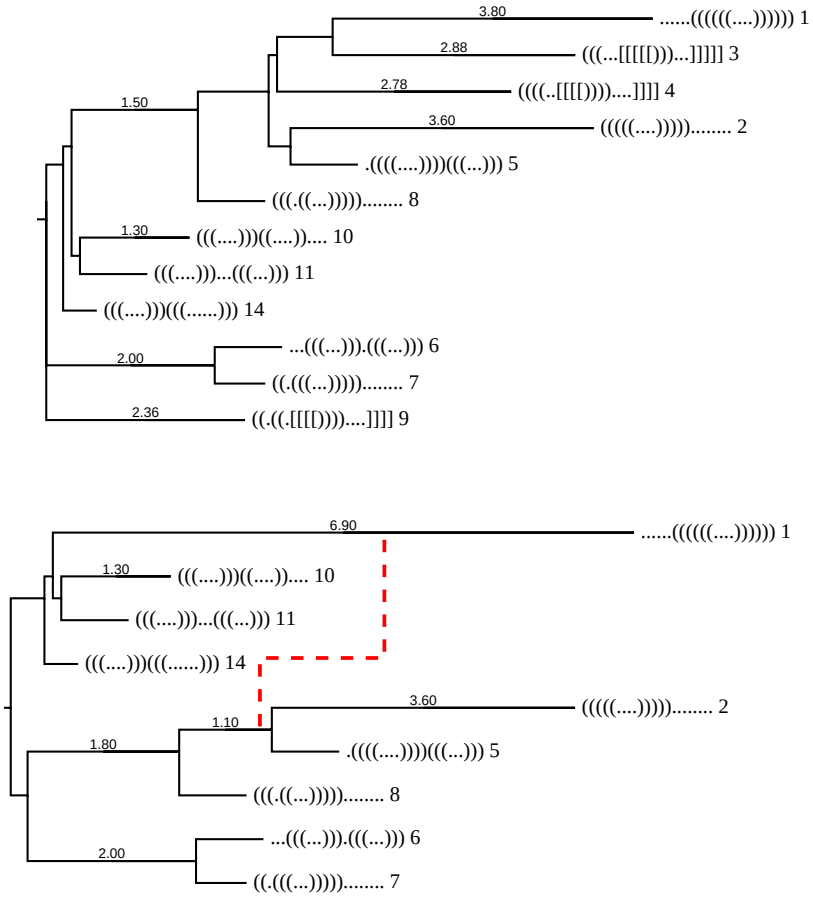


Figure 5.4: Comparison of both barrier trees. The red dashed line indicates where the pseudoknots introduce a shortcut.

The profiles of both folding paths provide a qualitative picture, however there are other alternative paths which also contribute to the overall transition rates. Thus it is useful to simulate the transition from structure 1 to structure 2 and vice versa, utilizing the kinetic folding algorithm. First passage times of 6000 trajectories were measured for the four different cases. The resulting curves confirm the assumptions inferred from the barrier tree.

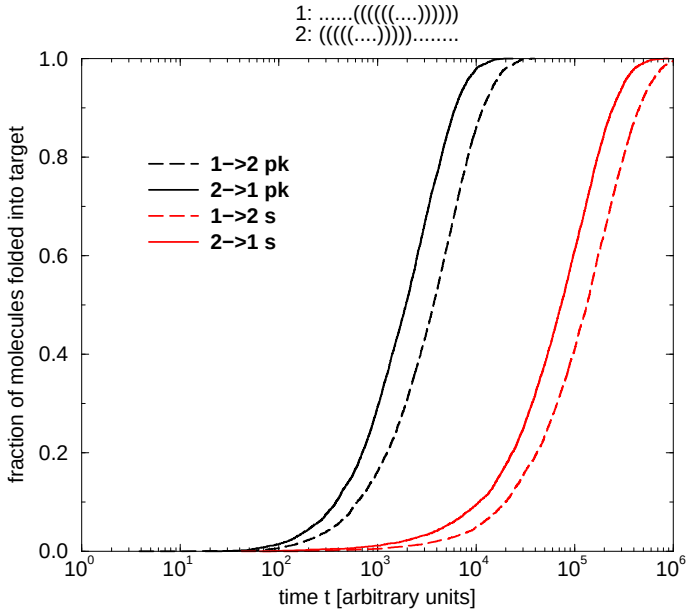


Figure 5.5: Transition kinetic of the mfe-structure and a metastable structure. Red lines: simple secondary structures. Black lines: secondary structures with pseudoknots

6 Conclusion and Outlook

The knowledge about the spatial conformation of functional RNA molecules is a crucial prerequisite to understand how they work.

In order to get a model that is theoretically and by computation easier to manage, the RNA secondary structure, an interesting class of contact structures, is introduced. Secondary structures provide a coarse graining of the 3D structure, by regarding base pair pattern only. Conventional secondary structures exclude overlapping base pair interactions by definition, and consequently reduce radically the number of possible base pair patterns. This strong restriction permits efficient algorithms to study RNA folding, but they totally ignore an important structural motif called RNA pseudoknots.

RNA pseudoknots mediate several biological functions, like translational and replicational control, others are necessary to form the reaction center in ribozymes. Therefore it is desirable to consider pseudoknots in theoretical models and prediction algorithms. The extension of the conventional secondary structure concept raises problems for the computational handling, caused by an enormous increase of possible base pair patterns and the fact that stereochemical constraints come into effect. To circumvent this problems, we focus on a strongly restricted type of pseudoknot - the hairpin-pseudoknot. Moreover the restrictions are motivated by the fact, that h-pseudoknot are by far the most abundant of all known pseudoknots, and that at least an approximate energy model exists. In particular dynamic programming algorithms become feasible in terms of time and memory demand due to the restrictions.
Dynamic programming was for a long time considered as incompatible with dynamic programming until Rivas and Eddy [51] published an algorithm capable to handle a multitude of different pseudoknot types. However, their method is rather prohibitive for longer sequences ($\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ mem-

ory) and neglects stereo-chemistry completely. The introduction of restricted h-pseudoknots reduces requirements to $\mathcal{O}(mn^3)$ time and $\mathcal{O}(mn^2)$ memory where m is a constant, depending on the structural freedom we approve to the h-pseudoknot. Consequently several valuable methods become realizable, like complete suboptimal folding. The complete suboptimal folding algorithm produces all secondary structures within a given energy interval above the most stable structure.

Pseudoknots are assumed to adopt an important role in several kinetically determined phenomena, like conformational switches. Thus an already existing method, introduced by Christoph Flamm *et.al.* [12,13] was adjusted. With the help of this rejection free Monte Carlo-type algorithm, folding trajectories can be calculated. A crucial component for this stochastic simulation is the choice of the move set for inter-converting secondary structures. The move set lays down the topology of the folding landscape by defining which secondary structures are neighbors of each other. It encodes the set of structural changes that RNAs can undergo at moderate activation energies. The algorithm uses the most elementary move set on the level of secondary structures, the closing and opening of a single base pair. A neighbor generating algorithm is consequently an integral part of this method. In fact it is the only extension necessary to include pseudoknots. For an efficient implementation a more flexible data-structure than for simple secondary structures was devised.

Moreover, the neighbor generating algorithm is utilized for the so called barriers algorithm, which provides information about the local minima of an energy landscape and how they are connected. To that end also suboptimal structures, produced by the previously mentioned complete suboptimal folding method, are needed.

The combination of all pseudoknot-adapted algorithms gives a comprehensive tool to study how pseudoknots change the folding landscape of RNA sequences. We could show, that pseudoknots can play an important role as intermediates in the folding path of conformational switches. They are able to lower the energy barrier between the two alternative conformations and accelerate transition rates significantly.

Restrictions also make sense for more complex pseudoknots, in fact they become even more important than in the case of h-pseudoknots. The missing energy model and the unsolved evaluation of stereo-chemical constraints are severe problems for the handling of complex pseudoknots. Dynamic programming algorithms can only be implemented efficiently on the basis of a simple energy model or strong restrictions. However until sufficient experimental results are available, it remains unclear if pseudoknot thermodynamics follows a simple-enough energy model.

A The Pseudoknot Database

In this section, a brief overview about the diversity of known pseudoknot base pair patterns is given. The reference for this overview is a database maintained by the Institute of Chemistry at the University Leiden [63]. There are 191 different pseudoknots in the database. 148 of them are h-pseudoknots, 13 are i-pseudoknots, 10 are i-pseudoknots with more that one stem in the pseudoknot loops, 4 are kissing hairpin complexes and 16 are more exotic pseudoknots. The h-pseudoknots are of special interest, because throughout this work we mainly focus on this most simple type of pseudoknot. Thus we give the frequencies of the occurring loop-sizes $L1, L2$ and $L3$ and stem-sizes $S1$ and $S2$.

The loop-size distribution of $L3$ shows, that almost all h-pseudoknots are actually h_0-pseudoknots, only 5 structures do not fit into this restriction. In the case of $L2$ four structure with big loop-sizes $(58, 75, 83, 261)$ where omitted because there are gaps in the database entry and they are very likely to form secondary structures. Stack-size $S1$ is wider distributed than $S2$, and loop-size $L2$ wider than $L1$. This corresponds well with the fact that $L2$ has to bridge $S1$ and $L1$ has to bridge $S2$. 35 of all h-pseudoknots exhibit one interior loop or bulge. 28 of them can be captured by a restriction where only symmetric interior loop of size 2 and bulges of size 1 are allowed.

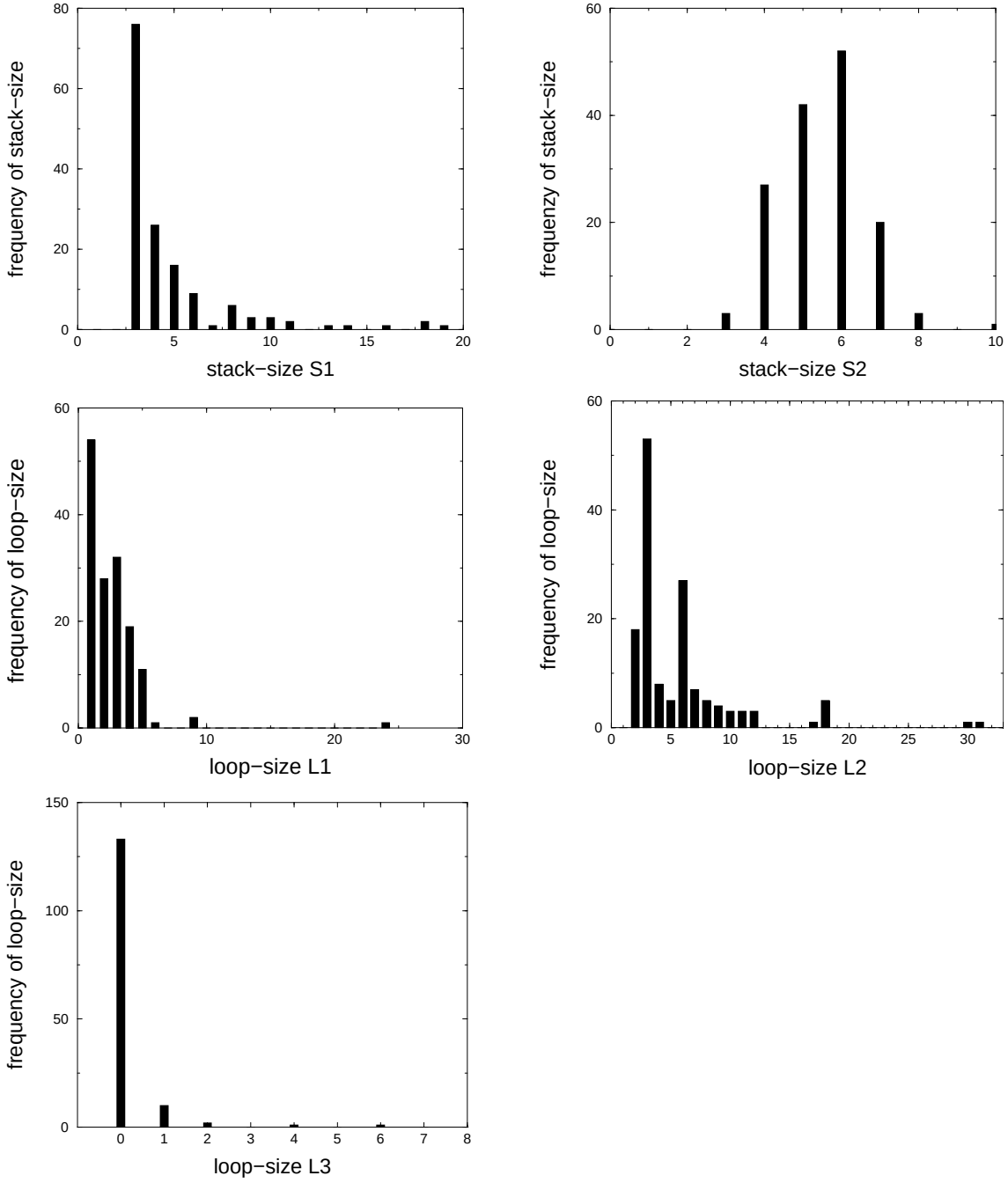


Figure A.1: Frequencies of loop- and stack-sizes of h-pseudoknots

Bibliography

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 1104:45–62, 2000.

[2] V. P. Antao and I. Tinoco. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetralooops. *Nucl.Acid.Res*, 20(4):819–824, 1992.

[3] P. Babitzke and C. Yanofsky. Reconstitution of bacillus subtilis trp attenuation in vitro with trap, the trp RNA-binding attenuation protein. *Proc Natl Acad Sci*, 1:133–137, 1993.

[4] I. Brierley, N. J. Rolley, A. J. Jenner, and S. C. Inglis. Mutational analysis of the RNA pseudoknot component of a coronavirus ribosomal frameshifting signal. *J. Mol. Biol*, 229:889–902, 1991.

[5] J. W. Brown. Structure and Evolution of Ribonuclease P RNA. *Biochemie*, 73:689–697, 1991.

[6] T. Cech. RNA as an enzyme. *Scientific American*, 11:76–84, 1986.

[7] M. Chamorro, N. Parkin, and H. E. Varmus. An RNA Pseudoknot and an Optimal Heptameric Shift Site Are Required for Highly Efficient Ribosomal Frameshifting on a Retroviral Messenger RNA. *Proc Natl Acad*, 89:713–717, 1992.

[8] E. B. T. Dam, C. W. A. Pleij, and L. Bosch. RNA Pseudoknots and Translational Frameshifting on Retroviral, Coronaviral and Luteoviral RNAs. *Virus Genes*, 4:121–136, 1990.

[9] B. Deiman and C. W. A. Pleij. Pseudoknots: A vital feature in viral RNA. *Semin Virol*, 8:166–175, 1997.

[10] J. D. Dinman, T. Icho, and R. B. Wickner. A -1 Ribosomal Frameshifting in a Double-stranded RNA Virus of Yeast Forms a Gag-Pol Fusion Protein. *Proc Natl Acad Sci U S A*, 88:174–178, 1991.

[11] G. Fayat, J. Mayaux, C. Sacerdot, M. Fromant, M. Springer, M. Grunberg-Manago, and S. Blanquet. Escherichia coli phenylalanyl-tRNA synthetase operon region. evidence for an attenuation mechanism. identification of the gene for the ribosomal protein l20. *JMB*, 3:239–261, 1983.

[12] C. Flamm. *Kinetic Folding of RNA*. PhD thesis, University Vienna, 1998.

[13] C. Flamm, W.Fontana, I. L. Hofacker, , and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[14] A. C. Forster and S. Altman. Similar Cage-shaped Structures for the RNA Component of All Ribonuclease P and Ribonuclease MRP Enzymes. *Cell*, 62:407–409, 1990.

[15] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[16] D. R. Gallie, J. N. Feder, R. T. Schmike, and V. Walbot. Functional Analysis of the Tobacco Mosaic Virus tRNA-like Structure in Cytoplasmic Gene Regulation. *Nucleic Acids*, 19:5031–5036, 1991.

[17] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[18] D. T. Gillespie. A general method for nummerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[19] C. Guerrier-Takada, K. Gardiner, T. M. N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[20] A. P. Gultyaev, F. van Batenburg, and C. W. A. Pleij. An approximation of loop free energy values of RNA h-pseudoknots. *RNA*, 5:609–617, 1999.

[21] E. S. Haas, D. P. Morse, J. W. Brown, J. F. Schmidt, and N. R. Pace. Long-range Structure in Ribonuclease P RNA. *Science*, 254:853–856, 1991.

[22] C. Haslinger and P. F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bul. Math. Biol.*, 1:1–33, 1998.

[23] L. He, R. Kierzek, J. SantaLucia, A. Walter, and D. Turner. Nearest-neighbour parameters for G-U mismatches. *Biochemistry*, 30:11124, 1991.

[24] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucleic acids research*, 12:67–74, 1984.

[25] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci., USA, Biochemistry*, 86:7706–7710, 1989.

[26] G. Joyce. Building the RNA world: evolution of catalytic RNA in the laboratory. In T. Cech, editor, *Molecular Biology of RNA. UCLA Symposium on Molecular and Cellular Biology*, pages 361–371. New York: Alan R.Liss, 1988.

[27] G. Joyce. Amplification, mutation, and selection of catalytic RNA. *Gene*, 82:85–87, 1989.

[28] G. F. Joyce. RNA evolution and the origins of life. *Nature*, 338:217–224, 1989.

[29] G. F. Joyce. The rise and fall of the RNA world. *The New Biologist*, 3:399–407, 1991.

[30] K. Kawasaki. Diffusion constants near the critical point for time-dependent Ising models. *Phys. Rev.*, 145:224–230, 1966.

[31] D. Konings and P.Hogeweg. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597–614, 1989.

[32] K. A. LeCuver and D. M. Crothers. Kinetics of an RNA conformationsl switch. *Proc. Natl. Acad. Sci.*, 91:3373–3377, 1993.

[33] R. B. Lyngso, M. Zucker, and C. N. S. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15:440–445, 1999.

[34] J. M. M. Wu and D. H. Turner. A periodic table of symmetric tandem mismatches in RNA. *Biochemistry*, volume 34::2304–11, 1995.

[35] H. D. Madhani and C. Guthrie. A novel base-pairing interaction between u2 and u6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, 5:803–817, 1992.

[36] R. Mans, C. Pleij, and L. Bosch. Transfer RNA-like Structures: Structure, Function and Evolutionarz Significance. *Eur J Biochem*, 201:303–324, 1991.

[37] R. Mans, M. H. V. Steeg, P. Verlaan, C. Pleij, and L. Bosch. Mutational Analysis of the Pseudoknot in the tRNA-like Structure of Turnip Yellow Mosaic Virus RNA. Aminoacylation Efficiency and RNA Pseudoknot Stability. *J Mol Biol*, 223:221–232, 1992.

[38] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092, 1953.

[39] F. Michel and E.Westhof. Modelling of the Three-dimmensional Architecture of Group I Catalytic Introns Based on Comparative Sequence Anaysis. *J Mol Biol*, 216:585–610, 1990.

[40] D. Moazed and H. F. Noller. Transfer RNA Shields Specific Nucleotides in 16S Ribosomal RNA from Attack by Chemical Probes. *Proc Natl Acad Sci U S A*, 47:985–994, 1986.

[41] S. Morse and D. E. Draper. Purine-purine mismatches in RNA helices: evidence for protona ted ga pairs and next-nearest neighbor effects. *Nucleic Acids Res.*, 23::302–6, 1995.

[42] C. Papanicolau, M. Gouy, and J. Ninio. An energy model that predicts the correct folding of the tRNA and the 5S RNA molecules. *Nucl. Acid. Res.*, 12:31–44, 1984.

[43] C. Philippe, C. Portier, M. Mougel, M. Grunberg-Manago, J. P. Ebel, B. Ehresmann, and C. Ehresmann. Target site of escherichia coli ribosomal protein S15 on its messenger RNA. *J Mol Biol*, 211:415–426, 1990.

[44] C. W. A. Pleij. Pseudoknots a New Motiv in the RNA Game. *Trends Biochem Sci*, 15:143–147, 1990.

[45] C. W. A. Pleij, K. Rietveld, and L. Bosch. A new principle of RNA folding based on pseudoknotting. *Nucl. Acids Res*, 13:1717–1731, 1985.

[46] D. Poland and H. A. Scheraga. *Theory of helix coil transitions in biopolymers*. New York and London: Academic Press, 1970.

[47] T. Powers and H. F. Noller. A Functional Pseudoknot in 16S Ribosomal RNA. *EMBO*, 10:2203–2214, 1991.

[48] J. D. Puglisi, J. R. Wyatt, and I. Tinocco. RNA Pseudoknots. *Acc Chem Res*, 24:152–158, 1991.

[49] H. Putzer and N. Gendron. Co-ordinate expression of the two threonyl-tRNA synthetase genes in bacillus subtilis: control by transcriptional antitermination involving a conserved regulatory sequence. *EMBO*, 11:3117–3127, 1992.

[50] A. L. N. Rao, T. W. Dreher, L. E. Marsch, and T. C. Hall. Telomeric Function of the tRNA-like Structure of Brome Mosaic Virus RNA. *Proc Natl Acad Sci*, 86:5335–5339, 1989.

[51] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

[52] E. Rivas and S. R. Eddy. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, 16:334–340, 2000.

[53] T. B. S. Ebel and A. N. Lane. Thermodynamic stability and solution conformation of tandem g a mismatches in RNA and RNA/DNA hybrid complexes. *Eur. J. Biochem.*, 220::703–15, 1994.

[54] W. Saenger. *Principles of Nucleic-Acid Structure*. Springer-Verlag, New York, first edition, 1984.

[55] W. Salser. Globin messenger RNA sequences - analysis of base-pairing and evolutionary implications. *Cold Spring Harbour Symp. Quant. Biol.*, 42:985, 1977.

[56] P. Schimmel. RNA Pseudoknots that Interact with Components of the Translation Apparatus. *Cell*, 58:9–12, 1989.

[57] M. J. Serra, T. J. Axenson, and D. H. Turner. A model for the stabilities of RNA hairpins based on a study on the sequence dependence of stability for hairpins of six nucleotides. *Biochemystry*, 33::14289–965., 1994.

[58] M. J. Serra, M. H. Lyttle, T. J. Axenson, C. A. Schadt, and D. H. Turner. RNA hairpin loop stability depends on the closing base pair. *Nucleic Acids Res.*, 21:3845–9, 1993.

[59] S. Spiegelman. An approach to the experimental analysis of precellular evolution. *Quart. Rev. Biophys.*, 17:213–219, 1971.

[60] W. Stockmayer and H. Jacobson. Intramolecular reaction in polycondensations. *J. Chem. Phys.*, 18:1600–1606, 1950.

[61] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Annual Review of Biophysics and Biophysical Chemistry*, 17:167–192, 1988.

[62] T. H. Tzeng, C. L. Tu, and J. A. Bruenn. Ribosomal Frameshifting Requires a Pseudoknot in the Saccharomyces cerevisiae Double-stranded RNA Virus. *J Virus*, 66:999–1006, 1992.

[63] F. H. D. van Batenburg, A. P. Gultyaev, and C. W. A. Pleij. Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res*, 28:201–204, 2000.

[64] A. E. Walter, D. H. Turner, J. Kim, M. Lyttle, P. Muller, D. H. Mathews, and M. Zucker. Coaxial stacking of helices enhances binding of oligoribunucleotides and improves prediction of RNA folding. *Proc. Natl Acad Sci.*, 91::9218–22, 1994.

[65] A. E. Walter, M. Wu, and D. Turner. The stability and structure of tandem g-a mismatches in RNA depends on closing base pairs. *Biochemistry*, 33::9218–22, 1994.

[66] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Studies*, 1:167 – 212, 1978.

[67] M. S. Waterman and T. F. Smith. Rapid dynamic programming methods for RNA secondary structures. *Adv. Appl. Math.*, 7:455–464, 1986.

[68] A. M. Weiner and N. Maizels. tRNA-like Structures Tag the 3' ends of Genomic RNA Molecules for Replication: Implications for the Origin of Protein Sybthesis. *Proc Natl Acad Sci*, 84:7383–7387, 1987.

[69] N. Wills, R. F. Gesteland, and J. F. Atkins. Evidence that a Downstream Pseudoknot is Required for Translational Readthrough of the Moloney Murine Leukemia Virus Gag Stop Codon. *Proc Natl Acad Sci U S A*, 88:6991–69995, 1991.

[70] C. R. Woese and R. R. Gutell. Evidence for Several Higher Order Structural Elements in Ribosomal RNA. *Proc Natl Acad Sci U S A*, 86:3119–3122, 1989.

[71] I. Wool, Y. Endo, Y.-L. Chan, and A. Glck. Structure, function and evolution of mammalian ribosomes. *Biochem. Cell Biol.*, 73:933–947, 1995.

[72] S. Wuchty, W.Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

Curriculum vitae

Mag. Christian Haslinger
* 23.3.1972, St.Pölten

1978 – 1982	Volksschule in St.Pölten
1982 – 1986	Hauptschule in St.Pölten
1986 – 1991	Höhere Technische Lehranstalt St.Pölten, Fachrichtung Informatik
6/1991	Matura an der HTL
10/1991	Beginn Studium der Biologie an der Universität Wien
10/1992 – 12/1997	Studium der Biochemie an der Universität Wien
11/1996 – 12/1997	Diplomarbeit am Institut für Theoretische Chemie der Universität Wien bei Prof. Dr. Peter Stadler
12/1997	2. Diplomprüfung mit Auszeichnung, Sponsion zum Mag. rer. nat.
2/1998 – 2/1999	Zivildienst
2/1999 – 3/2001	Dissertation am Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien bei Prof. Dr. Peter Schuster