universität
wien

DIPLOMARBEIT

RNA Secondary Structure Prediction including Pseudoknots

angestrebter akademischer Grad
Magister der Naturwissenschaften (Mag. rer. nat.)

Verfasser: Wolfgang Beyer
Matrikelnummer: 9925920
Studienrichtung: Molekulare Biologie (A490)
Betreuer: Ao. Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Wien, im Oktober 2010

Danksagung

An dieser Stelle möchte ich all jenen herzlich danken, die zum Gelingen dieser Arbeit beigetragen haben:

Ivo Hofacker und Christoph Flamm danke ich für die unkomplizierte Betreuung und die kompetente Unterstützung bei meiner Diplomarbeit sowie für die Aufnahme in die Arbeitsgruppe am TBI.

Allen anderen Mitarbeitern am Institut danke ich für ihre Hilfsbereitschaft beim Lösen meiner diversen größeren und kleineren Probleme und für die angenehme Arbeitsatmosphäre.

Meinen Eltern danke ich dafür, daß sie mir dieses Studium ermöglicht haben und mich in jeder Hinsicht unterstützt haben, wo sie nur konnten.

Abstract

RNAs are very important biological molecules. Previously they were thought of as being only the intermediary between DNA, which carries the genetic information, and proteins, which catalyze biochemical reactions. Today we know about the existence of diverse classes of RNAs which exhibit catalytic functions themselves. The function of an RNA molecule is dependent on its three-dimensional structure (the tertiary structure), which is in turn dependent on the base pairing within the RNA molecule (the secondary structure).

In order to draw functional conclusions from the linear sequence of an RNA molecule (the primary structure), one would ideally be able to predict the whole three-dimensional fold based on the sequence alone. But because the folding process of RNA is mainly a hierarchical process, with the secondary structure forming before any tertiary interactions, the secondary structure can already be used as a starting point for functional analysis. Therefore prediction of the secondary structure of RNAs is a central problem in bioinformatics.

The majority of all RNA base pairs are perfectly nested, meaning that all nucleotides enclosed by a specific base pair do not interact with any nucleotides outside of this base pair. This property allows the decomposition of the whole RNA secondary structure into simpler and independent substructures called loops, for which free energy parameters exist. The most common approach to predicting RNA secondary structures is based on dynamic programming, which relies heavily on this loop decomposition.

A certain group of RNA secondary structures called pseudoknots, of which more and more have been discovered in recent years, do not allow this simplification. In a pseudoknot nucleotides within a loop form base pairs with nucleotides outside of the loop, violating the condition of perfectly nested secondary structures. Pseudoknots are therefore more difficult and more expensive to handle computationally and the standard RNA secondary structure prediction algorithms simply do not take pseudoknots into account.

Approaches for predicting pseudoknots have only been developed in recent years, some of them based on dynamic programming, others on heuristic methods. In this diploma thesis I present *PKplex*, a new dynamic programming based algorithm for the prediction of RNA secondary structures including pseudoknots. After describing the basic idea behind *PKplex* and its implementation, the algorithm is then evaluated against a large set of known RNA pseudoknots and its performance compared with other published algorithms.

Zusammenfassung

RNAs sind sehr wichtige Biomoleküle. Früher sah man in ihnen nur die Zwischenstufe zwischen DNA, dem Träger der genetischen Information, und Proteinen, den Katalysatoren biochemischer Reaktionen. Heute wissen wir von der Existenz verschiedenster Klassen von RNAs, die selbst katalytische Eigenschaften haben. Die Funktion eines RNA-Moleküls ist von seiner dreidimensionalen Struktur (der Tertiärstruktur) abhängig, die wiederum von den Basenpaarung innerhalb des RNA-Moleküls (der Sekundärstruktur) abhängig ist.

Um von der linearen Sequenz (der Primärstruktur) auf die Funktion eines RNA-Moleküls schließen zu können, sollte man im Idealfall in der Lage sein, allein von der Sequenz die komplette dreidimensionale Struktur vorhersagen zu können. Weil aber RNA-Faltung als hierarchischer Prozess betrachtet werden kann, wobei sich die Sekundärstruktur vor jeglichen tertiären Interaktionen ausbildet, kann schon die Sekundärstruktur als Ausgangspunkt für die funktionelle Analyse dienen. Dementsprechend ist RNA-Sekundärstrukturvorhersage ein zentrales Problem der Bioinformatik.

Der Großteil aller RNA-Basenpaare ist perfekt verschachtelt, was bedeutet, daß alle Nukleotide, die von einem Basenpaar umschlossen sind, nicht mit Nukleotiden außerhalb dieses Basenpaars interagieren. Diese Eigenschaft erlaubt es, die gesamte RNA Sekundärstruktur in einfachere und voneinander unabhängige Substrukturen, die sogenannten Loops, für deren freie Energien man Parameter kennt, zu zerlegen. Dynamic Programming, der am häufigsten verwendete Ansatz zur RNA-Sekundärstrukturvorhersage, ist auf diese Loop-Zerlegung angewiesen.

Pseudoknoten, von denen man in letzter Zeit immer mehr entdeckt hat, sind RNA-Strukturen, die diesen vereinfachenden Schritt nicht zulassen. Bei einem Pseudoknoten formen Nukleotide innerhalb eines Loops Basenpaare mit Nukleotiden außerhalb des Loops und verletzen damit die Bedingung der perfekt verschachtelten Sekundärstrukturen. Deshalb ist die Berücksichtigung von Pseudoknoten rechnerisch komplizierter und aufwändiger und herkömmliche Algorithmen zur RNA-Sekundärstrukturvorhersage schließen Pseudoknoten der Einfachkeit halber aus.

Erst in den letzten Jahren wurden Ansätze zur Vorhersage von Pseudoknoten entwickelt, die entweder auf Dynamic Programming oder auf heuristischen Methoden beruhen. In dieser Diplomarbeit präsentiere ich *PKplex*, einen neuen, Dynamic Programming-basierten Algorithmus zur Vorhersage von RNA Sekundärstrukturen mit Pseudoknoten. Zuerst wird die grundlegende Idee hinter *PKplex* und ihre Umsetzung beschrieben, und dann wird der Algorithmus auf einen großen Datensatz bekannter RNA Pseudoknoten angewandt und seine Ergebnisse mit denen anderer publizierter Algorithmen verglichen.

Contents

1 Introduction 1

2 RNA - Biological Background 5

2.1 RNA Composition 5
2.2 RNA Functions 8

3 RNA Secondary Structure 13

3.1 Definitions 14
3.2 Representations 16
  3.2.1 Squiggle Plot 16
  3.2.2 Dot-Bracket Notation 16
  3.2.3 Arc Plot 17
  3.2.4 Circular Graph 18
  3.2.5 Dot Plot 18
  3.2.6 Mountain Plot 20
3.3 RNA Pseudoknots 20
  3.3.1 Pseudoknot Types 20
  3.3.2 Examples of Pseudoknots 24

4 RNA Energy Models 27

4.1 The Loop-based RNA Energy Model 27
4.2 Energy Models for RNA Pseudoknots 29
4.3 The Cao-Chen Energy Model for H-type Pseudoknots 30

5 RNA Secondary Structure Prediction Algorithms 33

5.1 Maximizing the Number of Base Pairs 33
5.2 Minimizing the Free Energy 34
5.3 The Partition Function 38

5.4 Structure Prediction Including Pseudoknots 40

6 PKplex 47

6.1 Algorithm 47

6.1.1 Calculation of Accessibility 50

6.1.2 Calculation of the Interaction Energy 51

6.2 Time Complexity and Implementation 53

7 Results and Discussion 55

7.1 Results 55

7.1.1 Accuracy Measures 55

7.1.2 Results of PKplex for Selected Sample Sequences . . . 56

7.1.3 Comparison with Other Algorithms 58

7.1.4 Data Sets 60

7.1.5 Prediction Accuracy 60

7.1.6 Computational Performance 62

7.2 Discussion 64

8 Conclusion and Outlook 69

Bibliography 71

CV 81

List of Figures

2.1 Chemical structure of the RNA building blocks 6
2.2 RNA base pairs 7
2.3 Hierarchic folding of an RNA molecule 8
2.4 The central dogma of molecular biology 9
2.5 Diverse functions of RNA 11

3.1 Loop types in RNA secondary structures 15
3.2 RNA squiggle plot 17
3.3 RNA arc plot 17
3.4 Circular RNA graph 18
3.5 RNA dot plot 19
3.6 RNA mountain plot 21
3.7 H-type pseudoknot 22
3.8 Kissing hairpin type pseudoknot 23
3.9 Bi-secondary structure 24
3.10 Non-planar pseudoknot 24

4.1 RNA loop decomposition 28
4.2 Pseudoknot covered by the Cao-Chen energy model 31

5.1 Structure decomposition in the Nussinov algorithm 33
5.2 Recursion for $F_{i,j}$ 35
5.3 Recursion for $C_{i,j}$ 36
5.4 Recursion for $M_{i,j}$ 36
5.5 Recursion for $M^1_{i,j}$ 37
5.6 Pseudoknot boundaries in *pknotsRG* 41
5.7 The pseudoknot classes covered by different algorithms 43

6.1 Pseudoknotted and pseudoknot-free structure of the same RNA molecule 48
6.2 Pseudoknot decomposition 52
6.3 Pseudocode for the calculation of the interaction energy. . . . 52
6.4 Pseudocode with rearranged loops 54

7.1 `HIVRT32` secondary structure 56
7.2 `MMTV` secondary structure 57
7.3 `telo.human` secondary structure 59
7.4 Runtime comparison 63
7.5 Pseudoknot classes covered by *PKplex* and other algorithms . 65
7.6 Loops involved in pseudoknot formation 66

List of Tables

7.1 Data sets for testing the pseudoknot prediction algorithms . . 60
7.2 Results of pseudoknot prediction with different algorithms . . 61

Chapter 1

Introduction

Not many decades ago RNA was thought of as mainly being messenger RNA, the passive intermediary between the gene-encoding DNA in the nucleus and the ribosomes, which translate the genetic information into the amino acid sequence of proteins, which then carry out the actual biochemical functions. This view has since then changed a lot:

RNAs do not only carry information, they are functionally active units themselves. They can act regulatorily or exhibit catalytic activity, both functions previously only attributed to proteins. In addition to tRNAs and rRNAs, which are vital for protein synthesis, entire classes of functional RNAs involved in diverse processes such as RNA splicing, gene regulation and chromosome structure have been discovered. This versatility of RNA has even lead to the formulation of the RNA world hypothesis, which states that our current biological world based on DNA for information storage and proteins for enzymatic activity evolved out of an era in which both of these functions were fulfilled by RNA [Gilbert, 1986; Joyce, 1989, 1991].

The function of an RNA molecule depends strongly on its three-dimensional structure, which is in turn dependent on its sequence. RNA structure formation is mainly a hierarchical process which can be separated into two steps [Brion and Westhof, 1997]: The first step is the formation of the secondary structure consisting of the set of base pairs between individual pairs of nucleotides in the RNA sequence. The second step, the formation of the tertiary structure, consists of the bending and folding of the secondary structure leading to the final three-dimensional fold of the RNA sequence.

The majority of all RNA base pairs are non-crossing, which means that for every base pair (i, j) with $i < j$ there is no base pair (k, l) $(k < l)$ with $k < i < l < j$ or $i < k < j < l$. Secondary structures elements which do not fulfill this condition are called pseudoknots. Pseudoknots are therefore defined as structures where bases enclosed by at least one base pair form base pairs with bases outside of the enclosing base pair. Pseudoknots can be interpreted as being part of both secondary and tertiary interactions: on the one hand the formation of nucleotide base pairs is the defining feature of secondary structure elements, on the other hand pseudoknots often form between parts of the RNA molecule that seem to be spatially distant from each other if only the pseudoknot-free secondary structure is taken into account, thereby directly influencing the RNA tertiary structure. It is known that pseudoknots play an important functional role in many RNA mediated processes. Examples include self-splicing group I introns [Cech, 1988], ribosomal RNAs [Cannone et al., 2002], Ribonuclease P [Brown, 1996], various prion mRNAs [Barette et al., 2001], transfer messenger RNAs [Andersen et al., 2006], viral pseudoknots involved in genome replication or ribosomal frameshifting [Giedroc and Cornish, 2009], and telomerase RNAs [Staple and Butcher, 2005].

Predicting the structure of RNA or protein biopolymers from sequence information alone is an important area in bioinformatics. While the sequence information is abundantly available, gaining the functionally relevant structural information experimentally requires extensive laboratory work. Since all necessary information determining the three-dimensional structure is in principle already contained in the linear sequence of the biopolymers' building blocks a theoretical approach seems obvious.

The RNA tertiary structure is difficult to obtain experimentally and computationally intractable to predict, because the secondary structure base pairing and base stacking energies already contribute the major part of the energy gained by folding the RNA. Tertiary structure interactions only play a minor role energy-wise and are therefore a lot more difficult to predict. Nevertheless, the RNA secondary structure is often sufficient to perform a successful functional analysis and is therefore generally accepted as a valid starting point.

The commonly used algorithms for predicting RNA secondary structures are based on thermodynamic models, searching for the structure with minimum free energy (MFE)[Zuker, 2000]. Despite their importance, pseudoknots are excluded in the standard approach, because their absence allows the use of fast and efficient dynamic programming routines. A free implementation of these algorithms is provided by the *Vienna RNA Package* [Hofacker et al., 1994], which is available at `http://www.tbi.univie.ac.at/~ivo/RNA/`.

While including arbitrary pseudoknots into the analysis has been shown to be NP-complete [Akutsu, 2000; Lyngsø and Pedersen, 2000], advances in predicting pseudoknots have nevertheless been made. Some approaches reduced the time complexity of the problem by using a simpler energy model, but still considering all possible pseudoknots [Akutsu, 2000; Tabaska et al., 1998; Ruan et al., 2004]. Others restricted the types of predictable pseudoknots to improve the computational cost of their algorithms [Rivas and Eddy, 1999; Dirks and Pierce, 2003; Lyngsø and Pedersen, 2000; Reeder and Giegerich, 2004]. Finally, heuristic approaches have also been employed [Gultyaev et al., 1995; Cai et al., 2003; Xayaphoummine et al., 2003; Ruan et al., 2004; Ren et al., 2005; Andronescu et al., 2010; Sperschneider and Datta, 2010].

In this diploma thesis I am presenting *PKplex*, a new dynamic programming algorithm for predicting RNA secondary structures with pseudoknots. I am starting in chapter 2 with a review of RNA and its biological relevance and talk about RNA structures in chapter 3. Chapter 4 focuses on RNA energy models and chapter 5 presents existing RNA folding algorithms, some of them including pseudoknots, some of them not. *PKplex* is introduced in chapter 6 and its results and performance evaluated in chapter 7. Finally, an outlook and conclusion is given in chapter 8.

Chapter 2

RNA - Biological Background

2.1 RNA Composition

Ribonucleic acids (RNAs) are linear biopolymers and one of the most important class of macromolecules in any living cell. An RNA molecule is built by a chain of nucleotides, its monomeric building blocks, which consist of a nitrogenous hetero-cyclic purine or pyrimidine base, the pentose sugar ribose and a phosphate group. The base, generally either one of the purines adenine (A) or guanine (G), or one of the pyrimidines cytosine (C) or uracil (U), is attached to the 1' carbon atom of the ribose as depicted in Figure 2.1. A phosphate group links the 3' carbon of one nucleoside with the 5' carbon of the next via a phosphodiester bond creating the sugar-phosphate backbone of RNA. According to the carbon atom not linked to another nucleotide the two ends of an RNA strand are called 5'- and 3'-end.

RNA is very similar to DNA (deoxyribonucleic acid), but differs in four main ways: First, DNA is double-stranded while RNA is generally single-stranded. Second, RNAs are usually shorter than DNAs. Third, instead of the sugar ribose DNA contains deoxyribose, which has no hydroxyl group attached to the 2' carbon. This causes DNA to be chemically more stable than RNA because it is less prone to hydrolysis. And fourth, instead of

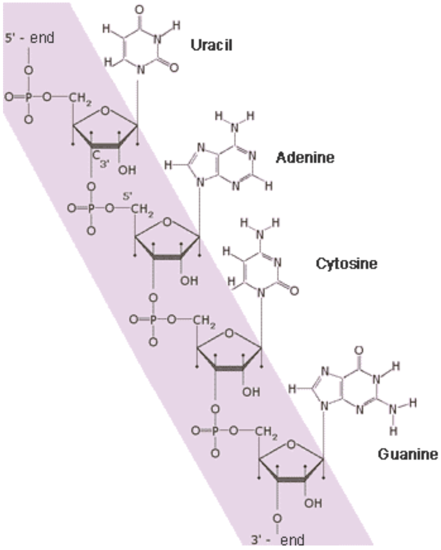

Figure 2.1: Chemical structure of the RNA building blocks. The bases adenine (A), guanine (G), cytosine (C) and Uracil (U) are linked to the sugar-phosphate backbone (highlighted in purple). (Image reproduced from www.mathcell.ru)

uracil the fourth base in DNA is the chemically very similar pyrimidine thymine. In addition, while both RNA and DNA can form helices, there exist different helix types depending on the exact spatial arrangement of the atoms in the helix. For DNA, the most common form is B-DNA while RNA typically exists in an A-DNA like conformation, also called A-RNA.

The primary structure of an RNA molecule is the nucleotide sequence, which is usually presented as a sequence of the letters A, G, C and U starting from the 5'-end through to the 3'-end. This four letter encoding of a sequence can be easily stored in and retrieved from databases and is used as starting point for various bioinformatical methods of analysis. Typical RNA sequence lengths vary from not much more than a dozen nucleotides to several million nucleotides.

The bases in a nucleic acid can form hydrogen bonds to other bases, thereby creating base pairs. This base pairing mechanism is dependent on the base types involved, not all combinations of two bases can form bonds under normal conditions. The most common base pairs AU and GC (and their inverses UA and CG) are called Watson-Crick base pairs in honor of James

Watson and Francis Crick who discovered the base pairing mechanism during their effort to determine the three-dimensional structure of DNA in 1953 [Watson and Crick, 1953]. An AU base pair is made up of two hydrogen bonds, whereas a GC base pair contains three hydrogen bonds, which is one of the reasons why the latter base pair is more stable than the first (Figure 2.2). The energetically weaker GU base pairs also occur frequently and are called wobble pairs.

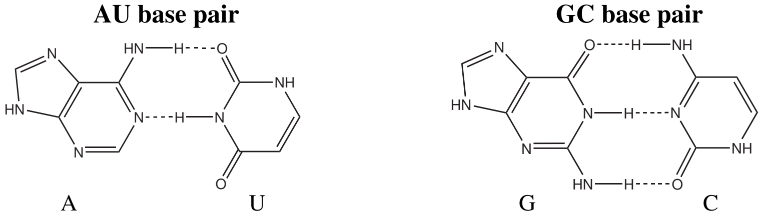

Figure 2.2: AU and GC base pairs. The purines adenine (A) and guanine (G) are shown on the left hand side of base pair, the pyrimidines uracil (U) and cytosine (C) are on the right. Dashed lines indicate hydrogen bonds. (Image reproduced from [Lorenz, 2007])

DNA usually consists of two complementary strands which can form base pairs from end to end, thus creating its famous double helix structure. Because of its single-strandedness RNA is able to form intramolecular base pairs. Multiple successive base pairs form helical stems interspersed with unpaired loop regions. The resulting structure is referred to as the RNA secondary structure. Figure 2.3b shows an example of an RNA secondary structure. The properties of secondary structures will be covered in more detail in chapter 3.

The embedding of an RNA molecule with its secondary structure in three-dimensional space is called the tertiary structure. This three-dimensional fold is often the result of stabilizing non-standard base pairs, triple base pairs and backbone-loop interactions. These interactions are usually weaker than the interactions responsible for the canonical base pairs of the secondary structure and as a consequence the secondary structure contributes the bigger part of the stabilizing energy of the whole RNA fold. RNA folding can therefore be seen as a hierarchical process with the base pairs forming first, before the full three-dimensional arrangement of the RNA molecule in its tertiary structure takes place [Tinoco et al., 1999]. This hierarchical nature

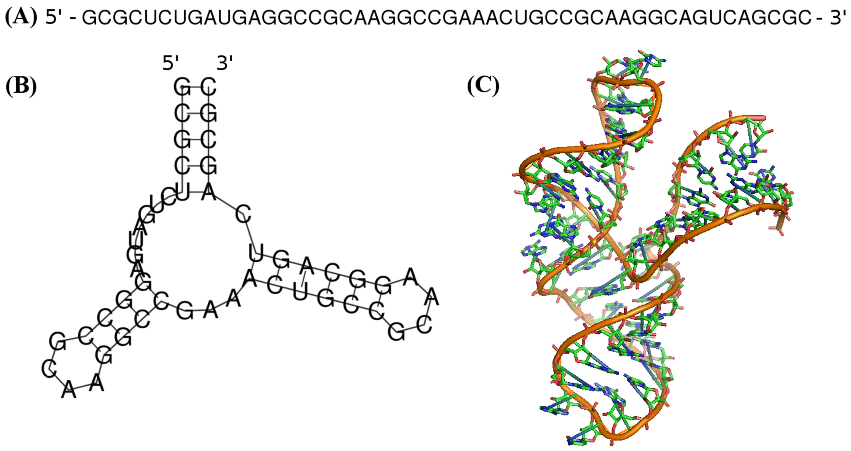
(A) 5' - GCGCUCUGAUGAGGCCGCAAGGCCGAAACUGCCGCAAGGCAGUCAGCGC - 3'
(B)
5'
3'
(C)

Figure 2.3: Hierarchic folding of an RNA molecule. (A) The primary structure is the nucleotide sequence. (B) Formation of base pairs leads to the secondary structure. (C) The tertiary structure consists of the complete three-dimensional fold of the molecule. (The sequence of the RNA hammerhead ribozyme was taken from [Tuschl et al., 1994]; secondary structure predicted with the *Vienna RNA package*; 3D-structure based on PDB-ID 1RMN and displayed with *PyMOL*).

of RNA folding is also the reason why it makes sense to draw conclusions about the function of an RNA molecule from its secondary structure prediction without knowing the precise tertiary fold, which is much more difficult to predict due to its smaller energy contribution to the fold. These functional predictions are usually done by comparing the predicted secondary structure to the structures of RNAs with known functions.

2.2 RNA Functions

For decades the central dogma of molecular biology - formulated by Francis Crick [Crick, 1958, 1970] - heavily influenced the opinion on the role of RNAs in cellular processes. The dogma, which is often summarized as "DNA makes RNA makes protein", states that the general flow of information in a cell goes from DNA to RNA to protein with some exceptions like reverse transcription being allowed under special circumstances (see Figure 2.4). According to this view the genetic information stored in the genomic DNA located in the nucleus is transcribed into messenger RNA (mRNA), which in turn transmits this information to the ribosomes in the cytoplasm, where it is translated into a protein sequence with the mRNA serving as a template.

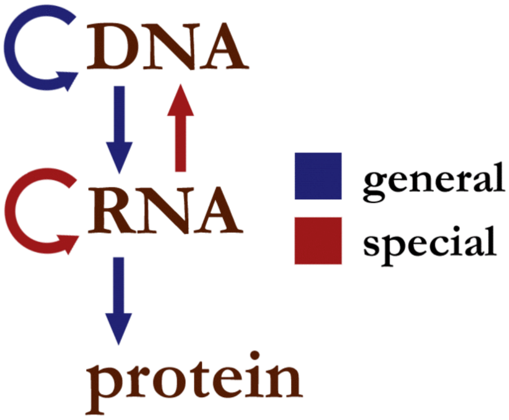

Figure 2.4: The central dogma of molecular biology. The general transfers of sequence information occur under normal circumstances in most cells. The special transfers only occur under specific circumstances.

This protein-centric view of life with all RNAs being just mRNAs, an intermediate between DNA and protein, has changed only gradually. Altman and Cech showed that not only proteins, but RNAs too could catalyze biological reactions [Guerrier-Takada et al., 1983; Cech et al., 1981]. They were awarded the Nobel Prize in Chemistry in 1989 for the discovery of ribozymes, which are RNAs showing enzymatic properties.

Today we know a whole number of diverse classes of RNAs which are transcribed from DNA, but are not translated into proteins. Among the best known of those non-coding RNAs (ncRNAs) are transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), both involved in protein synthesis. tRNAs transport amino acids to the ribosomes and act as adaptors to translate the mRNA triplet code into the according amino acid. Ribosomes, the sites of protein synthesis, are complexes made out of both proteins and RNAs. It has been shown that the catalytic activity of ribosomes, which is the formation of the peptide bond between amino acids, is solely performed by rRNAs while the ribosomal proteins are mostly responsible for the structural integrity of the complex.

Ribonuclease P (RNAse P) cleaves RNA and is responsible for the processing of the 5'-leader sequence of precursor tRNAs to form mature tRNAs [Guerrier-Takada et al., 1983]. Small nuclear RNAs (snRNAs), which are associated together with proteins in complexes called small nuclear ribonucleoproteins (snRNPs) are involved in mRNA intron splicing, regulation of transcription and telomer maintenance [Valadkhan, 2005]. Small nucleolar RNAs (snoRNAs) are responsible for targeting the site for chemical modifications (methylation or pseudouridylation) of nucleotides of other RNAs [Bachellerie et al., 2002]. Transfer-messenger RNA (tmRNA) is a bacterial RNA with tRNA- and mRNA-like properties. It rescues ribosomes that have stalled in the middle of protein synthesis by recycling the stalled ribosome, adding a proteolysis-inducing tag to the unfinished polypeptide, and facilitating the degradation of the aberrant messenger RNA [Keiler et al., 1996].

MicroRNAs (miRNAs) were first discovered in 1993 [Lee et al., 1993]. Together with small interfering RNAs (siRNAs) [Hamilton and Baulcombe, 1999; Elbashir et al., 2001] they play a crucial role in the RNA interference (RNAi) pathway, which most commonly results in post-transcriptional gene silencing, but in some cases RNAi can be activating as well. Today RNAi is a valuable research tool, both in vitro and in vivo because this pathway enables researchers to suppress specific genes of interest in a simple and cheap way by introducing synthetic short double-stranded RNAs into cells. The 2006 Nobel Prize in Physiology or Medicine was awarded to Andrew Fire and Craig Mello for their work on RNAi in the nematode worm *C. elegans* [Fire et al., 1998]. These examples highlight the versatility of RNA (see Figure 2.5) and show that RNAs can catalyze chemical reactions of various types including phosphoryl group transfers, isomerization of C-C bonds and hydrolytic reactions.

RNAs are the only known biological macromolecules with the ability to act as both carrier of genetic information as well as catalytically active substance, thereby combining genotype and phenotype in one molecule. This property has lead to the formulation of the RNA world hypothesis [Gilbert, 1986; Joyce, 1989, 1991], which proposes some form of Darwinian evolution based on RNA alone, prior to our current world of life based on DNA as carrier of information and protein as catalytically active substance.

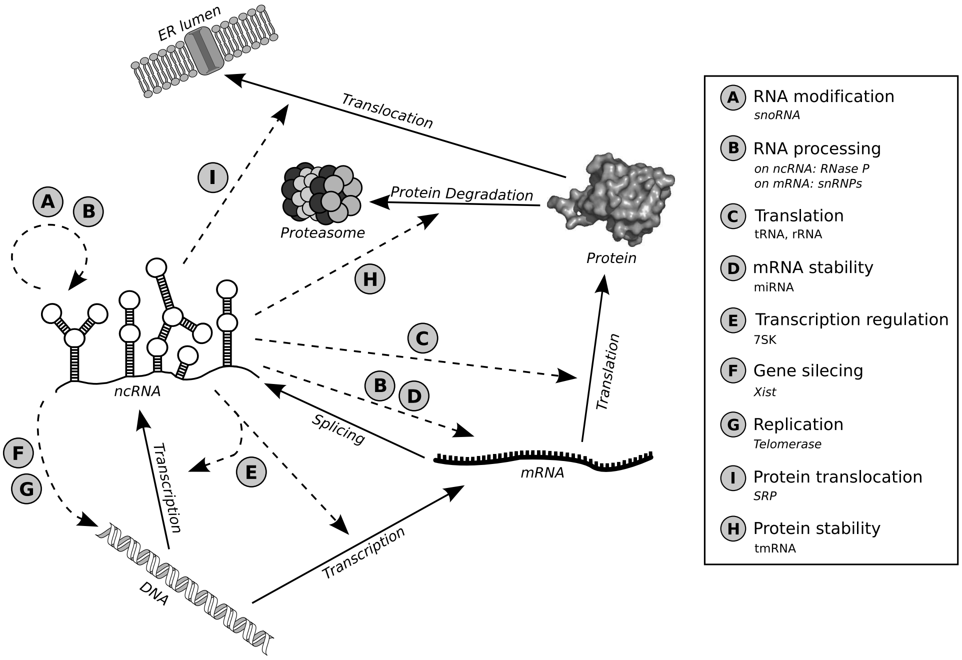

Figure 2.5: Summary of the biological processes RNA molecules are involved in. (Image reproduced from [Gruber, 2007])

Another piece of evidence for the importance of RNA comes from the huge amount of sequence data which has become available only in recent years. Protein coding genes amount to only about 1.5% of the human genome and the estimated number of human protein-coding genes is 23000, which is a lot lower than expected [Stein, 2004]. This number does not appropriately reflect the increased complexity of humans compared to much simpler organisms such as *C. elegans* with about 20000 genes. Combined with the knowledge that a big fraction of the human genome is being transcribed [Birney et al., 2007; Johnson et al., 2005], while many transcripts lack protein-coding potential, this results in the assumption that functional RNAs comprise a significant part of the human genome. John Mattick states that the complexity of higher organisms cannot solely be achieved by proteins, but that there has to be an additional layer of regulatory ncRNAs [Mattick, 2003].

All these findings have contributed to a switch away from a protein dominated view of biological life and to an increased effort in scientific RNA research over the last decades. Our knowledge of RNAs and especially functional RNAs is still very limited and new discoveries will lead to a better understanding of the fundamental processes of cellular life. This makes the analysis and prediction of RNA structures and functions an important, challenging and interesting task.

Chapter 3

RNA Secondary Structure

Solving the full structure problem for RNA is difficult, because the number of degrees of freedom of the RNA chain is very high. However, there are several reasons why the secondary structure can be used as a valid substitute for the full three-dimensional fold:

- The secondary structure with its base pairing and base stacking energies is responsible for the major part of the free energy of folding.
- RNA folding is a hierarchical process causing the secondary structure to be the starting point for the tertiary folding.
- RNA secondary structure provides distance constraints for the formation of the full tertiary fold.
- RNA secondary structures are evolutionary conserved and can be used to successfully predict RNA function.

RNA secondary structures are discrete, easy to visualize and compare, and can be handled efficiently by computational methods. All these factors contribute to the fact that RNA secondary structure prediction is an important and popular task in bioinformatics.

3.1 Definitions

An RNA sequence is defined as a string s over the nucleotide alphabet $\{A, C, G, U\}$. n denotes the length of s and (i, j) with $i < j$ denotes base pairing between s_i and s_j, the nucleotides at positions i and j. A secondary structure is defined as a set S of base pairs with the allowed base pairs being (A, U), (G, C), (G, U) and their reversals. For any two base pairs $(i, j) \in S$ and $(k, l) \in S$ with $i < k$, the following must hold [Waterman and Smith, 1978]:

1. $j - i > 3$.
2. $i = k$ if and only if $j = l$.

The first condition imposes a minimal hairpin loop size of three nucleotides and the second condition states that each base can be part of at most one base pair and forbids e.g. triple base pair interactions. A secondary structure S is called pseudoknot-free if for any two base pairs $(i, j) \in S$ and $(k, l) \in S$ the condition

3. Either $i < j < k < l$ or $i < k < l < j$.

also holds. Otherwise S contains at least one pseudoknot. In a pseudoknot-free secondary structure all base pairs either precede each other or are properly nested, while pseudoknots consist of overlapping base pairs.

A base k is called immediately interior to the base pair (i, j) if $i < k < j$ and if there is no base pair (p, q) such that $i < p < k < q < j$. A base pair (p, q) is called immediately interior to a base pair (i, j) if p and q are immediately interior to (i, j) [Zuker and Sankoff, 1984].

All bases immediately interior to the same base pair (i, j) form the loop enclosed by the exterior pair (i, j). The external loop is defined as the set of bases which are not immediately interior to any base pair, i.e. all nucleotides not enclosed by any base pair.

Loops are characterized by their size u, the number of unpaired bases in the loop, and by their degree k. $k - 1$ is the number of base pairs in the loop (not counting the enclosing base pair), therefore k is the number of base pairs delimiting the loop (including the enclosing base pair).

The different loop types, which are depicted in Figure 3.1, are the building blocks for the loop-based RNA energy model described in section 4.1. Loops of degree 1 have no immediately interior base pair and are called hairpin loops. Loops with a degree of 2 and a size of 0 are stacked (base) pairs. Loops with $k = 2$ and $u > 0$ are called interior loops. Bulge loops are a special case of interior loops where the immediately interior base pair (p, q) lies directly next to the enclosing base pair (i, j), i.e. $p = i + 1$ or $q = j - 1$. Loops with a degree greater than 2 are called multiloops.

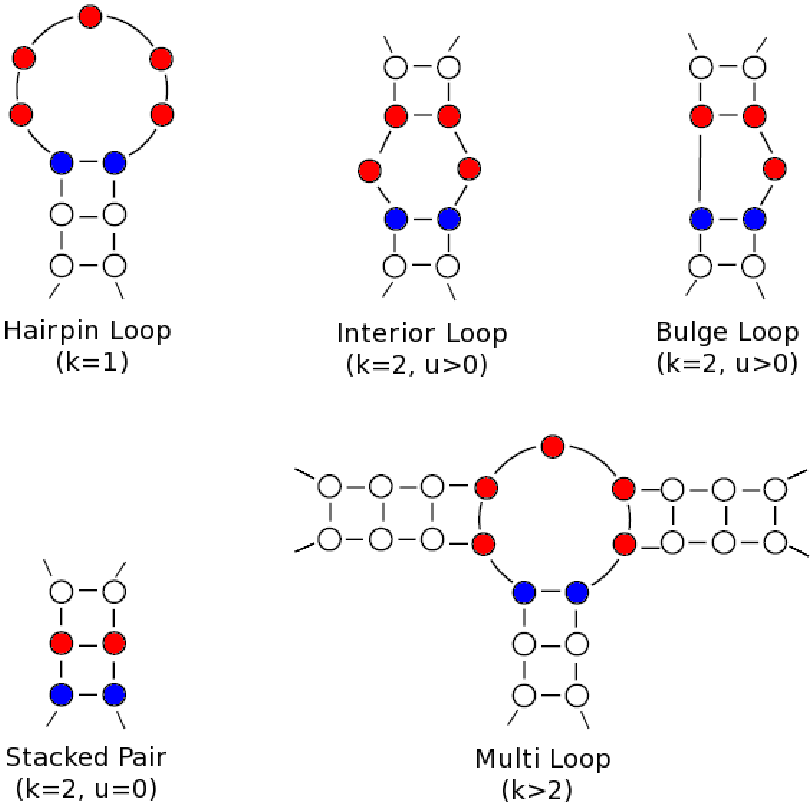

Figure 3.1: Loop types in RNA secondary structures. The bases being part of the same loop are colored in red, the respective enclosing base pairs in blue.

NcRNAs often have similar secondary structures without having very similar primary sequences. With the help of multiple sequence alignments and statistical covariance models the RNA secondary structures of known ncRNAs have been categorized into families which are based on evolution from a common ancestor. The known information about these ncRNA families is stored in the Rfam database, which is among other things very useful for comparing potential new functional RNAs with known ncRNAs [Griffiths-Jones et al., 2005].

3.2 Representations

Instead of as a list of base pairs, which is not intuitive to the human mind, RNA secondary structures are commonly displayed in one of the following ways:

3.2.1 Squiggle Plot

The most common representation of an RNA secondary structure is the so-called squiggle plot. The nucleotide labels are placed along a curved line representing the sugar-phosphate backbone and base pairs are visualized by (usually) short straight lines. Figure 3.2 shows an example. Only pseudoknot-free structures are guaranteed to result in a planar graph, i.e. a graph that can be drawn without any crossing lines.

3.2.2 Dot-Bracket Notation

The computer science community commonly uses the dot-bracket notation, where the RNA sequence is aligned next to a row of symbols: a dot represents an unpaired nucleotide and each base pair is symbolized by a pair of opening and closing brackets of the same type. For pseudoknot-free structures only one type of brackets is needed, pseudoknotted structures require more than one type. The following lines show an example of an RNA sequence and its secondary structure in dot-bracket notation:

```
Human Gln-tRNA
UAGGACGUGGUGUAGUAGGUAGCAUGGAGAAUGUUGAAUUCUCAGGGGUAGGUUCAAUUCCUAUAGUUCUAG
((((((...((((........)))).((((((.....))))))....(((((.......))))).)))))).
```

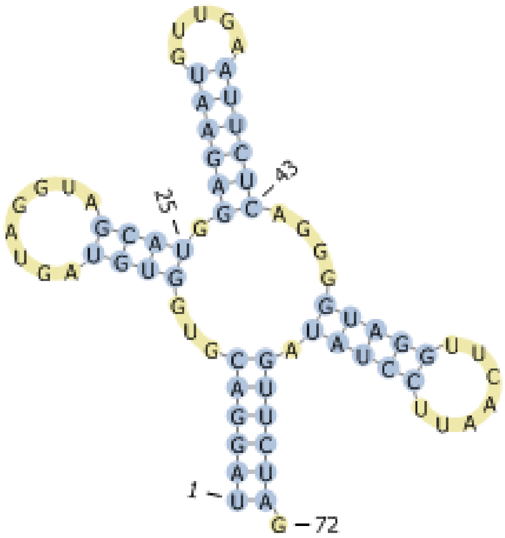

Figure 3.2: Squiggle plot of a human tRNA. (Image generated with Pseudoviewer3 [Byun and Han, 2009])

3.2.3 Arc Plot

Here the RNA backbone is drawn as a straight line and the nucleotides of each base pair are connected by an arc (see Figure 3.3). For pseudoknot-free secondary structures all arcs can be drawn on the same side of the backbone line without intersecting each other.

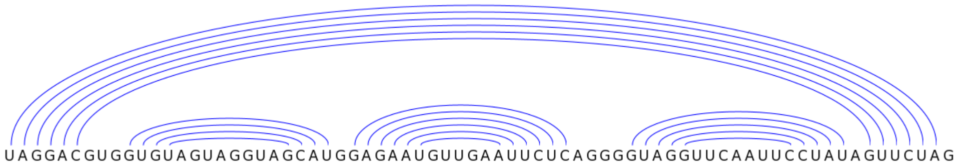
UAGGACGUGGUGUAGUAGGUAGCAUGGAGAAUGUUGAAUUCUCAGGGGUAGGUUCAAUUCCUAUAGUUCUAG

Figure 3.3: Arc plot of a human tRNA. (Image generated with jViz.Rna 2.0 [Wiese et al., 2005])

3.2.4 Circular Graph

In a circular RNA secondary structure graph the backbone is represented by a circle. The base pairs are symbolized by arcs in the interior of the circle exactly like in an arc plot (see Figure 3.4). Only pseudoknot-free secondary structures can be drawn in this way without any crossing of arcs. In graph theory, the formal equivalent of circular RNA graphs are called outerplanar graphs. Outerplanar graphs can be drawn in the plane without crossing of any edges in such a way that all of the vertices belong to the unbounded face of the drawing. Every outerplanar graph is planar, but not every planar graph is outerplanar.

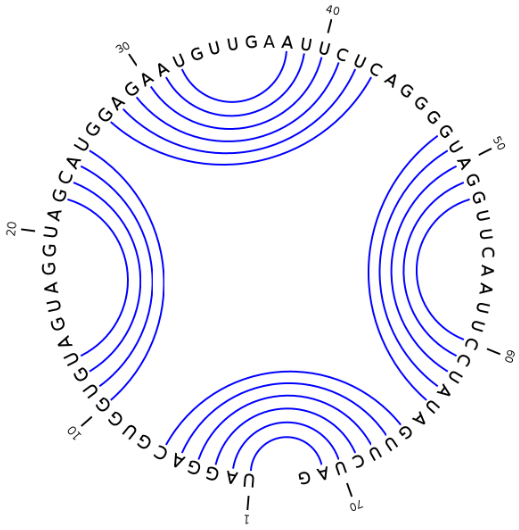

Figure 3.4: Circular graph of a human tRNA. (Image generated with jViz.Rna 2.0 [Wiese et al., 2005])

3.2.5 Dot Plot

Dot plots can convey more information about the RNA sequence than just a single secondary structure. In these two-dimensional graphs the nucleotide sequence is written along both x- and y-axis and the whole graph is divided into two parts by a diagonal from the top-left to the bottom-right corner (see Figure 3.5). In the lower left half of the plot, a dot at the intersection of column i and row j represents a base pair (i, j). Typically, the secondary structure with the minimum free energy is displayed in the lower left part. In the upper right half, more than one structure is visualized. The dot sizes in this part of the plot are proportional to the probabilities of the corresponding base pairs being formed. This allows the visualization of a weighted set of secondary structures such as the Boltzmann ensemble.

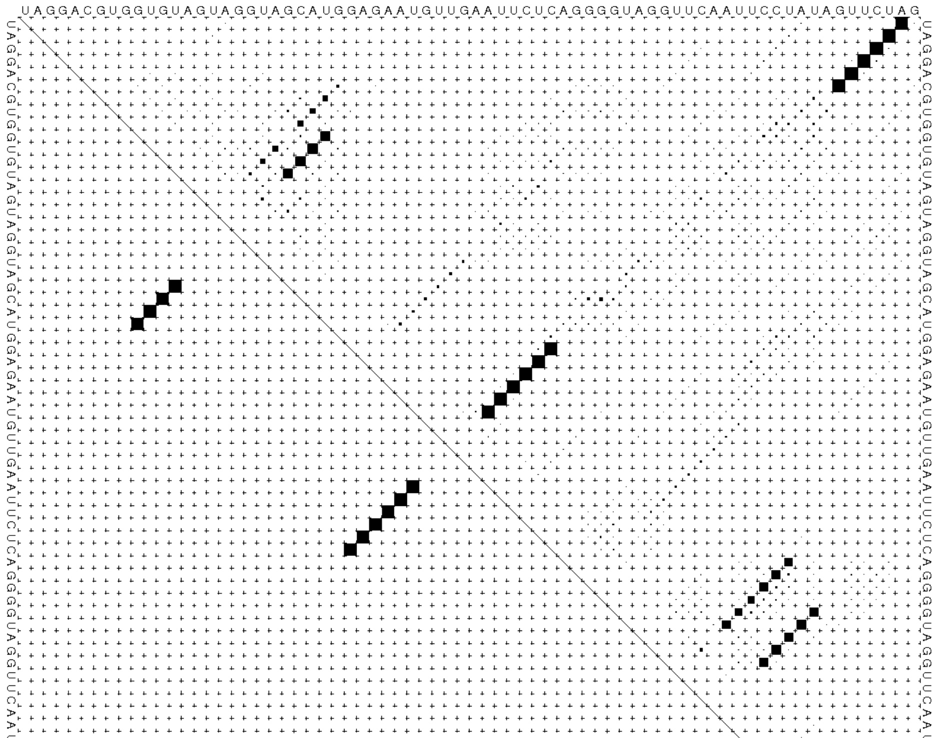

Figure 3.5: Dot plot of human tRNA. The single structure with minimum free energy is displayed in the lower left half, and the base pair probabilities in the upper right half. (Image generated with the *Vienna RNA package*)

3.2.6 Mountain Plot

A mountain plot is another two-dimensional graph to display an RNA secondary structure. The nucleotide index number is plotted against the x-axis and the number of enclosing base pairs against the y-axis (a nucleotide k is enclosed by a base pair (i, j) if $i < k < j$). The resulting plot usually looks similar to a mountain range (see Figure 3.6). Peaks correspond to hairpin loops, showing the unpaired bases enclosed by slopes representing the stem. Plateaus within a sloped region represent bulge loops or, if they are paired with another plateau on the other side of the mountain, interior loops. Mountain plots are not suited to display pseudoknotted secondary structures because pseudoknots cause mountain plots to be ambiguous.

3.3 RNA Pseudoknots

3.3.1 Pseudoknot Types

An RNA pseudoknot is a secondary structure element where unpaired bases within a loop pair with complementary bases in a single-stranded region outside the loop, therefore violating the condition of perfectly nested base pairing which only holds for pseudoknot-free secondary structures. Formally, an RNA secondary structure S is said to contain a pseudoknot if and only if there exist base pairs $(i, j) \in S$ and $(k, l) \in S$, such that $i < k < j < l$.

The simplest type of pseudoknot is called H-type pseudoknot. In these structures the bases in a hairpin loop interact with bases outside of the stem enclosing the hairpin. This generates a second stem and loop with the two stems stacking on top of each other, forming a quasi-continuous helix consisting of one continuous and one discontinuous strand (see Figure 3.7). Another typical and simple pseudoknot structure, the kissing hairpin, is produced by the unpaired bases in a hairpin loop interacting with the unpaired bases in another hairpin loop (see Figure 3.8).

H-type pseudoknots, kissing hairpins and most other observed pseudoknots are planar pseudoknots, i.e. they can be drawn on a plane (as a squiggle or arc plot) without any crossing of arcs. All bi-secondary structures, which

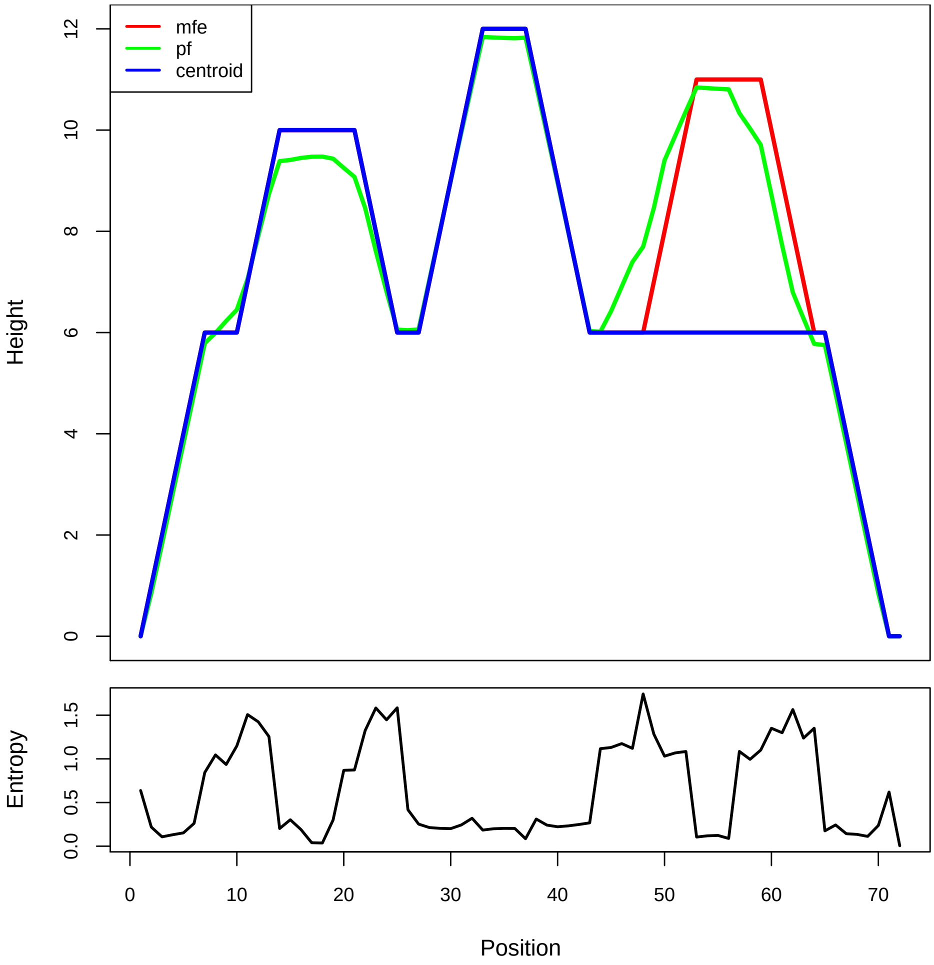

Figure 3.6: Mountain plot for the MFE structure (red), the thermodynamic ensemble of RNA structures (green), and the centroid structure (blue) of a human tRNA. The lower part of the figure displays the positional entropy of the sequence. (Image generated with the *RNAfold* web server [Gruber et al., 2008])

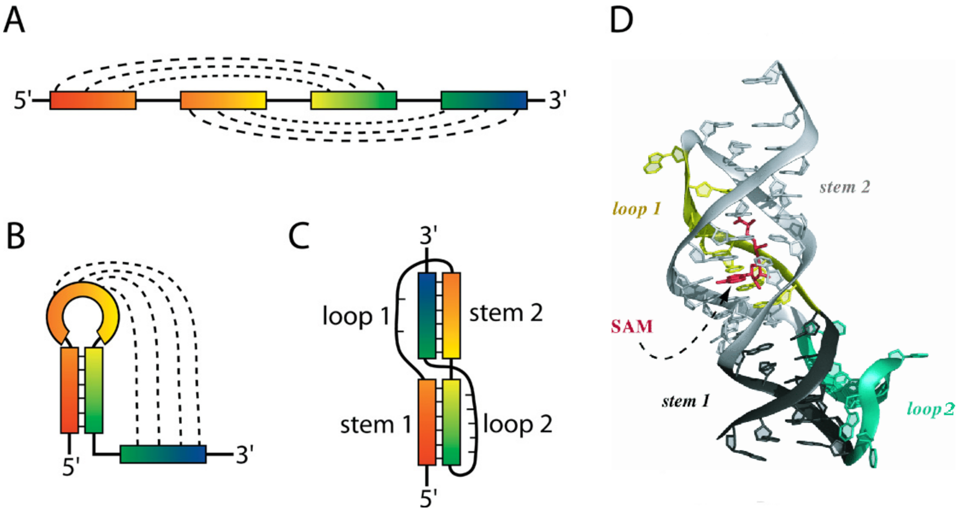

Figure 3.7: H-type pseudoknot architecture.
(A) Linear arrangement of the elements forming the pseudoknot. Base pairing indicated by dashed lines.
(B) Bases within the initial hairpin pair with bases outside of the loop.
(C) Schematic view of the final H-type pseudoknot fold.
(D) 3-dimensional fold of the SAM-II riboswitch, an H-type pseudoknot. The stacking helices are shown in black and gray, the connecting loops in yellow and green and the bound SAM in red.
(Images (A)-(C) reproduced from [Staple and Butcher, 2005], Image (D) reproduced from [Brierley et al., 2008])

are defined as the superposition of two disjoint pseudoknot-free secondary structures (see Figure 3.9), are planar. The converse is not true, since there are planar structures which are not bi-secondary. An example for such a planar, not bi-secondary structure would be a helix crossing scheme of ABCA'B'C' with X-X' denoting one or more consecutive base pairs. For circular RNAs, bi-secondary structures and planar structures are equivalent [Witwer et al., 2004].

Even more complex non-planar folds, which cannot be drawn in a plane without crossing of arcs, are possible as well. A helix crossing scheme such as ABCDA'C'B'D' results in the formation of a non-planar pseudoknot. So far, pseudoknots of this type are quite rare. An example is the ribosome binding site of the *E. coli* α operon (see Figure 3.10).

As an extension to bi-secondary structures, k-partite structures are the union of k pseudoknot-free sub-structures. These structures can be intuitively visualized by imagining a book. The spine of the book represents the RNA sequence, which is shared by all sub-structures. Each of the k pages of the book then contains the arcs of one of the k pseudoknot-free disjoint sub-structures. Rigorous definitions and the mathematical properties of bi-secondary and k-partite RNA secondary structures can be found in [Haslinger and Stadler, 1999].

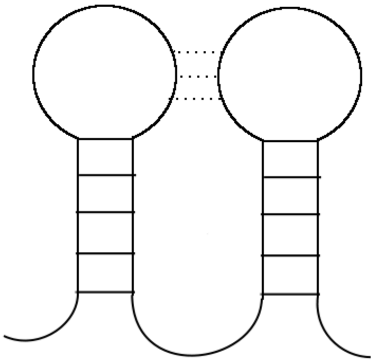

Figure 3.8: Kissing hairpin type pseudoknot. The dotted lines indicate the pseudoknot interaction.

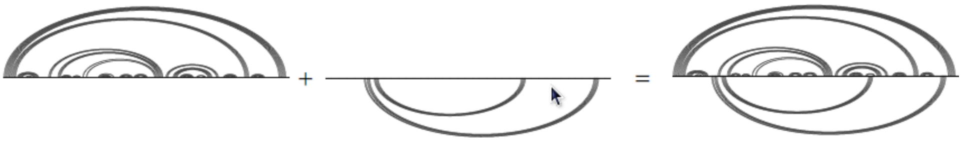

Figure 3.9: A bi-secondary structure is formed by the union of two disjoint pseudoknot-free sub-structures. (Image reproduced from [Witwer et al., 2004])

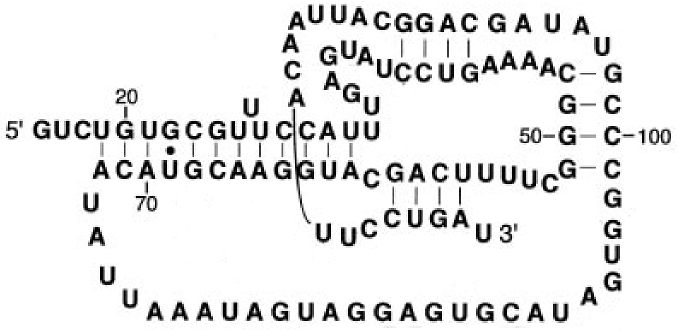

Figure 3.10: The non-planar pseudoknot of the ribosome binding site of the *E. coli* α operon (Image reproduced from [Schlax et al., 2001])

3.3.2 Examples of Pseudoknots

It is known that pseudoknots play an important role in many RNA mediated processes and often the pseudoknots are not only structurally relevant, but are directly responsible for the RNA molecule's function.

A lot of catalytically active RNAs have been discovered to contain pseudoknots. Human telomerase is a riboprotein complex. The $5'$ end of the telomerase RNA forms a highly conserved pseudoknot, which is required for the activity of the complex [Staple and Butcher, 2005]. Mutations within this H-type pseudoknot have been shown to be connected to the diseases autosomal dyskeratosis congenita [Marciniak et al., 2000] and aplastic anemia [Vulliamy et al., 2002]. Most of the ribosomal RNAs contain pseudoknots as well, although in this case they do not seem to sit at the catalytically active site and are assumed to be of mainly structural importance [Cannone et al., 2002]. Another ribozyme is Ribonuclease P, which is responsible for cleaving of a precursor sequence on tRNA molecules. Its catalytically active subunit contains a pseudoknot which is required for the RNase P's function [Brown, 1996; Mann et al., 2003]. The versatile tmRNAs, which are involved in translation by rescuing stalled ribosomes and facilitating the degradation of aberrant messenger RNAs, contain multiple pseudoknots. Although these pseudoknots are evolutionarily conserved, they seem to be not directly involved in the tmRNA's function [Andersen et al., 2006; Nameki et al., 2000]. Other pseudoknots were found in the mRNA of prion proteins [Barette et al., 2001].

In eukaryotic life, introns have to be removed from pre-mRNA to get mature mRNA. This is usually done by the spliceosome riboprotein. Some introns, however, are self-cleaving, catalyzing their removal from the pre-mRNA on their own [Staple and Butcher, 2005; Cech, 1988]. In one group of such introns, the group I self-splicing introns, the catalytic core consists of a pseudoknot [Adams et al., 2004]. Self-cleaving RNA pseudoknots can also be found in various viral genomes: the pseudoknotted self-cleaving Hepatitis delta virus ribozyme is essential for the virus' replication. It is responsible for cutting the replicated multi-genome RNA strand into genome-length units required for virus packaging. It is the fastest-known self-cleaving ribozyme with a reaction rate of one per second [Brierley et al., 2007]. Pseudoknots can also be involved in other viral processes, such as the regulation of viral protein synthesis [Brierley et al., 2008].

Not only catalytically active RNAs contain pseudoknots. Pseudoknots are also involved in ribosomal frameshifting, again commonly found in viruses. Ribosomes normally translate mRNAs without shifting the translational reading frame, but specific pseudoknots together with "slippery" nucleotide sequences can cause the ribosome to change its reading frame during translation [Giedroc and Cornish, 2009; Staple and Butcher, 2005]. Viruses typically have small genomes with very densely packed information because there is no room for bigger genomes in the virus envelope. A lot of viruses are therefore employing a -1 frameshifting mechanism which allows two proteins to be encoded within one genomic region. Because this frameshifting is necessary for the survival of all retroviruses, the pseudoknots involved in this mechanism are attractive targets for drug development. Other examples of viruses with ribosomal frameshift inducing pseudoknots are the severe respiratory syndrome (SARS) coronavirus and other coronaviruses, the mouse mammary tumor virus, the beet western yellow virus and the pea enation mosaic virus.

RNA pseudoknots have been discovered in nearly every organism and have been found to be of functional relevance for ribozymes, self-splicing introns, ribonucleoprotein complexes, viral genomes and other biological systems. The determination of their structure is therefore of great importance.

Chapter 4

RNA Energy Models

In order to predict RNA secondary structures a method of comparing and evaluating different structures is needed. RNA energy models condense the whole secondary structure information into a single number representing the structure's free energy. Energy models therefore serve as scoring functions for RNA folding algorithms. The simplest scoring method employed in the first secondary structure prediction algorithm was to just maximize the number of base pairs [Nussinov and Jacobson, 1980]. A structure with more base pairs than another was assumed to be energetically favorable. This approach has been replaced by the loop-based energy model, which evaluates RNA structures by decomposing them into loops and assigning energy values to these loops (RNA loops have been described in section 3.1).

4.1 The Loop-based RNA Energy Model

The loop-based energy model scores different secondary structures according to their free energy. This idea in its principles has first been proposed about 30 years ago [Waterman and Smith, 1978; Waterman, 1978; Zuker and Stiegler, 1981]. Today's standard model [Mathews et al., 1999; Turner and Mathews, 2009] has been refined several times since then.

Any RNA secondary structure S can be unambiguously decomposed into loops as shown in Figure 4.1. S consists of all base pair enclosed loops $L_{i,j}$

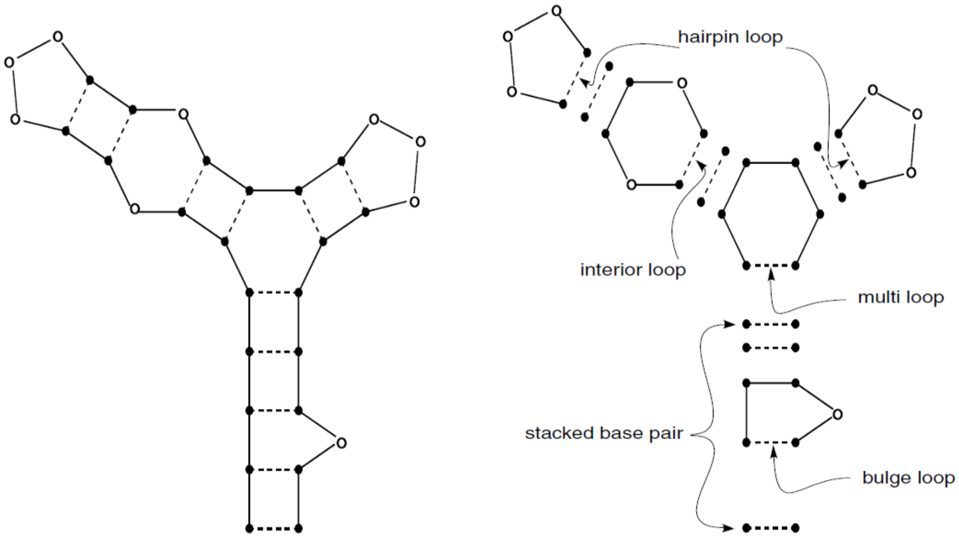

Figure 4.1: RNA Loop Decomposition. The right hand side shows the loop decomposition of the secondary structure depicted on the left hand side. The enclosing base pairs of each loop are indicated by dashed lines. (Image adapted from [Flamm, 1998])

and the external loop L_0 which contains all nucleotides not enclosed by any base pair.

$$S = L_0 \bigcup \left(\bigcup_{(i,j)\in S} L_{i,j} \right)$$

The energy E of an RNA structure is then assumed to be the sum of the energy contributions of all loops of the structure:

$$E(S) = E(L_0) + \sum_{(i,j)\in S} E(L_{i,j})$$

Since absolute energy values are impossible to determine, energy differences between unfolded and folded states in solution are considered. Base pairs are formed by the creation of hydrogen bonds between bases. These hydrogen bonds themselves do not contribute a lot to the free energy of RNAs in solution, since unfolded RNA molecules with no base pairs still form hydrogen bonds with the surrounding water molecules instead. But base pairing also causes adjacent base pairs to stack on top of each other creating stems of stacked base pairs. Base pair stacking is an energetically very favorable interaction. Single bases next to stacking base pairs, i.e. at the end of a stem, can stack onto the stem as well. These dangling end interactions are energetically favorable too. On the other hand, base pairing causes the formation of loops, and in loops the free movement of the nucleotide chain is restricted by the fixed end points, which are the bases forming the pair. This leads to an energetically unfavorable destabilizing entropic effect of all loops.

The concrete energy parameters for the different structure elements have been derived empirically by RNA oligomer folding experiments. The values for stacked bases and small hairpin, internal and bulge loops have been tabulated explicitly. Longer loops are assigned an estimated logarithmically increasing penalty as derived from polymer folding theory. For reasons of computational efficiency, multiloops are usually scored with an affine energy model with a large penalty for the initiation of a multiloop and a smaller penalty for each additional stem added.

4.2 Energy Models for RNA Pseudoknots

Because pseudoknotted RNA secondary structures contain overlapping base pairs, loop decomposition does not work for them, at least not in the way as described above. Therefore the pseudoknotted parts of secondary structures cannot be evaluated with the standard loop-based energy model. In addition, there is also only very limited experimental data on pseudoknot energies available, making the estimation of pseudoknot energies even more difficult.

For pseudoknot-free structures, steric considerations are of no concern, because all secondary structures that can be decomposed into loops are sterically possible. This is not true for pseudoknotted structures and checking whether a suggested pseudoknotted fold is sterically possible is no trivial task. Even for the simple H-type pseudoknots there are restrictions concerning the lengths of the different loops depending among other factors on whether those loops are crossing the minor or the major grove of the RNA. The first pseudoknot energy model dealing with these problems was presented by Gultyaev [Gultyaev et al., 1999].

Consequently, a good energy model for pseudoknots has to be substantially more complex than the current model for pseudoknot-free RNA structures. In the absence of any measured parameters for pseudoknots, most pseudoknot prediction algorithms resort to estimating the energy parameters by using simplified energy models. A common approach is to treat pseudoknots similar to how multiloops are treated in the standard loop decomposition model. This means that the energy associated with a pseudoknot is described by the following linear equation:

$$E_{pk} = \beta_1 + \beta_2 B_p + \beta_3 U_p \tag{4.1}$$

β_1 is a penalty for introducing a pseudoknot, which depends on whether the pseudoknot is embedded in the exterior loop, in a multiloop, or in another pseudoknot. B_p is the number of base pairs that border the interior of the pseudoknot, β_2 is the penalty for each such base pair, U_p is the number of unpaired bases inside the pseudoknot and β_3 is the penalty for each unpaired base.

4.3 The Cao-Chen Energy Model for H-type Pseudoknots

Song Cao and Shi-Jie Chen recently presented a more advanced pseudoknot energy model based on applying polymer physics to the evaluation of pseudoknot loops [Cao and Chen, 2006, 2009]. Loop free energies are made up of enthalpic and entropic contributions, which are depending on factors such as temperature and ionic strength. In Cao and Chen's model the enthalpic contributions are captured by base pairing and stacking energies and their main effort was to estimate the loop entropies of pseudoknot loops. Unfortunately, their model is only applicable to H-type pseudoknots in which the two helices are directly stacked on top of each other and to the more general case in which the two helices are connected by a loop of up to 6 nucleotides. This restriction unfortunately still severely limits the range of pseudoknots that can be evaluated with their model.

For a given pseudoknot defined by the lengths of its two stems and three loops Cao and Chen employed a three-vector virtual-bond-based RNA conformational model to enumerate all possible loop conformations on a grid, while accounting for excluded volume effects. This information was used to compute the loop entropies and following that their free energies (see Figure 4.2). These pre-computed results are stored in tables as loop entropy parameters and can be easily looked up during the evaluation of an actual pseudoknot.

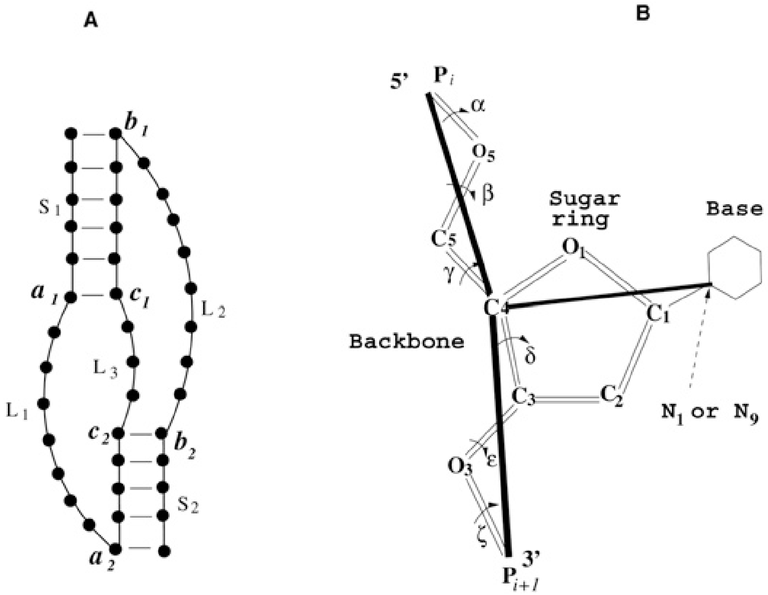

Figure 4.2: (A) A schematic view of a pseudoknot covered by the Cao-Chen model. It is made up of two stems and three loops.
(B) The three vector virtual bond model involves the bonds $P_i - C_4$, $C_4 - P_{i+1}$ and $C_4 - N_1$ (pyrimidine) or $C_4 - N_9$ (purine). (Image reproduced from [Cao and Chen, 2009])

The entropy values of different pseudoknot loops have not been measured experimentally yet and have often been ignored or treated in a simplified way by previous models [Ren et al., 2005; Dirks and Pierce, 2003]. Using the values provided by Cao and Chen allows more accurate calculations of the free energies of H-type and some other pseudoknots. There are still a lot of other pseudoknot types where this model is not applicable and a more general energy model for pseudoknotted RNA secondary structures would be highly desirable.

Chapter 5

RNA Secondary Structure Prediction Algorithms

5.1 Maximizing the Number of Base Pairs

The goal of the first algorithm for RNA secondary structure prediction was to find the structure with the maximum number of base pairs. This algorithm was published by Ruth Nussinov [Nussinov and Jacobson, 1980] based on the idea by Michael Waterman [Waterman and Smith, 1978; Waterman, 1978] to use a dynamic programming approach. The basic principle of dynamic programming is to decompose the overall problem into a number of simpler and smaller subproblems for which optimal solutions can be found.

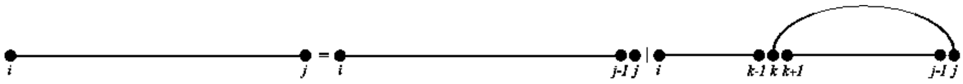

Figure 5.1: Decomposition of RNA structures in the Nussinov algorithm. (Image reproduced from [Gruber, 2007])

We start with an RNA sequence s of length n. $M(i, j)$ denotes the maximum number of base pairs of the subsequence $s_i, \ldots, s_j$. The main idea is that $M(i, j)$ can be calculated recursively in the following way: s_j, the last nucleotide of the subsequence, is either unpaired or paired to some base s_k. In the former case $M(i, j)$ equals $M(i, j-1)$. In the latter case $s_i, \ldots, s_j$ is divided into the intervals $s_i, \ldots, s_{k-1}$ and $s_{k+1}, \ldots, s_{j-1}$ for which the maximum number of base pairs is already known. A graphical representation of this decomposition is shown in Figure 5.1. It produces to the following recursion:

$$M(i,j) = \max \begin{cases} M(i, j-1) \\ \max\limits_{\substack{i \le k < j \\ (k,j) \in S}} \left(M(i, k-1) + M(k+1, j-1) + 1\right) \end{cases}$$

The matrix $M(i,j)$ is filled proceeding from shorter to longer subsequences. The value stored in $M(1,n)$ contains the maximum number of base pairs for the whole sequence. The corresponding structure is then deduced via backtracking. This means going backwards through the calculated matrix reconstructing the path and therefore the list of base pairs for $M(1,n)$.

Although the matrix can be stored as a triangular matrix the algorithm still requires $\mathcal{O}(n^2)$ memory space. The runtime scales with $\mathcal{O}(n^3)$ because the algorithm iterates over i, j and k.

This non-thermodynamic base pair maximization model is too simple too give realistic RNA secondary structure predictions, but it serves as a starting point for more sophisticated algorithms. The basic principle of dynamic programming with backtracking stays the same, but base pair maximization is replaced by thermodynamic considerations.

5.2 Minimizing the Free Energy

Thermodynamic energy models are based on the loop decomposition described in section 4.1. Each loop is assigned an energy value and the energy of the whole structure is the sum of the energies of all loops making up the structure. The goal of the folding algorithm is to find the secondary structure with the minimum free energy:

$$F_{\min} = \min_{S \in \mathcal{S}} (F(S)) \tag{5.1}$$

$F(S)$ denotes the free energy of a structure $S \in \mathcal{S}$, with $\mathcal{S}$ being the set of all possible secondary structures S.

MFE folding was first described by Michael Zuker and Patrick Stiegler [Zuker and Stiegler, 1981]. They make a major simplification by regarding multiloops as bifurcation loops, which means that they do not assign an energy to the multiloop itself, they just add up the energies of its constituent loops.

Today's models explicitly take multiloops into account. It makes no sense to determine multiloop energy values experimentally and then use tabulated values in the energy model because due to a combinatorial explosion the number of possible multiloops is simply too big. The models therefore often use a linear approach for the energy of a multiloop $\mathcal{M}$:

$$\mathcal{M} = a + b \cdot k + c \cdot u$$

a is the cost of initiating a multiloop, b is the penalty for each helix protruding from the loop, k is the degree of the loop, c is the penalty for each unpaired base in the loop, and the loop size u is the number of unpaired bases.

The standard energy model, as implemented for example in the *Vienna RNA package* [Hofacker et al., 1994], is based on the model by Zuker and Stiegler. In addition to scoring multiloops with the described linear approach, this model also decomposes multiloops unambiguously. This ensures that every structure is encountered exactly once, which is important for calculating the partition function as described in the next section.

The array $F_{i,j}$ stores the minimum free energy of all possible structures on the subsequence $s_i, \ldots, s_j$. The base s_i is either unpaired or paired with a base s_k (see Figure 5.2), leading to the following recursion for $F_{i,j}$:

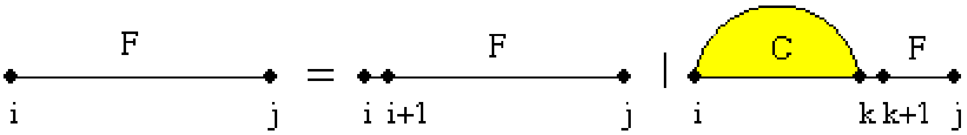

Figure 5.2: Recursion for $F_{i,j}$. (Image adapted from [Lorenz, 2007])

$$F_{i,j} = \min \begin{cases} F_{i+1,j} \\ \min\limits_{i<k\leq j} C_{i,k} + F_{k+1,j} \end{cases}$$

$C_{i,j}$ stores the minimum energy of the subsequence $s_i, \ldots, s_j$ provided that s_i and s_j form a base pair. Every base pair encloses either a hairpin loop, an interior loop, or a multiloop (see Figure 5.3). The recursion for $C_{i,j}$ therefore looks like this:

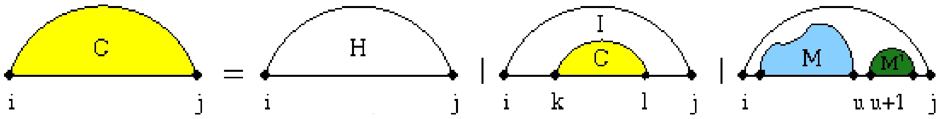

Figure 5.3: Recursion for $C_{i,j}$. (Image adapted from [Lorenz, 2007])

$$C_{i,j} = \min \begin{cases} \mathcal{H}(i,j) \\ \min\limits_{i<k<l<j} \mathcal{I}(i,j;k,l) + C_{k,l} \\ \min\limits_{i+1<u<j-1} M_{i+1,u} + M^1_{u+1,j-1} + a \end{cases}$$

The energy values $\mathcal{H}(i,j)$ for hairpin loops and $\mathcal{I}(i,j;k,l)$ for interior loops are tabulated. Multiloops are decomposed into a left part M, which contains at least one base pair, and a right part M^1 containing exactly one base pair. a is the cost of initiating a multiloop.

In $M_{i,j}$, the final base s_j is either unpaired, leading to a penalty of c for the enclosing multiloop, or it pairs with a base s_u, causing a penalty b for a base pair within a multiloop. $s_i, \ldots, s_{u-1}$ is then either unpaired ($u - i$ unpaired bases within a multiloop) or contains at least one base pair (and therefore has a minimum energy of $M_{i,u-1}$)(see Figure 5.4). This results in the following recursion for M:

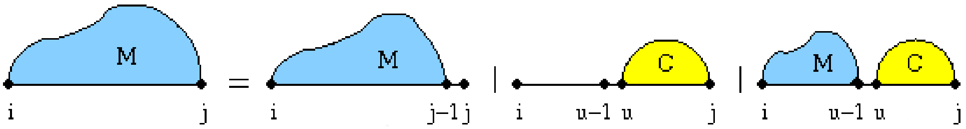

Figure 5.4: Recursion for $M_{i,j}$. (Image adapted from [Lorenz, 2007])

$$M_{i,j} = \min \begin{cases} M_{i,j-1} + c \\ \min\limits_{i \leq u < j} C_{u,j} + b + c(u-i) \\ \min\limits_{i < u < j} C_{u,j} + b + M_{i,u-1} \end{cases}$$

b is the penalty for a base pair within a multiloop and c the penalty for an unpaired base within a multiloop. The recursion for $M^1_{i,j}$ is simple, since s_i is paired and there is no further base pair. So s_j is either unpaired or it is paired with s_i (see Figure 5.5):

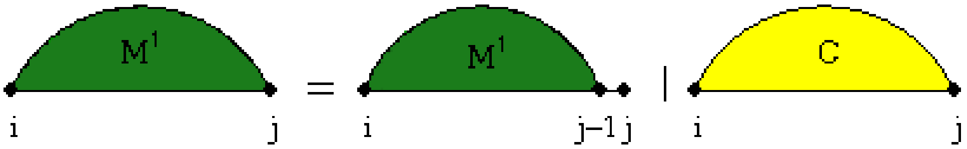

Figure 5.5: Recursion for $M^1_{i,j}$. (Image adapted from [Lorenz, 2007])

$$M^1_{i,j} = \min \begin{cases} M^1_{i,j-1} + c \\ C_{i,j} + b \end{cases}$$

All matrices are filled gradually, starting with the smallest feasible subsequences, which are pentanucleotides due to the minimum loop size for a hairpin loop which has to enclose at least three bases. Once the matrices are filled, the MFE can be found in $F_{1,n}$ and the corresponding MFE structure is determined via backtracking. The algorithm requires $\mathcal{O}(n^2)$ memory. When dealing with interior loops, the algorithm has to evaluate $\mathcal{I}(i, j; k, l)$ for all combinations of i, j, k and l. This would result in a time complexity of $\mathcal{O}(n^4)$. To reduce this time complexity and because very large interior loops are biologically irrelevant and energetically unfavorable, the maximum loop size of interior loops is usually set to some constant, resulting in an overall time complexity of $\mathcal{O}(n^3)$.

In addition, modern RNA secondary structure prediction algorithms usually take dangling ends into account. The bases immediately next to the end of helices can stack onto the helix and contribute favorably to the energy of the structure. Still missing from most modern approaches, however, is the inclusion of pseudoknots. While more and more pseudoknotted structures have been experimentally discovered, they are still being ignored by most RNA secondary structure prediction algorithms and software packages.

5.3 The Partition Function

MFE calculations only return information about a single secondary structure, the one with minimum free energy. But this is not the only secondary structure occurring in nature, in non-equilibrium states it might not even be the most common structure. RNAs are synthesized in an unfolded state and only later fold into a secondary structure. The multidimensional folding space contains many local minima which can be occupied by the RNA molecule. In order to answer questions about the space of all possible secondary structures, the likelihood of particular structures, or the frequency of structures with specific features, one has to look at the partition function Z. The equilibrium partition function contains information about the whole set $\mathcal{S}$ of possible secondary structures and is defined as

$$Z = \sum_{S \in \mathcal{S}} e^{\frac{-F(S)}{RT}} \tag{5.2}$$

with $F(S)$ denoting the free energy of a structure $S \in \mathcal{S}$. T is the absolute temperature and R the gas constant (a transformation of the Boltzmann constant if dealing with energies per mol). After introducing $\beta = \frac{1}{RT}$ for simplicity, the relative frequency $P(S)$ of a specific secondary structure S in equilibrium is then given by

$$P(S) = \frac{e^{-\beta F(S)}}{Z} \tag{5.3}$$

Similarly, the probability of various features, such as the frequency of a specific base pair, can be calculated by adding the probabilities of all structures containing that feature.

The number of possible secondary structures and therefore the number of summands in the partition function grows exponentially with increasing sequence length [Waterman, 1978]. Therefore it seems to be impossible to calculate Z in polynomial time, but McCaskill introduced a dynamic programming algorithm in [McCaskill, 1990] which achieves this computation. McCaskill's partition function algorithm is very similar to the MFE algorithm described previously and is implemented in the folding routines of the *Vienna RNA package.* The minima in the MFE algorithm are replaced by sums in the partition function algorithm, because every possible secondary structure and not just the one with minimum free energy contributes to the partition function. The additivity of free energies causes in a similar fashion the multiplicativity of the Boltzmann-weighted contributions to the partition function. Using the same decomposition of secondary structures as in the MFE approach described in section 5.2 is only possible because that decomposition is unambiguous and does not count any structures more than once. This property is not needed for MFE calculations, as only the structure with minimum energy is relevant, but it is necessary for the partition function because every single structure contributes to Z. Overall, this results in the following recursion for the partition function:

$$\begin{aligned}
Q_{i,j} &= Q_{i+1,j} + \sum_{i<k\leq j} Q^B_{i,k} \cdot Q_{k+1,j} \\
Q^B_{i,j} &= e^{-\beta\cdot\mathcal{H}(i,j)} \\
&\quad + \sum_{i<k<l<j} e^{-\beta\cdot\mathcal{I}(i,j;k,l)} \cdot Q^B_{k,l} \\
&\quad + \sum_{i+1<u<j-1} Q^M_{i+1,u} \cdot Q^{M^1}_{u+1,j-1} \cdot e^{-\beta\cdot a} \\
Q^M_{i,j} &= Q^M_{i,j-1} \cdot e^{-\beta\cdot c} \\
&\quad + \sum_{i\leq u<j} e^{-\beta\cdot(u-i)\cdot c} \cdot Q^B_{u,j} \cdot e^{-\beta\cdot b} \\
&\quad + \sum_{i<u<j-1} Q^M_{i,u} \cdot Q^B_{u+1,j} \cdot e^{-\beta\cdot b} \\
Q^{M^1}_{i,j} &= Q^{M^1}_{i,j-1} \cdot e^{-\beta\cdot c} + Q^B_{i,j} \cdot e^{-\beta\cdot b}
\end{aligned} \tag{5.4}$$

The matrix $Q_{i,j}$ is the analogon to $F_{i,j}$ in the MFE algorithm, storing the partition function of the subsequence $[i,j]$. Similarly $Q^B_{i,j}$ corresponds to $C_{i,j}$ and stores the partition function of the subsequence $[i,j]$ with (i,j) forming a base pair. Q^M and Q^{M^1} are equivalent to M and M^1 and are responsible for the calculation of multiloops. By summation over all Boltzmann-weighted energies, starting with the smallest subsequences, the algorithm generates the partition function $Z = Q_{1,n}$ of the whole sequence.

5.4 Structure Prediction Including Pseudoknots

The commonly used MFE-based RNA secondary structure prediction algorithms rely on the condition that the structure can be unambiguously decomposed into independent loops, i.e. that all secondary structure motifs are non-crossing and self-contained, which allows the application of the dynamic programming approach. Despite their importance, pseudoknots are usually excluded in this approach because they violate the restriction to non-crossing structure elements. While secondary structure prediction of pseudoknot-free structures with the MFE approach requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space [Zuker and Stiegler, 1981; Lyngsø et al., 1999], allowing for arbitrary pseudoknots under the loop-based energy model has been shown to be NP-complete [Akutsu, 2000; Lyngsø and Pedersen, 2000].

Interestingly, if instead of using the loop-based energy model, one only wants to maximize the number of base pairs, the problem can be solved: Akutsu proposed a dynamic programming algorithm for base pair maximization that requires $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^3)$ space [Akutsu, 2000]. Another approach used for the prediction of RNA pseudoknots and originating from graph theory is maximum weighted matching with a time complexity of $\mathcal{O}(n^3)$ [Tabaska et al., 1998; Ruan et al., 2004]. Maximum weighted matching was also used by Witwer in the *hxmatch* algorithm, which is very useful for predicting pseudoknots in multiple RNA sequences with known covariances [Witwer et al., 2004].

Instead of simplifying the energy model, one can also keep the well-proven combination of dynamic programming and the loop-based energy model and restrict the types of possible pseudoknots. Rivas and Eddy designed a secondary structure prediction algorithm which requires $\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ space [Rivas and Eddy, 1999]. Other algorithms, further restricting the range of predictable pseudoknots, reduced the runtime to $\mathcal{O}(n^5)$ using $\mathcal{O}(n^4)$ or $\mathcal{O}(n^3)$ space [Dirks and Pierce, 2003; Lyngsø and Pedersen, 2000].

A modern example for a dynamic programming algorithm for the prediction of pseudoknots is *pknotsRG*, developed by Jens Reeder and Robert Giegerich, which computes the MFE structure for canonical simple recursive pseudoknots in $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ space [Reeder and Giegerich, 2004].

Reeder and Giegerich define simple pseudoknots as H-type pseudoknots. If the three loops of this simple pseudoknot are allowed to fold internally, even into further pseudoknots, the resulting structure is called a simple recursive pseudoknot. Every simple recursive pseudoknot has a canonical representative which can be handled by *pknotsRG*. A simple recursive pseudoknot is canonical if it fulfills the following conditions:

- Both strands in a helix must have the same length and must not contain any bulges.
- The helices making up the pseudoknot must both have the maximum possible length.
- If the two maximal helices would overlap, their boundary is fixed at an arbitrary point between them.

Simple recursive pseudoknots are defined by the eight end points of the four sequence intervals making up the two stems of the pseudoknot. The canonization rules reduce these eight independent variables to four (see Figure 5.6), allowing a dynamic programming recursion to find the MFE structure of an RNA sequence possibly containing canonical simple recursive pseudoknots with a runtime of $\mathcal{O}(n^4)$.

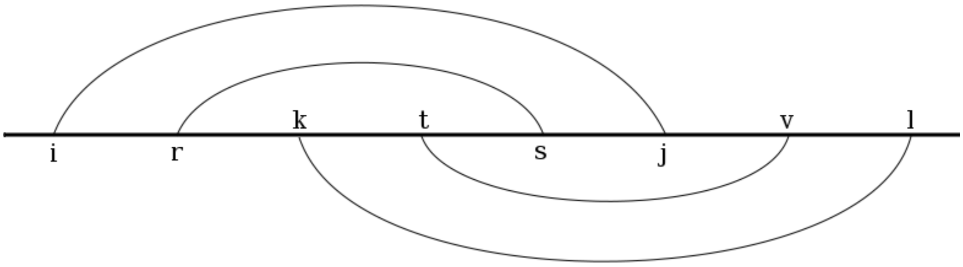

Figure 5.6: *pknotsRG* defines pseudoknots by their eight boundaries i, j, k, l, r, s, t and v. The canonization rules requiring bulge-free helices of maximum length reduce these eight variables to four (i, j, k and l).

Whenever possible, the energy model of *pknotsRG* uses the standard parameters for pseudoknot-free secondary structures. For pseudoknots, the stabilizing effect of nearest neighbor stacking energies of the pseudoknot helices and of dangling bases at the helix ends are being accounted for. The coaxial stacking of helices such as in H-type pseudoknots has also been incorporated into the energy model. Apart from that, pseudoknots are treated similar to multiloops in the standard energy model: they are assigned an initiation penalty of 9 kcal/mol. This value is relatively high in order to prevent the prediction of too many false positive pseudoknots. In addition, each unpaired nucleotide within a pseudoknot loop is penalized with 0.1 kcal/mol.

So far, *pknotsRG* is the fastest dynamic programming algorithm for the prediction of RNA pseudoknots. But it also restricts the type of pseudoknots, which it is able to predict, further than other dynamic programming algorithms (see Figure 5.7). While *pknotsRG* can handle more than one pseudoknot per sequence and even recursive pseudoknots, the biologically quite common cases of kissing hairpins or pseudoknots with helices containing bulges or interior loops are not covered by *pknotsRG*. Nevertheless, with its low computational complexity of $\mathcal{O}(n^4)$, it is the most useful dynamic programming pseudoknot prediction algorithm in many cases.

An alternative to dynamic programming are heuristic approaches. Heuristic methods can handle a wider class of pseudoknots and more complex energy models. However, they are not guaranteed to find the MFE structure. In the context of RNA secondary structure prediction including pseudoknots approaches using genetic algorithms [Gultyaev et al., 1995], stochastic context-free grammars [Cai et al., 2003], kinetic folding simulations [Xayaphoummine et al., 2003] and iterative stem adding procedures [Ruan et al., 2004] have all been employed.

A recent representative of the stem adding approach is *HotKnots*, which first generates a tree of energetically favorable structures by partially folding the input sequence with i and j paired, for all i and j with $j-1 > 3$ [Andronescu et al., 2010; Ren et al., 2005]. New base pairs are added using a dynamic programming algorithm leaving the already existing partial structures unchanged. For the calculation of the energy of folded structures *HotKnots* uses the model of Dirks and Pierce [Dirks and Pierce, 2003], which is based on the standard energy model for pseudoknot-free secondary structures. Its additional parameters concerning pseudoknots include three different penalties for initiating a pseudoknot, depending on whether the pseudoknot is not enclosed by another loop, nested within a multiloop, or nested within

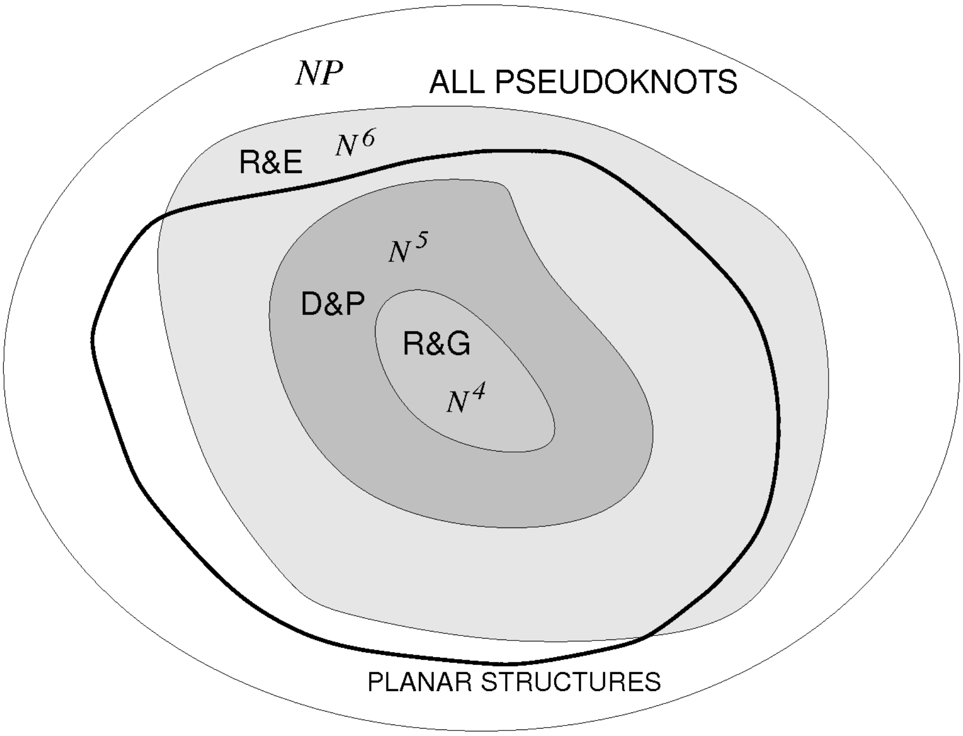

Figure 5.7: The classes of pseudoknots investigated by R&E [Rivas and Eddy, 1999], D&P [Dirks and Pierce, 2003] and R&G's *pknotsRG* [Reeder and Giegerich, 2004]. (Image reproduced from [Hofacker and Stadler, 2007])

another pseudoknot. Similar to the treatment of multiloops in the standard energy model, there are additional penalties for each unpaired base and each stem within a pseudoknot. The actual energy parameter values used in the latest version of *HotKnots* are not those published by Dirks and Pierce. Instead, improved values obtained by applying constraint generation and Boltzmann likelihood parameter estimation methods to a large dataset of RNA structures are used.

Pseudoknot detection algorithms are following another heuristic approach, in which pseudoknot candidates are generated and analyzed before folding the remaining sequence with the standard MFE approach. An example is *DotKnot* [Sperschneider and Datta, 2010], which uses the dot plot generated by the partition function of the *Vienna RNA package* as a starting point to pick stems for the construction of pseudoknots. First, two crossing stems are selected to form core H-type pseudoknots, which serve as the building blocks for more complex pseudoknots. Within the loops of these core H-type pseudoknots recursive secondary structure elements are allowed to form independently of each other, before the whole recursive H-type pseudoknot candidate is assembled and verified.

DotKnot uses three different pseudoknot energy models depending on the type of pseudoknot it encounters. In each case, the stabilizing effect of the pseudoknot helices are calculated via the standard loop decomposition energy model. Only the destabilizing entropic effect of the pseudoknot loops is then calculated via one of three methods: for pseudoknots with helices without bulge or interior loops and with a loop length between the two pseudoknot helices of only 0 or 1 nucleotide, the original energy model by Cao and Chen [Cao and Chen, 2006] is used. If the loop length is between 2 and 6 nucleotides, Cao and Chen's extended energy model [Cao and Chen, 2009] is used. Finally, for pseudoknots with a longer loop connecting the two pseudoknot helices or for pseudoknots with bulge or interior loops within their helices, a simple heuristic energy model is used. This heuristic model penalizes the initiation of a pseudoknot loop with 7 kcal/mol and each unpaired nucleotide within a pseudoknot loop with 0.1 kcal/mol.

The three algorithms described in this section, the dynamic programming algorithm *pknotsRG*, and the heuristic algorithms *HotKnots* and *DotKnot*, will be used as benchmarks to compare the results of *PKplex* against in section 7.1. *pknotsRG* was chosen because it is the leading dynamic programming solution to the problem of pseudoknot prediction. And although dynamic programming and heuristic algorithms are not fully comparable, *HotKnots* and *DotKnot* are included in the analysis as well to show the strengths and weaknesses of the different general approaches.

Chapter 6

PKplex

In this chapter I am describing *PKplex*, a new dynamic programming RNA secondary structure prediction algorithm which takes pseudoknots into account. In the *PKplex* model, the thermodynamics of an RNA pseudoknot essentially consist of two components: the energy necessary to make the residues of a potential pseudoknot accessible, i.e. unpaired, which is calculated with the algorithm used in *RNAplfold* [Bernhart et al., 2006], and the energy gained from the base pairing of the nucleotides involved in the pseudoknot interaction. A dynamic programming routine based on the *RNAplex* [Tafer and Hofacker, 2008] algorithm combines these two energy values to compute the optimal pseudoknot for a given RNA sequence.

In the next chapter I am taking a look at the strengths and limitations of *PKplex*, present the results of evaluating the algorithm on a broad set of known RNA structures, both with and without pseudoknots, and compare *PKplex* with other published RNA pseudoknot prediction algorithms.

6.1 Algorithm

The *PKplex* algorithm operates within the framework of the *Vienna RNA package* [Hofacker et al., 1994]. It uses the standard RNA energy model described e.g. by Mathews and Turner [Mathews et al., 1999; Turner and Mathews, 2009] and is based on the classic RNA-folding algorithm by Zuker and Stiegler [Zuker and Stiegler, 1981]. The recursions for the equilibrium partition function are based on those suggested by McCaskill [McCaskill, 1990].

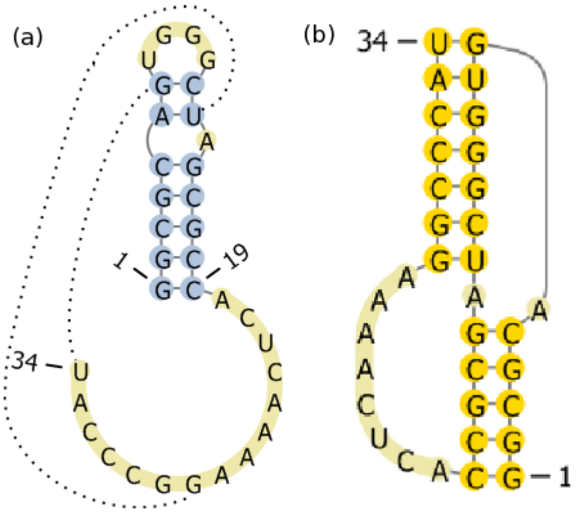

Figure 6.1: Two different secondary structures for the same RNA sequence, one without pseudoknots (a), and the other with a single big pseudoknot (b). The dotted lines in (a) indicate the area of the pseudoknot interaction shown in (b). (Image generated with Pseudoviewer3 [Byun and Han, 2009])

Even though there are plenty of RNAs which contain pseudoknots, the fraction of nucleotides which are part of a pseudoknot is generally pretty low. That is because the pseudoknot usually comprises only a small part of the whole sequence. Since we already know how to predict pseudoknot-free structures decently, the approach used in *PKplex* starts with a pseudoknot-free secondary structure and then potentially adds just a single pseudoknot. The actual kinetic RNA folding process in living cells often follows the same sequential order: the pseudoknot-free secondary structure is built first, and only then the sequence forms its pseudoknotted base pairs.

Other pseudoknot-prediction algorithms often have to avoid predicting too many false positive pseudoknots. *PKplex* somewhat bypasses this problem by predicting at most one pseudoknot per sequence. However, if we are interested in multiple pseudoknots on a single sequence, it is possible to extend *PKplex* in such a way as to run multiple times iteratively on the same sequence and therefore generating structures containing more than one pseudoknot.

PKplex constructs pseudoknots by starting with a pseudoknot-free secondary structure and then picking two non-overlapping intervals and letting the bases within those two intervals form pairs between each other - independent of the base pairing of the remainder of the sequence. If those two intervals do not lie within the same loop of the pseudoknot-free structure, this results in the formation of a pseudoknot (see Figure 6.1). This approach causes all secondary structures predicted by *PKplex* to contain at most one pseudoknot. Because the frequency of pseudoknots is low, multiple pseudoknots on a single sequence are very rare, and prediction accuracy does not suffer much by this simplification. In addition, during the construction of the pseudoknot, base pairing is only allowed between the two sequence intervals, but not within one of the intervals. The resulting pseudoknot stems can therefore contain bulge and interior loops, but multiloops are not possible. Since most known pseudoknot stems are quite short, this restriction should not impair the prediction quality by much either.

The secondary structures covered by *PKplex* can be decomposed into a pseudoknot-free structure and additional base pairs responsible for forming the pseudoknot, i.e. one of the stems of the pseudoknot. The free energy ΔG of an RNA structure is then calculated as the sum of ΔG_{pkfree} and ΔG_{pk}, the energy of the pseudoknot-free structure and the energy of the pseudoknot itself.

$$\Delta G = \Delta G_{pkfree} + \Delta G_{pk}$$

In the *PKplex* model, ΔG_{pk} consists of ΔG_{int}, the energy gained through the base pair interactions and c_{pk}, the cost of initiating a pseudoknot, which is constant.

$$\Delta G_{pk} = \Delta G_{int} + c_{pk}$$

ΔG_{pkfree} is the energy of the pseudoknot-free part of the secondary structure. The two intervals involved in the pseudoknot interaction cannot take part in the pseudoknot-free structure and therefore have to be accessible and unpaired. *PKplex* models ΔG_{pkfree} as

$$\Delta G_{pkfree} = \Delta G_{MFE} + \Delta G_u[k,i][j,l]$$

$$\Delta G_{pkfree} \approx \Delta G_{MFE} + \Delta G_u[k,i] + \Delta G_u[j,l]$$

with ΔG_{MFE} being the the pseudoknot-free MFE and $\Delta G_u[k,i]$ and $\Delta G_u[j,l]$ being the energies needed to render the two pseudoknot intervals $[k,i]$ and $[j,l]$ unpaired and accessible. The approximation of *PKplex* assumes that the energy of rendering two different non-overlapping intervals accessible is the sum of rendering each interval accessible individually, i.e. that these two processes are independent of each other.

In order to construct the optimum pseudoknot *PKplex* therefore searches for the interacting intervals $[k,i]$ and $[j,l]$, which minimize the expression $\Delta G_u[k,i] + \Delta G_u[j,l] + \Delta G_{int}$.

6.1.1 Calculation of Accessibility

$\Delta G_u[a,b]$, the energy necessary to render the sequence interval $[a,b]$ single-stranded and accessible can be calculated from $P_u[a,b]$, the probability that $[a,b]$ is unpaired via

$$\Delta G_u[a,b] = -RT \ln(P_u[a,b]) \tag{6.1}$$

$P_u[a,b]$ is calculated for all subsequences by *RNAplfold* [Bernhart et al., 2006; Bompfünewerer et al., 2008], which is part of the *Vienna RNA Package*, following the approach first used in *RNAup* [Mückstein et al., 2006; Mückstein et al., 2008].

The equilibrium partition function Z is defined as

$$Z = \sum_S e^{-\beta F(S)} \tag{6.2}$$

with $F(S)$ denoting the free energy of a secondary structure S and β denoting the inverse of the temperature times Boltzmann's constant. According to equation 5.3 we can use the partition function to calculate the probability of a sequence interval being unpaired via

$$P_u[a,b] = \frac{1}{Z} \sum_{S \in \mathcal{S}^u_{[i,j]}} e^{-\beta F(S)} \tag{6.3}$$

with $\mathcal{S}^u_{[i,j]}$ denoting the set of secondary structures with the interval $[i, j]$ being unpaired. $[i, j]$ is either part of the external loop of the secondary structure or part of a loop enclosed by a base pair (p, q). $P_u[i, j]$ can therefore be expressed as

$$P_u[i, j] = \frac{Z(1, i-1)Z(j+1, n) + \sum_{p<i} \sum_{j<q} \hat{Z}(p, q)Z_{pq}[i, j]}{Z(1, n)} \tag{6.4}$$

The first term handles the case of $[i, j]$ being part of the external loop by multiplying the partition functions of all sub-structures upstream and downstream of $[i, j]$. The second term handles the case of $[i, j]$ being part of a loop enclosed by the base pair (p, q). $\hat{Z}(p, q)$ denotes the partition function outside the base pair (p, q) and $Z_{pq}[i, j]$ the partition function inside (p, q) given that the interval $[i, j]$ is unpaired. The restricted partition functions $\hat{Z}(p, q)$ and $Z_{pq}[i, j]$ are then further decomposed according to the type of loop containing the unpaired interval $[i, j]$.

Depending of the details of the implementation one or more additional matrices compared to the partition function algorithm of McCaskill have to be stored. In *RNAplfold* the CPU requirements have been reduced to $\mathcal{O}(n^3)$ and for very long sequences a sliding window technique which only looks at sequence intervals of fixed length instead of at the whole sequence can be used [Bernhart, 2009]. The implementation of the *PKplex* algorithm directly calls *RNAplfold* to calculate $P_u[i, j]$ which is then transformed into $\Delta G_u[a, b]$ via equation 6.1.

6.1.2 Calculation of the Interaction Energy

ΔG_{pk} is calculated by a dynamic programming recursion based on the one used in *RNAplex* [Tafer and Hofacker, 2008] to evaluate RNA-RNA interactions. Every potential pseudoknot can be described as an interaction between two intervals $[k, i]$ and $[j, l]$ with both (i, j) and (k, l) forming base pairs. *PKplex* uses the table $C_{i,j,k,l}$ to store the best energy of an interaction between those two intervals. $C_{i,j,k,l}$ is filled via the following recursion (see Figure 6.2):

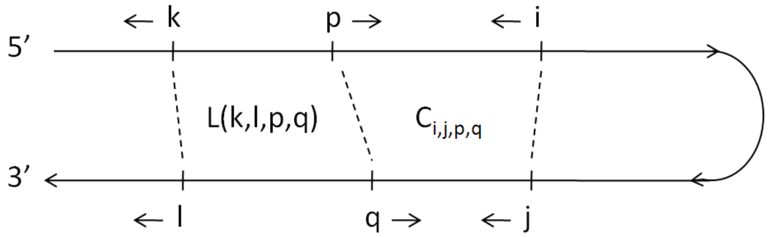

Figure 6.2: Decomposition of a pseudoknot spanning the intervals $[k, i]$ and $[j, l]$. (i, j), (k, l) and (p, q) each form a base pair. The minimum energy of an interaction between $[p, i]$ and $[j, q]$ has been already been calculated and is stored in the table $C_{i,j,p,q}$. $L(k, l, p, q)$, the energy of the loop enclosed by (k, l) and (p, q) can be calculated directly.

$$C_{i,j,k,l} = \min_{\substack{k<p\leq\min(i,k+v)\\ \max(j,l-v)\leq q<l}} (C_{i,j,p,q} + L(k, l, p, q)) \tag{6.5}$$

with $L(k, l, p, q)$ being the energy of a loop enclosed by the base pairs (k, l) and (p, q), and v being the maximum size of an interior/bulge loop. The pseudocode for this recursion is shown in Figure 6.3.

```
for(i=n...1)
  for(j=i+4...n)
    for(k=i-1...1)
      for(l=j+1...n)
        for(p=k+1...min(i,k+v))
          for(q=l-1...max(j,l-v))
            E = LoopEnergy(p,q;k,l)
            C(i,j,k,l) = min(C(i,j,k,l), C(i,j,p,q)+E)
```

Figure 6.3: Pseudocode for the calculation of the interaction energy.

$C_{i,j,k,l}$ is initialized by setting $C_{i,j,i,j}$ to the pseudoknot initialization cost constant c_{pk}. After the matrix $C_{i,j,k,l}$ is filled, $\Delta G_u[i, j]$ and $\Delta G_u[k, l]$, the energies of rendering the involved intervals unpaired and accessible, which have been previously calculated by *RNAplfold* and stored for later use, are added - resulting in $\Delta G(i, j, k, l)$, the free energy change for the complete pseudoknot interaction. If the minimum of $\Delta G(i, j, k, l)$ is greater than 0, no energetically favorable pseudoknot has been found and the optimum pseudoknot-free structure according to the *Vienna RNA package's RNAfold* is returned. Otherwise backtracking is employed to get the structure of the pseudoknot. The non-pseudoknot part of the structure is generated by *RNAfold*, using constraints to force the pseudoknot intervals $[k, i]$ and $[j, l]$ to be not involved in the formation of the remainder of the secondary structure.

6.2 Time Complexity and Implementation

The time complexity of calculating the accessibilities $\Delta G_u[a, b]$ is $\mathcal{O}(n^3)$. The recursion for ΔG_{pk} uses 6 loops for the 6 variables i, j, k, l, p and q. A naïve implementation would therefore require $\mathcal{O}(n^6)$ computational time. Since the pseudoknot helices of known pseudoknots are generally of very limited length (the longest continuous pseudoknotted helix in the full dataset is only 9 base pairs long), *PKplex* uses a parameter w to limit the maximum length of the intervals $[k, i]$ and $[j, l]$ involved in the pseudoknot formation. The CPU time required for the ΔG_{pk} recursion then scales with $\mathcal{O}(n^2 w^4)$, because i and j run over the whole sequence, whereas the ranges of values for k, l, p and q are limited by w. Adding up both components of the algorithm, the runtime for *PKplex* as a whole scales with $\mathcal{O}(n^3 + n^2 w^4)$.

Instead of storing the whole four-dimensional table $C_{i,j,k,l}$, it is sufficient to only store a single two-dimensional table $\hat{C}_{k,l}$ for every combination of i and j which can be discarded as soon as the calculations for the next combination of i and j are started. The memory consumption of the recursion for ΔG_{pk} is therefore $\mathcal{O}(w^2)$.

The pseudocode of Figure 6.3 shows that the loop energy $L(k, l, p, q)$ is calculated in every step of the innermost loop. Although the loop energy function is constant in its time complexity, the amount of times this function is called still makes this a time consuming step. $L(k, l, p, q)$ is called with exactly the same parameter values more than once for different combinations of i and j, therefore unnecessarily repeating some calculations.

I managed to reduce the overall runtime of *PKplex* by reordering the loops in such a way that $L(k, l, p, q)$ is called less often, because it is not part of the innermost loop anymore. This reordering requires the addition of a dimension to the storing table and keeping track of $C_{j,k,l}$. This increases the memory requirement of the ΔG_{pk} recursion to $\mathcal{O}(nw^2)$, which does not cause any problems since calculating the accessibilities already requires $\mathcal{O}(n^2)$ memory space. This loop rearrangement reduces the runtime due to fewer calls of the loop energy function. The time complexity of the whole algorithm stays unchanged though, since the loop energy function is of constant time. The pseudocode for the recursion after implementing the described changes is shown in Figure 6.4.

```
for(i=n...1)
  for(k=i-1...i-w)
    for(l=i+5...n)
      for(p=k+1...min(i,k+v))
        for(q=l-1...max(i+4,l-v))
          E = LoopEnergy(p,q;k,l)
          for(j=max(i+4,l-w)...q)
            C(j,k,l) = min(C(j,k,l), C(j,p,q)+E)
```

Figure 6.4: Pseudocode for the calculation of the interaction energy with rearranged loop order to reduce the runtime.

With a time complexity of $\mathcal{O}(n^3 + n^2w^4)$ *PKplex* should theoretically compare favorably with other dynamic programming algorithms for pseudoknot prediction. Both computation time and the quality of its results are analyzed in the following chapter.

Chapter 7

Results and Discussion

7.1 Results

7.1.1 Accuracy Measures

To evaluate the quality of a predicted secondary structure, the prediction is compared to the experimentally determined true structure (called reference structure). The following statistical measures are used to evaluate the prediction quality:

$$\begin{aligned}\text{Sensitivity} &= \frac{\text{correctly predicted base pairs}}{\text{reference structure base pairs}} \\ &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}\end{aligned} \tag{7.1}$$

$$\begin{aligned}\text{Selectivity} &= \text{Positive prediction value (PPV)} \\ &= \frac{\text{correctly predicted base pairs}}{\text{predicted base pairs}} \\ &= \frac{\text{true positives}}{\text{true positives} + \text{false positives}}\end{aligned} \tag{7.2}$$

$$\text{F-measure} = \frac{2 \times \text{Sensitivity} \times \text{PPV}}{\text{Sensitivity} + \text{PPV}} \tag{7.3}$$

The F-measure is the harmonic mean of sensitivity and selectivity. It is close to the arithmetic mean of sensitivity and PPV when those two numbers are close to each other, but smaller than their arithmetic mean when they are further apart.

7.1.2 Results of PKplex for Selected Sample Sequences

The 35-nucleotide sequence `HIVRT32` is an inhibitor of the human immunodeficiency virus type 1 reverse transcriptase (HIV-1-RT) and contains a simple H-type pseudoknot [Tuerk et al., 1992]. *PKplex* is able to perfectly predict its secondary structure, while the pseudoknot-free prediction with *RNAfold* only returns one of the two stems of the pseudoknot:

```
HIVRT32
Sequence:  UCAAGUAUUCCGAAGCUCAACGGGAAAAUGAGCUA
Reference: .......[[[[[.((((((.]]]]]...)))))).
PKplex:    .......[[[[[.((((((.]]]]]...)))))).
RNAfold:   .............((((((.........)))))).
```

Figure 7.1 shows squiggle plots of the reference structure/*PKplex* prediction and the *RNAfold* prediction and compares the two different structures in an arc plot. For `HIVRT32` *PKplex* simply adds a stem to the pseudoknot-free MFE structure. This addition does not change the structure of the remainder of the sequence at all since the bases involved in the additional pseudoknotted stem are already accessible in the MFE structure.

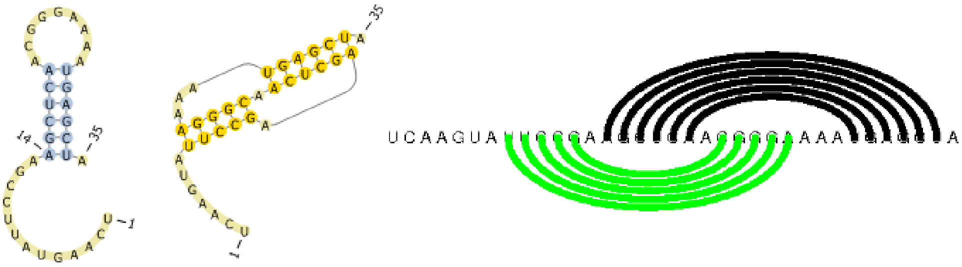

Figure 7.1: The secondary structure of `HIVRT32` as predicted by *RNAfold* (left) and its true secondary structure as predicted by *PKplex* (middle). The arc plot (right) shows the base pairs contained in both structures in black and the base pairs contained only in the *PKplex* prediction in green. (Squiggle plots generated with Pseudoviewer3 [Byun and Han, 2009], arc plot generated with the help of *hxmatch* [Witwer et al., 2004])

The sequence `MMTV` contains a short frameshifting pseudoknot of the mouse mammary tumor virus [Theimer and Giedroc, 2000]. *PKplex* again manages to predict the pseudoknot, but in this case it adds a single false positive base pair to one of the pseudoknot stems. Contrary to the first example though, *RNAfold* predicts a completely different structure for `MMTV`. The *PKplex* algorithm therefore first renders the site of the pseudoknot stem accessible and unpaired, then adds the pseudoknotted stem and finally lets *RNAfold* fold the remainder of the sequence, which results in a totally different fold overall. The following lines show the different structures for `MMTV` in dot-bracket notation:

```
MMTV
Sequence:   GGGGCAGUCCCCUAGCCCCACUCAAAAGGGGGAU
Reference:  [[[[[..((((((.]]]]]........)))))).
PKplex:     [[[[[.(((((((.]]]]]........)))))))
RNAfold:    ((((....))))...((((.((....))))))..
```

Figure 7.2 displays squiggle plots of the different secondary structures for `MMTV` and compares the structures in an arc plot.

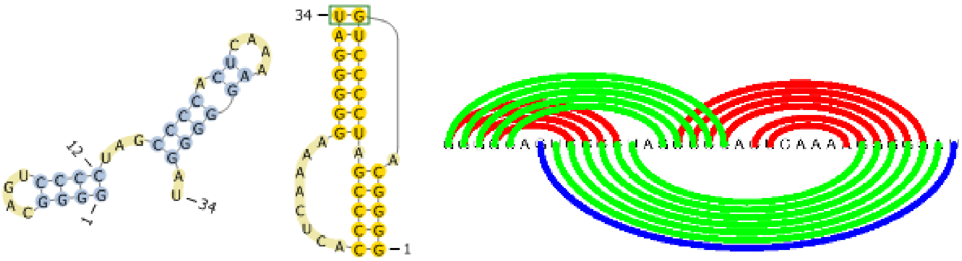

Figure 7.2: The secondary structure of `MMTV` as predicted by *RNAfold* (left) and by *PKplex* (middle). The single false positive base pair compared to the reference structure is marked with a green box. The arc plot (right) shows the base pairs of the *RNAfold* prediction in red and the base pairs of the *PKplex* prediction in green (true positives) and blue (false positives). (Squiggle plots generated with Pseudoviewer3 [Byun and Han, 2009], arc plot generated with the help of *hxmatch* [Witwer et al., 2004])

PKplex does not improve the prediction results compared to *RNAfold* for all sequences though. An example is the pseudoknot of the human telomerase RNA [Chen et al., 2000]. The following lines show the dot-bracket notation of reference structure, *PKplex* prediction and *RNAfold* prediction:

```
telo.human
GGGUUGCGGAGGGUGGGCCUGGGAGGGGUGGUGGCCAUUUUUUGUCUAACCCUAACUGAGAAGGGCGUAGGCGCCG
UGCUUUUGCUCCCCGCGCGCUGUUUUUCUCGCUGACUUUCAGCGGGCGGAAAAGCCUCGGCCUGCCGCCUUCCACC
```

```
GUUCAUUCUAGAGCAAACAAAAAAUGUCAGCUGCUGGCCCGUUCGCCCCUCCCGGGGA
Reference:
.................((((((((((((((.(((((...........................((((((((....
.(((((.......(((((((((........[[[[[[[[[)))))))))..)))))....))))).)))........
.....................]]].]]]]]]...)))))...))))))))))))))..
PKplex:
.......[[[[[[[[[.((((((((((((((.(((((....................(((((((((...(((((((
.((....))....)).))))))))))))))((((((......(((((.((......)).)))))]]]]]]]]]...
((((......))))...........))))))...)))))...))))))))))))))..
RNAfold:
.................((((((((((((((.(((((((((((((...........((.(((((.((..((.((((
.((((((.....((((.(((((....((.....))....))))).))))))))))..))))))..)))))))))..
((((......))))..))))))))..........)))))...))))))))))))))..
```

Arc plots comparing these structures can be seen in Figure 7.3. *PKplex* adds a pseudoknot stem for `telo.human` which does not occur in the reference structure. This predicted pseudoknot causes the remainder of the structure to fold differently than in the pseudoknot-free *RNAfold* prediction. Overall, the *PKplex* prediction is worse than the *RNAfold* prediction for `telo.human` as measured by sensitivity (0.50 vs. 0.68), selectivity (0.40 vs. 0.54) and F-measure (0.44 vs. 0.60).

These three examples offer only a snapshot of the predictions of *PKplex*. For a proper evaluation of the algorithm, we compared its results with those of four other secondary structure prediction algorithms over a big dataset containing both pseudoknotted and pseudoknot-free RNA structures.

7.1.3 Comparison with Other Algorithms

I am comparing the results of *PKplex* with the results of three other programs for pseudoknot prediction and with one pseudoknot-free secondary structure prediction software: *pknotsRG* by Reeder and Giegerich [Reeder and Giegerich, 2004] is based on a dynamic programming algorithm and requires $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ space. *HotKnots* uses a heuristic approach and has been described by Ren et al [Ren et al., 2005]. I am using version 2.0 with improved parameter values [Andronescu et al., 2010]. *DotKnot* by Sperschneider and Datta [Sperschneider and Datta, 2010] is a heuristic pseudoknot detection algorithm. It only predicts pseudoknots, therefore I use *RNAfold* to predict the remaining pseudoknot-free structure. And fi-

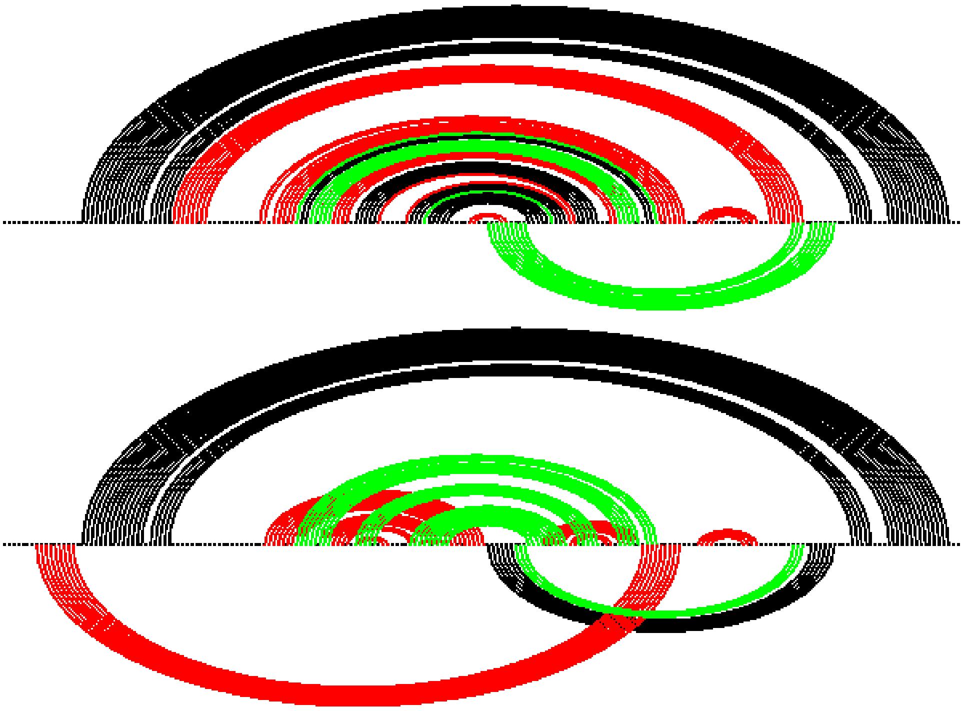

Figure 7.3: Arc plots comparing the reference structure for `telo.human` to the predictions of *RNAfold* (top) and *PKplex* (bottom). In both plots base pairs contained in both structures (true positives) are black, base pairs contained only in the reference structure (false negatives) are green, and base pairs contained only in the structure prediction (false positives) are red. (Images generated with the help of *hxmatch* [Witwer et al., 2004])

nally, the *Vienna RNA package's RNAfold* [Hofacker et al., 1994], which only predicts pseudoknot-free structures, is included in the analysis as well.

7.1.4 Data Sets

I evaluated the different pseudoknot prediction approaches by applying them to RNA sequences with known secondary structures and comparing the predictions with the reference structures. The full data set consists of 2253 sequences, both pseudoknotted and pseudoknot-free, and is the union of the S-Train and S-Test data sets used in [Andronescu et al., 2010]. The sequence-structure pairs were originally taken from the RNA STRAND v2.0 database [Andronescu et al., 2008] and from Pseudobase [van Batenburg et al., 2001]. About 80% of the sequences are pseudoknot-free, other properties are shown in Table 7.1. To evaluate the prediction quality on different types of structures the data set was split into four subsets: short pseudoknotted sequences (length <100 nucleotides [nt]), short pseudoknot-free sequences, long pseudoknotted sequences and long pseudoknot-free sequences.

Data set	Description	Number of sequences	Percentage (%)	Average length	Standard deviation
Complete	All sequences	2253	100.0	74.1	40.8
Short PK	<100nt, pseudoknotted	327	14.5	46.7	19.6
Short PKfree	<100nt, pseudoknot-free	1358	60.3	57.4	23.7
Long PK	≥100nt, pseudoknotted	98	4.3	148.2	41.4
Long PKfree	≥100nt, pseudoknot-free	470	20.9	125.2	23.9

Table 7.1: Properties of the data sets used for testing the pseudoknot prediction algorithms.

7.1.5 Prediction Accuracy

Table 7.2 shows the results of applying each of the five different secondary structure prediction algorithms on the test dataset and its subsets. The parameter values for *PKplex* were chosen as follows: c_{pk}, the pseudoknot initialization cost, was set to 8.1 kcal/mol, which results in a good balance between the quality and the quantity of the pseudoknots returned by *PKplex*. The maximum length w of the interacting intervals was set to 12 and the maximum size v of internal and bulge loops was set to 10. Larger values did not improve the quality of results and only increased the runtime of *PKplex*.

Dataset		PKplex	pknotsRG	DotKnot	HotKnots	RNAfold
Complete	Sensitivity	0.748	0.735	0.694	0.791	0.699
	PPV	0.690	0.696	0.656	0.762	0.684
	F-measure	0.718	0.715	0.675	0.776	0.691
	Containing PK (%)	25.5	22.4	52.2	17.1	0
	Runtime (s)	192	140	2100	33290	20.9
ShPK	Sensitivity	0.712	0.779	0.797	0.744	0.501
	PPV	0.741	0.779	0.801	0.766	0.659
	F-measure	0.726	0.779	0.799	0.755	0.569
	Containing PK (%)	55.1	71.9	79.8	64.2	0
	Runtime (s)	10.8	6.0	186	367	2.1
ShPKfree	Sensitivity	0.774	0.743	0.701	0.814	0.752
	PPV	0.700	0.695	0.651	0.776	0.705
	F-measure	0.735	0.718	0.675	0.794	0.728
	Containing PK (%)	12.4	7.2	42.6	7.7	0
	Runtime (s)	64.7	32.0	796	2354	12.4
LoPK	Sensitivity	0.640	0.594	0.606	0.597	0.496
	PPV	0.601	0.556	0.597	0.599	0.510
	F-measure	0.620	0.558	0.601	0.598	0.503
	Containing PK (%)	80.6	60.2	92.9	53.1	0
	Runtime (s)	28.3	33.7	502	16671	2.1
LoPKfree	Sensitivity	0.722	0.719	0.619	0.798	0.727
	PPV	0.642	0.668	0.582	0.755	0.678
	F-measure	0.679	0.692	0.600	0.776	0.701
	Containing PK (%)	31.5	24.0	51.9	4.3	0
	Runtime (s)	91.6	69.9	677	13983	7.7

Table 7.2: Results of applying the five different prediction algorithms on the data sets.

For the complete dataset the quality of *PKplex'* results as judged by the F-measure is about the same as the quality of the results derived with the only other dynamic programming approach, *pknotsRG*. Looking at the heuristic approaches, *HotKnots* returns better results and *DotKnot* returns worse results than *PKplex*. But the computation time is a lot longer for the heuristic approaches than for the dynamic programming approaches. Interestingly, even the pseudoknot-free secondary structure prediction algorithm of *RNAfold* delivers better results than *DotKnot* and is not far behind the other algorithms.

When analyzing the predictions for the different subsets of sequences it turns out that *RNAfold* performs worse than all pseudoknot-prediction algorithms on the data sets that contain pseudoknotted sequences, whereas the opposite is not the case: the secondary structure prediction algorithms that include pseudoknots do not necessarily perform worse than *RNAfold* on data sets that do not contain any pseudoknotted sequences. For ShPKfree, the subset containing short pseudoknot-free sequences, *PKplex* returns marginally better results than *RNAfold*. This means that for the sequences where *PKplex* wrongly predicts a pseudoknot, the quality of these predictions is on average still better than the prediction quality of *RNAfold* for these sequences.

PKplex performs equally well over all four subsets. The F-measure for the data sets containing longer sequences is lower than that for the shorter sequences. That is a trend that is true for all methods and a consequence of the fact that secondary structure prediction difficulty increases with sequence length.

7.1.6 Computational Performance

The $\mathcal{O}(n^3)$ algorithm of *RNAfold* takes only about 21 s to fold all 2253 sequences of the complete data set (on an Intel Core 2 Quad Q6600 with 2.4 GHz and 4 GB RAM). The dynamic programming methods *pknotsRG* and *PKplex* take 140 s and 192 s respectively, while both heuristic methods compare unfavorably in this regard with runtimes of 2100 s for *DotKnot* and 33290 s for *HotKnots* (Table 7.2).

Figure 7.4 shows the runtime of the different secondary structure prediction algorithms when applied to 20 random RNA sequences of various lengths. *PKplex* compares favorably, especially for longer sequences, where it is the fastest of all tested pseudoknot prediction algorithms. For even longer sequences, such as the 1542 bp 16 s *E. coli* rRNA, the differences in runtime are substantial. While *PKplex'* runtime is 98 s, *DotKnot* already requires 1404 s and *pknotsRG* 2811 s.

To summarize this analysis, *PKplex* is the fastest of the tested pseudoknot prediction algorithms. For shorter sequences the other dynamic programming algorithm *pknotsRG* is a little bit faster than *PKplex*, but for longer sequences, where differences in runtime are a lot more important, *PKplex* is faster due to its lower time complexity of $\mathcal{O}(n^3 + n^2w^4)$ versus $\mathcal{O}(n^4)$ for *pknotsRG*. This low time complexity makes *PKplex* the fastest of all known dynamic programming pseudoknot prediction algorithms. The heuristic approaches cannot compete with either of the two tested dynamic programming algorithms regarding computation time and *HotKnots* in particular is impractical for longer RNA sequences.

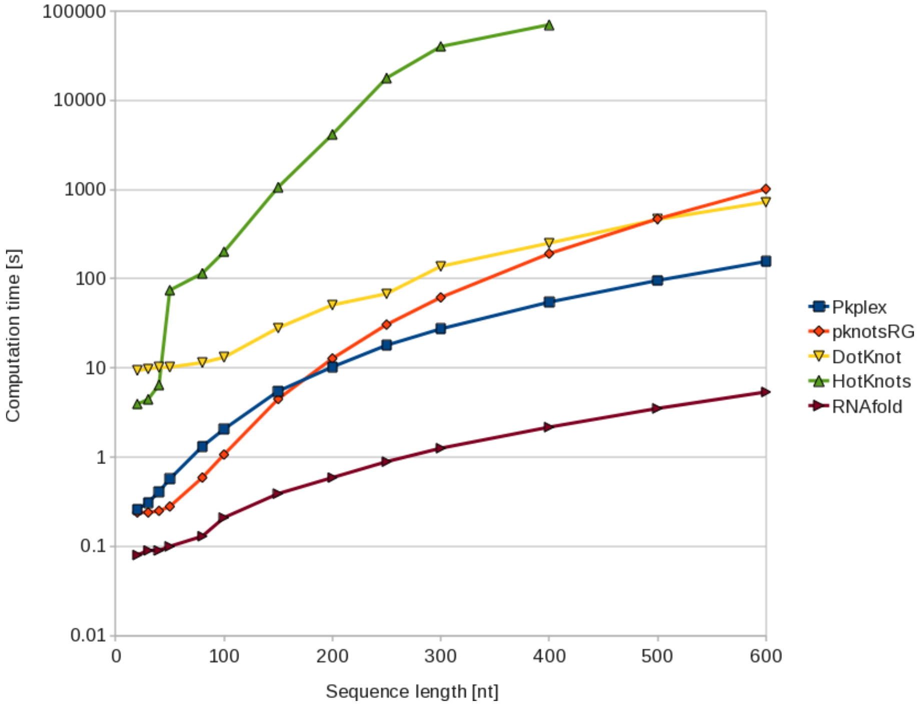

Figure 7.4: Comparison of the runtime of the secondary structure prediction algorithms for folding 20 random RNA sequences of different lengths.

The bandwidth of differences in the quality of the results of the tested methods is a lot smaller than the differences in runtime. The F-measure of the results of calculating the secondary structures of all 2253 sequences in the data set varies between 0.67 and 0.78 for the different methods. *HotKnots*, which is the slowest algorithm by far, returns the best results. It is followed by the faster dynamic programming approaches *PKplex* and *pknotsRG*. *RNAfold* and *DotKnot* occupy the lower end of that spectrum. As a consequence I recommend using *HotKnots* for pseudoknot prediction of short RNA sequences. For longer sequences or bigger datasets where *HotKnots* becomes impractical due to its long runtime, or if one wants to use a dynamic programming approach instead of a heuristic approach, *PKplex* is the recommended method for pseudoknot prediction.

7.2 Discussion

The general idea behind the *PKplex* algorithm - calculating the cost of rendering two sequence intervals accessible and adding the gain of letting those two intervals form base pairs among each other - is very simple, but it also implies some assumptions and restrictions of the algorithm which I want to mention in this section:

To calculate $\Delta G_u[a,b]$, the energy necessary to make a sequence interval accessible, *PKplex* employs the algorithm used in *RNAplfold*. This algorithm takes an approach based on the partition function, whereas the other parts of *PKplex* are based on MFE calculations. During the development process I also tried an MFE based approach to calculate $\Delta G_u[ab]$, but I found that this did not make a difference on the quality of the results. Because the implementation based on the partition function is faster, there is no reason to use an MFE-based approach instead. *PKplex* therefore returns the energetically most favorable pseudoknot for the ensemble of pseudoknot-free structures, not the most favorable pseudoknot for the pseudoknot-free MFE structure.

In the *PKplex* algorithm, pseudoknots are constructed by forming base pairs between two continuous sequence intervals in addition to the regular pseudoknot-free secondary structure. This results in H-type pseudoknots, kissing hairpins and other pseudoknots with only a single helix violating the condition of perfectly nested structures. The pseudoknotted structures returned by *PKplex* are therefore all bi-secondary structures. Non-planar pseudoknots or k-partite structures with $k > 2$ cannot be predicted by *PKplex*. For the same reason *PKplex* cannot predict multiple pseudoknots on a single sequence either. Because a high percentage of the known pseudoknotted sequences are planar and contain only a single pseudoknot, this is not a severe restriction. It should also be possible to adapt the *PKplex* algorithm to run multiple times on the same sequence, enabling it to predict more complex multiple stem pseudoknots.

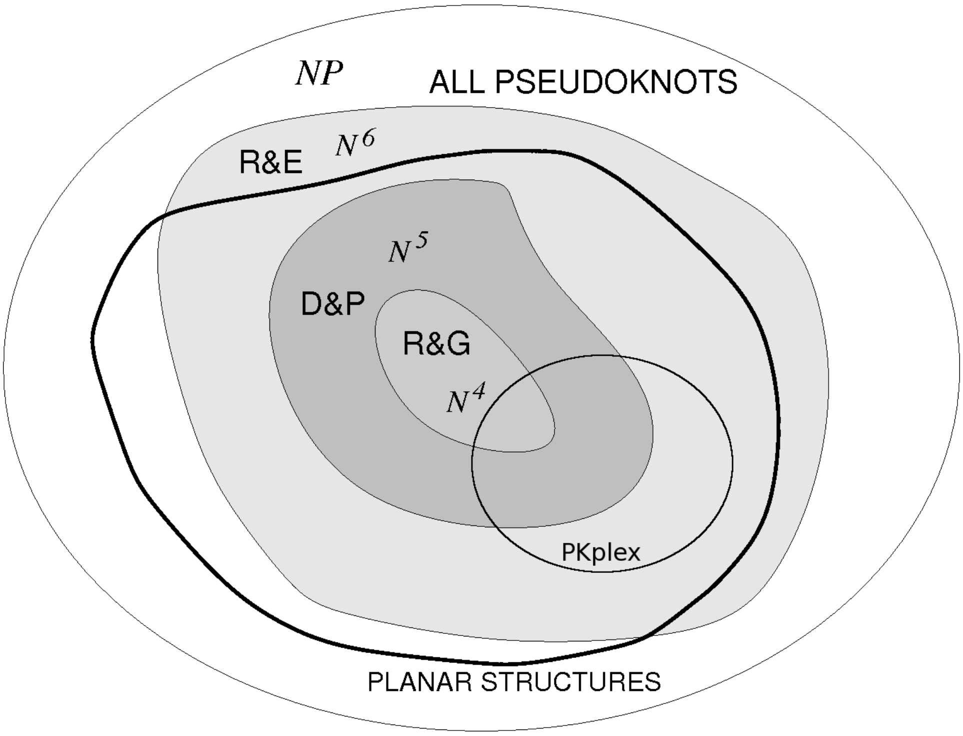

Figure 7.5: The runtime requirements and classes of pseudoknots covered by R&E [Rivas and Eddy, 1999], D&P [Dirks and Pierce, 2003], R&G's *pknotsRG* [Reeder and Giegerich, 2004] and *PKplex*. (Image adapted from [Hofacker and Stadler, 2007])

pknotsRG has other restrictions regarding the pseudoknots it is able to predict: on the one hand, it cannot handle kissing hairpins or bulge or internal loops within pseudoknot helices, but on the other hand it is able to generate structures with more than one pseudoknot and even recursive pseudoknots.

Other dynamic programming algorithms such as the one suggested in [Rivas and Eddy, 1999] can handle more general pseudoknots, but at the cost of a severely higher runtime making these algorithms quite impractical to use. A comparison of the covered pseudoknot classes and runtime requirements of different dynamic programming algorithms for pseudoknot prediction is shown in Figure 7.5. In contrast to dynamic programming algorithms, heuristic algorithms for pseudoknot prediction are not by their very nature forced to reduce the search space, i.e. restrict the types of pseudoknots they can handle. But this advantage comes at the cost of not being guaranteed to find the optimal solution with minimum energy within their search space.

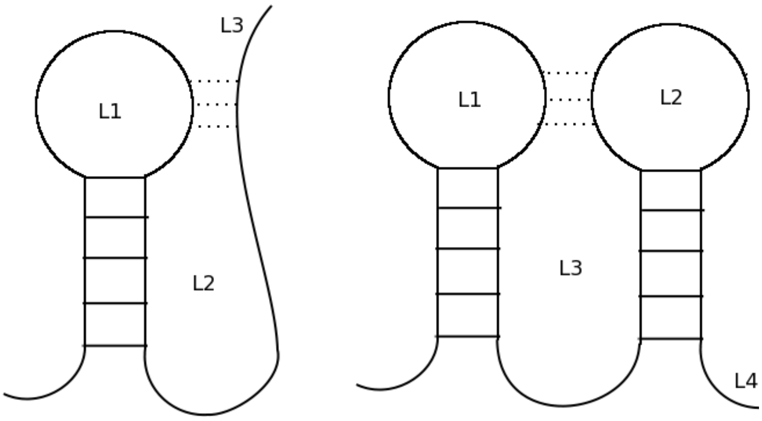

Figure 7.6: The formation of a pseudoknot (dotted lines) changes three loops in the left sequence (hairpin loop L1, L2 created by the pseudoknot and the external loop L3) and four loops in the right sequence (hairpin loops L1 and L2, L3 created by the pseudoknot and the external loop L4).

Whenever two unpaired sequence intervals form a pseudoknot, at least three secondary structure elements are directly involved (see Figure 7.6): the two loops enclosing the pseudoknot stems and the new loop created by the pseudoknot formation. The pseudoknot energy model of *PKplex* relies on the additivity of secondary structure elements and therefore assumes that the energies of those loops are not changed by the formation of the pseudoknot. As a compensation the pseudoknot initialization cost c_{pk} is introduced, which is a constant and does not depend on the actual composition of the loops changed by the pseudoknot.

In an effort to increase the prediction quality of *PKplex*, I analyzed the loops containing the bases involved in the pseudoknot formation. I found that *PKplex* significantly overestimates the amount of pseudoknot stems forming within multiloops. This lead to the creation of an energy model in which the pseudoknot initiation cost c_{pk} depends on the loop type enclosing the pseudoknot stems, with an increased penalty for multiloops. But while this reduced the amount of pseudoknot stems within multiloops in the predictions of *PKplex*, the overall prediction accuracy did not increase and I returned to the original simpler energy model.

The pseudoknot energy models of *pknotsRG*, *HotKnots* and *DotKnot* are described in section 5.4. While not as simple as the single constant used in *PKplex*, these models are not very complex either and mostly resemble the linear multiloop valuation in the standard RNA models: in addition to the pseudoknot initiation constant, they penalize longer loops by adding an energy penalty for every unpaired base and/or for every base pair within the pseudoknot loop. These models are not necessarily better than the simple constant used in *PKplex*, and in fact all these energy models can be considered makeshift and in need of improvement, because they do not properly account for the pseudoknot loop entropies. The problem is not that a more complex pseudoknot loop energy model would increase the computational costs too much, the heuristic algorithms at least could easily handle such a model. The problem rather is that the pseudoknot loop entropies are largely unknown and no proper model for them exists, apart from the one by Cao and Chen described in section 4.3 [Cao and Chen, 2009], but unfortunately their model is only applicable to a subset of the pseudoknots predicted by each of the four algorithms.

An advantage of the *PKplex* algorithm, which is caused by the design of the dynamic programming routine to calculate the energy of an interaction between two intervals, is the possibility to calculate suboptimal pseudoknotted secondary structures at no additional cost. These structures can then be further analyzed by either manual inspection or some automated filtering to potentially further improve the quality of the results.

The pseudoknot energy model by Cao and Chen was used as such a filter for improving the quality of the prediction results. The suboptimal structures predicted by *PKplex* were evaluated with the Cao and Chen energy model and the prediction with the best energy according to this evaluation was picked. But because the Cao and Chen energy model can only handle H-type pseudoknots, a lot of the candidate suboptimal structures could not be evaluated with it. This resulted in complicated rules for picking the best structure out of the suboptimal candidates depending on whether the Cao and Chen model could handle all, some, one or none of the candidate structures. Overall, this strategy did not improve the quality of the results significantly, most likely due to the problems with the limited applicability of the energy model. Other modifications that failed to improve the prediction quality include disallowing bulge and interior loops for the pseudoknot stems or restricting the loop lengths of the loops involved in the pseudoknot formation.

Chapter 8

Conclusion and Outlook

RNAs are a very important class of molecules for the living world. The unofficial creation of the research area now known as molecular biology was the discovery of the structure of DNA, the carrier of genetic information in 1953. For the first couple of decades following this discovery, RNA was only seen as occupying the less interesting role of passively carrying information from DNA to proteins. Almost all research was focused on the latter two molecule classes and RNAs were neglected. Only after the discovery that there existed RNAs which could exhibit catalytic activity themselves, RNA gained the attention of the scientific community that it deserves.

More and more functional RNAs are still being discovered today. The function of an RNA molecule strongly depends on its structure, and while a great number of RNA sequences are known and easily available in various sequence data bases, determining RNA structures requires a lot of experimental work. Therefore, predicting the structure of an RNA from its sequence has the potential to accelerate the scientific progress and save a lot of experimental work as well as money. Theoretically predicting the structure of RNA sequences cannot replace classic wet lab work, but it can filter the huge amount of potential functional RNAs to the most promising candidates, which can then be analyzed experimentally.

Most RNA secondary structure prediction algorithms are based on a dynamic programming approach, breaking apart the structure prediction problem into a number of smaller and simpler subproblems which can eventually be solved easily, and then constructing the solution to the bigger problem out of the solutions to the smaller problems. This approach depends on the loop decomposition of RNA structures requiring RNA structures to be perfectly nested. No bases enclosed by a base pair are allowed to bind with bases outside of the enclosing base pair. This restriction is only true for pseudoknot-free secondary structures, but more and more experimentally determined RNA structures contain pseudoknots, which violate this restriction.

In the effort to include pseudoknots into structure prediction algorithms it has been found that predicting pseudoknotted RNA structures without any restrictions is an NP-complete problem. Only after simplifying the energy model or by restricting the allowed classes of pseudoknots are current dynamic programming algorithms able to predict RNA pseudoknots. Some heuristic algorithms have also been employed for the prediction of pseudoknots, but in contrast to dynamic programming algorithms they are not guaranteed to find the optimum solution.

In this diploma thesis I introduced *PKplex*, a dynamic programming based approach within the MFE framework to predict RNA secondary structures with pseudoknots. The simple idea is to render two sequence intervals unpaired and accessible and let them interact with each other while the rest of the sequence is folded by a conventional pseudoknot-free folding algorithm. *PKplex* achieves a time complexity of $\mathcal{O}(n^3 + n^2w^4)$ and a space complexity of $\mathcal{O}(nw^2)$. Testing this approach on a large set of pseudoknotted and pseudoknot-free sequences shows that *PKplex* generates results that are qualitatively as good as those of other pseudoknot prediction algorithms while being computationally less expensive. The *PKplex* algorithm has been implemented in *C* as an extension to the *Vienna RNA package*.

Especially for longer pseudoknotted sequences the prediction quality of all available algorithms is not satisfying yet. The consensus is that this is less due to the computational effort required to handle longer sequences, but due to the fact that there is no general RNA energy model available yet which takes pseudoknots into account. Pseudoknots contain loops that are no hairpin-, interior-, bulge- or multiloop and that do not allow a simple loop decomposition like in the case of pseudoknot-free structures. Cao and Chen have created an energy model for H-type pseudoknots with energies for the pseudoknotted loops occurring in these structures but the energies of more general pseudoknotted loops are still unknown. An energy model which is able to handle a wider range of pseudoknots would potentially allow the creation of new algorithms for the prediction of RNA pseudoknots with greater accuracy.

Availability

The source code of *PKplex* and the *Vienna RNA package* are available on the web server of the Institute for Theoretical Chemistry of the University of Vienna at `http://www.tbi.univie.ac.at/~wolfgang/PKplex/` and at `http://www.tbi.univie.ac.at/~ivo/RNA/`.

Bibliography

P.L. Adams, M.R. Stahley, M.L. Gill, A.B. Kosek, J. Wang, and S.A. Strobel. Crystal structure of a group I intron splicing intermediate. *RNA*, 10 (12):1867, 2004.

T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.

E.S. Andersen, M.A. Rosenblad, N. Larsen, J.C. Westergaard, J. Burks, I.K. Wower, J. Wower, J. Gorodkin, T. Samuelsson, and C. Zwieb. The rmRDB and SRPDB resources. *Nucleic Acids Research*, 34:163–168, 2006.

M. Andronescu, V. Bereg, H.H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, 9(1):340, 2008.

M.S. Andronescu, C. Pop, and A.E. Condon. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, 16(1): 26, 2010.

J.P. Bachellerie, J. Cavaillé, and A. Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84(8):775–790, 2002.

I. Barette, G. Poisson, P. Gendron, and F. Major. Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic Acids Research*, 29(3):753–758, 2001.

S. Bernhart. Personal communication, 2009.

S.H. Bernhart, I.L. Hofacker, and P.F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614, 2006.

E. Birney, J. Stamatoyannopoulos, A. Dutta, and R. Guigo. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, 2007.

A.F. Bompfünewerer, R. Backofen, S.H. Bernhart, J. Hertel, I.L. Hofacker, P.F. Stadler, and S. Will. Variations on RNA folding and alignment: lessons from Benasque. *Journal of Mathematical Biology*, 56(1):129–144, 2008.

I. Brierley, S. Pennell, and R.J.C. Gilbert. Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nature Reviews Microbiology*, 5(8):598–610, 2007.

I. Brierley, R. Gilbert, and S. Pennell. RNA pseudoknots and the regulation of protein synthesis. *Biochemical Society Transactions*, 36:684–689, 2008.

P. Brion and E. Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

J.W. Brown. The ribonuclease P database. *Nucleic Acids Research*, 24(1): 236, 1996.

Y. Byun and K. Han. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. *Bioinformatics*, 25(11):1435, 2009.

L. Cai, R.L. Malmberg, and Y. Wu. Stochastic modeling of RNA pseudoknotted structures: a grammatical approach. *Bioinformatics*, 19(Suppl 1):i66–i73, 2003.

J.J. Cannone, S. Subramanian, M.N. Schnare, J.R. Collett, L.M. D'Souza, Y. Du, B. Feng, N. Lin, L.V. Madabusi, K.M. Müller, et al. The Comparative RNA Web(CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1):2, 2002.

S. Cao and S.J. Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research*, 34(9):2634, 2006.

S. Cao and S.J. Chen. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA*, 15(4):696, 2009.

T. Cech. Conserved sequences and structures of group I introns: building an active site for RNA catalysis - A review. *Gene*, 73:259–271, 1988.

T.R. Cech, A.J. Zaug, and P.J. Grabowski. In vitro splicing of the ribosomal RNA precursor of Tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27(3):487–496, 1981.

J.L. Chen, M.A. Blasco, and C.W. Greider. Secondary structure of vertebrate telomerase RNA. *Cell*, 100(5):503–514, 2000.

F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

F.H.C. Crick. The biological replication of macromolecules. In *Symp. Soc. Exp. Biol*, 1958.

R.M. Dirks and N.A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.

S.M. Elbashir, J. Harborth, W. Lendeckel, A. Yalcin, K. Weber, and T. Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, 2001.

A. Fire, S.Q. Xu, M.K. Montgomery, S.A. Kostas, S.E. Driver, and C.C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998.

C. Flamm. *Kinetic Folding of RNA*. PhD thesis, Universität Wien, 1998.

D.P. Giedroc and P.V. Cornish. Frameshifting RNA pseudoknots: structure and mechanism. *Virus research*, 139(2):193–208, 2009.

W. Gilbert. Origin of life: the RNA world. *Nature*, 319(6055), 1986.

S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S.R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33(Database Issue):D121, 2005.

A. Gruber. Strategies for measuring evolutionary conservation of RNA secondary structures. Master's thesis, Universität Wien, 2007.

A.R. Gruber, R. Lorenz, S.H. Bernhart, R. Neubock, and I.L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Research*, 36(Web Server issue): W70, 2008.

C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

A.P. Gultyaev, F.H.D. van Batenburg, and C.W.A. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of Molecular Biology*, 250(1):37–51, 1995.

A.P. Gultyaev, F.H. van Batenburg, and C.W. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5(5):609, 1999.

A.J. Hamilton and D.C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950, 1999.

C. Haslinger and P.F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*, 61(3):437–467, 1999.

I.L. Hofacker and P.F. Stadler. RNA secondary structures. In T. Lengauer, editor, *Bioinformatics: From Genomes to Therapies*, pages 438–489. Wiley-VCH, 2007.

I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

J.M. Johnson, S. Edwards, D. Shoemaker, and E.E. Schadt. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *TRENDS in Genetics*, 21(2):93–102, 2005.

G.F. Joyce. RNA evolution and the origins of life. *Nature*, 338(6212):217–224, 1989.

G.F. Joyce. The rise and fall of the RNA world. *The New Biologist*, 3(4): 399, 1991.

K.C. Keiler, P.R.H. Waller, and R.T. Sauer. Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. *Science*, 271(5251):990, 1996.

R.C. Lee, R.L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.

R. Lorenz. Secondary structure prediction for circular RNAs. Master's thesis, Universität Leipzig, 2007.

R.B. Lyngsø and C.N.S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

R.B. Lyngsø, M. Zuker, and C.N. Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6):440, 1999.

H. Mann, Y. Ben-Asouli, A. Schein, S. Moussa, and N. Jarrous. Eukaryotic RNase P:: Role of RNA and Protein Subunits of a Primordial Catalytic Ribonucleoprotein in RNA-Based Catalysis. *Molecular cell*, 12(4):925–935, 2003.

R.A. Marciniak, F.B. Johnson, and L. Guarente. Dyskeratosis congenita, telomeres and human ageing. *Trends in Genetics*, 16(5):193–195, 2000.

D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

J.S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, 2003.

J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

U. Mückstein, H. Tafer, J. Hackermuller, S.H. Bernhart, P.F. Stadler, and I.L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177, 2006.

U. Mückstein, H. Tafer, S.H. Bernhart, M. Hernandez-Rosales, J. Vogel, P.F. Stadler, and I.L. Hofacker. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. *Bioinformatics Research and Development*, pages 114–127, 2008.

N. Nameki, T. Tadaki, H. Himeno, and A. Muto. Three of four pseudoknots in tmRNA are interchangeable and are substitutable with single-stranded RNAs. *FEBS letters*, 470(3):345–349, 2000.

R. Nussinov and A.B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309, 1980.

J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5(1):104, 2004.

J. Ren, B. Rastegari, A. Condon, and H.H. Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11 (10):1494, 2005.

E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285 (5):2053–2068, 1999.

J. Ruan, G.D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20(1):58, 2004.

P.J. Schlax, K.A. Xavier, T.C. Gluick, and D.E. Draper. Translational Repression of the Escherichia coli α Operon mRNA. *Journal of Biological Chemistry*, 276(42):38494, 2001.

J. Sperschneider and A. Datta. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Research*, 38(7):e103, 2010.

D.W. Staple and S.E. Butcher. Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6), 2005.

L.D. Stein. Human genome: end of the beginning. *Nature*, 431(7011):915–916, 2004.

J.E. Tabaska, R.B. Cary, H.N. Gabow, and G.D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691, 1998.

H. Tafer and I.L. Hofacker. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657, 2008.

C.A. Theimer and D.P. Giedroc. Contribution of the intercalated adenosine at the helical junction to the stability of the gag-pro frameshifting pseudoknot from mouse mammary tumor virus. *RNA*, 6(3):409, 2000.

I. Tinoco et al. How RNA folds. *Journal of Molecular Biology*, 293(2): 271–281, 1999.

C. Tuerk, S. MacDougal, and L. Gold. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proceedings of the National Academy of Sciences of the United States of America*, 89(15): 6988, 1992.

D.H. Turner and D.H. Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 2009.

T. Tuschl, C. Gohlke, T.M. Jovin, E. Westhof, and F. Eckstein. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266(5186):785, 1994.

S. Valadkhan. snRNAs as the catalysts of pre-mRNA splicing. *Current Opinion in Chemical Biology*, 9(6):603–608, 2005.

F.H.D. van Batenburg, A.P. Gultyaev, and C.W.A. Pleij. PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Research*, 29(1):194, 2001.

T. Vulliamy, A. Marrone, I. Dokal, and P.J. Mason. Association between aplastic anaemia and mutations in telomerase RNA. *The Lancet*, 359 (9324):2168–2170, 2002.

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

M.S. Waterman and T.F. Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42(3-4):257–266, 1978.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, Apr 1953.

K.C. Wiese, E. Glen, and A. Vasudevan. JViz.Rna - A Java tool for RNA secondary structure visualization. *IEEE Transactions on Nanobioscience*, 4(3), 2005.

C. Witwer, I.L. Hofacker, and P.F. Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(2):66–77, 2004.

A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15310, 2003.

M. Zuker. Calculating nucleic acid secondary structure. *Current Opinion in Structural Biology*, 10(3):303–310, 2000.

M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, 1984.

M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133, 1981.

Mag. Wolfgang Beyer

Personal Information	Address:	Hasnerstraße 105/26, A-1160 Vienna, Austria
	Phone:	+43 (0) 650-5858999
	Email:	wolfi.b@gmx.at
	Date of Birth:	Feb 10th, 1980 in Vienna, Austria
	Nationality:	Austrian
Education	Since 2003	**University of Vienna** **Molecular Biology** specializing in Bioinformatics, Structural Biology and Genetics
	Spring 2007	**Universidade Nova de Lisboa (Portugal), Exchange Semester** Research internship topic: Modeling Metabolic Pathways
	1999 - 2004	**Vienna University of Economics and Business Administration** **Degree in Business Administration, graduation with distinction** Diploma thesis: “The Quadrature Method in Comparison with Other Methods for Valuing Barrier Options” Graduation average: 1.5 Majors: Finance, Controlling
	Spring 2002	**University of Technology Sydney (Australia)** **Exchange Semester, Bachelor of Business course** Coursework includes: International Financial Management, Derivative Securities
	1998 – 1999	**Military Service**
	1990 - 1998	**Secondary School, Gymnasium und Realgymnasium Wien 8** Graduation average: 1.0
Experience	Oct 2004 - Feb 2005	**Altea Trading Company LLC** Proprietary Futures Trader
	Oct 2002 – Dec 2002	**Merrill Lynch (Frankfurt)** Internship at the European Equity Sales department
Language & Computer Skills		First language German, fluent English, basic French and Portuguese C, C++, HTML, Latex, Visual Basic, Bloomberg
Awards & Activities		Academic Scholarship for 2001/2002 Member of the Austrian team at the International Mathematical Olympiad 1998 in Taipeh, Taiwan Youth Work (Scout leader since 1998)

Vienna, October 22nd, 2010