On the Expected Free Energy of RNA Molecules

Markus E. Nebel, Anika Scheid*
Fachbereich Informatik, Technische Universität Kaiserslautern
Gottlieb-Daimler-Straße 48, D-67663 Kaiserslautern, Germany
{nebel,a_scheid}@informatik.uni-kl.de

Abstract

This article focuses on the analytical analysis of the free energy in a realistic model for RNA secondary structures. In fact, the free energy in a stochastic model derived from a database of small and large subunit ribosomal RNA (SSU and LSU rRNA) data is studied. A common thermodynamic model for computing the free energy of a given RNA secondary structure, as well as stochastic context-free grammars and generating functions are used to derive the desired results. These results include asymptotics for the expected free energy and for the corresponding variance of a random RNA secondary structure. The quality of our model is judged by comparing the derived results to the used database of SSU and LSU rRNA data. At the end of this article, it is discussed how our results could be used to help on identifying good predictions of RNA secondary structure.

1 RNA Secondary Structure

Ribonucleic acid (RNA) is a single-stranded *nucleotide polymer* (also called *oligonucleotide* or *polynucleotide*). The basis structural units (monomers) of RNA are formed by *nucleotides*. In RNA, each nucleotide is a molecule consisting of a phosphate group, a sugar group (ribose) and one of the four bases adenine (A), cytosine (C), guanine (G) and uracil (U). Ribose is a 5-carbon sugar (pentose) and in RNA, it has a cyclic form, which is called the *ribose ring*. The five carbon atoms of the ribose ring are numbered in clockwise order, and the ith carbon is called the i' carbon of the ribose ring. In each nucleotide, the base is bound to the $1'$ carbon of the ribose ring.
An RNA single-strand is formed by linking together the nucleotide units. More precisely, the linear structure of the RNA molecule, a chain consisting of four different types of nucleotides, is formed by creating *phosphodiester bonds*. In such bonds, the phosphate group (at the $5'$ carbon of the ribose ring) of the first nucleotide is attached to the hydroxyl group (at the $3'$ carbon of the ribose ring) of the next nucleotide. In nature, RNA strands are always extended at the $3'$ end, which implies that they grow in the $5' \to 3'$ direction. Details can be found, for example, in [CB00].
The specific sequence of bases along the RNA chain is called the *primary structure* of the molecule. The primary structure of an RNA molecule is essentially one-dimensional and is usually modeled as a string over the alphabet $\Sigma = \{A, C, G, U\}$, i.e. it is represented as a sequence of letters $r_1 r_2 \dots r_n$, where r_i is either A, C, G or U. By convention, strings representing the primary structure of RNA molecules are written in the $5' \to 3'$ direction, which means that they are written with the $5'$ end at the left to the $3'$ end at the right.
In vivo, single-stranded RNA chains bend and twine about themselves. The reason for this behaviour is that, in addition to the phosphodiester bonds between neighbored bases in the RNA chain, two bases that are not neighbored may form other, weak chemical bonds, called *hydrogen bonds*. More precisely, the complementary bases adenine (A) and uracil (U) resp. cytosine (C) and guanine (G) form stable base pairs with each other by creating hydrogen bonds. These base pairs are called *Watson-Crick pairs*. In addition to these stable Watson-Crick base pairs, there may occur weaker base pairs, called GU *wobble pairs*, which are formed by the non-complementary bases guanine (G) and uracil (U). All these pairs (Watson-Crick and GU wobble pairs) are called *canonical* base pairs, as they are most common. Other pairs, called *non-canonical* base pairs, may also occur, but they are not as stable as the canonical ones.
Since base pairs may be formed arbitrarily, the linear RNA chain is folded into a three-dimensional conformation, called the *tertiary structure* of the RNA molecule, which determines the biochemical activity of the molecule. It is costumary in science to simplify the study of the tertiary structure of an RNA molecule by allowing only non-crossing base pairs such that the corresponding molecule remains planar. Accordingly, this restriction yields a two-dimensional conformation, called the *secondary structure* of the

molecule. By investigating secondary structures of RNA instead of the corresponding tertiary structures, the focus of attention is hence set only on what base pairs are involved, and not on the three-dimensional conformation of the RNA chain.
Finally, it should be mentioned that there are different kinds of RNA playing different roles. In fact, SSU and LSU rRNA, which will be considered in the sequel to derive a realistic model for RNA secondary structure, are only one such type. For more detailed information on the molecular structure and the functions of the different types of RNA, see for example [AJL+02].

2 Definitions and Prior Results

Following the convention that RNA sequences are written in the $5' \to 3'$ direction, we number the bases of an RNA sequence from 1 (called the $5'$ terminus) to n (the $3'$ terminus). This leads to the following definition of a secondary structure of size n:

Definition 2.1 ([ZMT99]) *A* secondary structure $\mathbf{S}$ *of size n is a finite set (possibly empty) of* base pairs. *A base pair between i and j ($1 \leq i < j \leq n$) is denoted by $i.j$. A few constraints are imposed:*

1. *Two base pairs, $i.j$ and $i'.j' \in \mathbf{S}$ are either identical, or else $i \neq i'$ and $j \neq j'$. Thus base triplets are deliberately excluded from the definition of secondary structure.*

2. *Pseudoknots are prohibited. That is, if $i.j$ and $i'.j' \in \mathbf{S}$, then, assuming $i < i'$, either $i < i' < j' < j$ ($i.j$ includes $i'.j'$) or $i < j < i' < j'$ ($i.j$ precedes $i'.j'$) .*

3. *Sharp U-turns are prohibited. A U-turn, called hairpin loop, must contain at least 3 bases. That is, if $i.j \in \mathbf{S}$, then $|j - i| \geq 4$.*

According to constraint 1 of Definition 2.1, each i occurs either in exactly one pair or in no pairs, and i is described as *paired* or *unpaired*, accordingly. Pseudoknots [PB89, AvdBvBP90, GW90, DPD92, Ple94], formed by two base pairs $i.j$ and $i'.j'$ satisfying $i < i' < j < j'$, are often considered as belonging to the tertiary structure and are not permitted in secondary structures according to Definition 2.1, due to constraint 2. Additionally, by constraint 3 of Definition 2.1, the stereochemical constraint that i and j cannot base pair if $|j - i| < 4$ is included into the definition of a secondary structure of size n. This means that a U-turn must contain at least three bases, since in nature, U-turns containing less than three unpaired bases are impossible and do not form. Constraints 1 to 3 of Definition 2.1 limit the number of possible foldings of a given RNA molecule in a very significant way. However, Definition 2.1 still allows an exponential number of biologically impossible structures, since any two bases are allowed to pair. But as we already know, hydrogen bonding can occur only between the bases A and U or between G and C, with a weaker bond possible between the two bases G and U. To distinguish between paired and unpaired bases resp. double-stranded and single-stranded regions in RNA secondary structures, we will use the following definition:

Definition 2.2 ([ZMT99]) *A group of two or more consecutive*[1] *base pairs is called a* helix. *The first and last are the closing base pairs of the helix. They may be written as $i.j$ and $i'.j'$, where $i < i' < j' < j$. Then $i.j$ is called the* external closing base pair *and $i'.j'$ is called the* internal closing base pair.

Hence, any secondary structure $\mathbf{S}$ can be decomposed into single-stranded regions and helices. But for our further investigations, we additionally need to distinguish between different kinds of single-stranded regions. Therefore, we first have to consider the following definition:

Definition 2.3 ([Zuk86]) *Any subset of a secondary structure $\mathbf{S}$ is also a secondary structure, and is called a* substructure. *The substructure $\mathbf{S}_{ij}$ for $1 \leq i < j \leq N$ is defined as*

$$\mathbf{S}_{ij} = \{i'.j' \in \mathbf{S} : i \leq i' < j' \leq j\}.$$

We can decompose any given secondary structure $\mathbf{S}$ in a unique way into a number of substructures such that each position is contained in exactly one such substructure:

[1] A group of $k \geq 1$ consecutive base pairs means k base pairs $(i+1).(j-1), \ldots, (i+k).(j-k)$ such that neither the two bases $(i+k+1)$ and $(j-k-1)$ nor the two bases i and j (if existing) form together a base pair.

Definition 2.4 (k-loop decomposition [ZS84, Zuk86]) *If $i.j$ is a base pair in the secondary structure $\mathbf{S}$ and if $i < k < j$, we say that k is* accessible *from $i.j$ if there is no $i'.j'$ in $\mathbf{S}$ such that $i < i' < k < j' < j$. Similarly, if $k.l$ is also in $\mathbf{S}$, we say that the base pair $k.l$ is* accessible *if both k and l are accessible. The set of $(k-1)$ base pairs and k' unpaired bases accessible from $i.j$ is called the* k-loop *(or* k-cycle*) closed by $i.j$. The (possibly empty) set of base pairs in a k-loop constitute the* interior *base pairs of the k-loop. The closing base pair is called the* exterior *base pair. k' is called the* size *of the k-loop. The collection of $(k-1)$ base pairs and k' unpaired bases which are accessible from no base pair (the* exterior *or* free *base pairs and bases) is called the* null k-loop *or* exterior loop*. It is easy to see that any secondary structure $\mathbf{S}$ decomposes the sequence $1, 2, \ldots, n$ uniquely into k-loops $s_0, s_1, s_2, \ldots, s_m$, where s_0 is the null k-loop and $m > 0$ iff $\mathbf{S}$ is nonempty* [2].
Biochemists have developed their own nomenclature for k-loops. The various cases and subcases are given as follows:

1. *$k = 1$: A 1-loop is called a* hairpin loop*.*
2. *$k = 2$: Let $i'.j'$ be the base pair accessible from $i.j$. Then the 2-loop is called*
   (a) *a* stacked pair*, if $i' - i = 1$ and $j - j' = 1$,*
   (b) *a* bulge (loop) *if $i' - i > 1$ or $j - j' > 1$, but not both, and*
   (c) *an* interior loop[3] *if $i' - i > 1$ and $j - j' > 1$.*
3. *$k \geq 3$: These k-loops are called* multi-branched loops*,* multiple loops *or simply* multiloops*.*

In the style of [ZMT99], the loop closed by a base pair $i.j$ will be denoted by $\mathbf{L}(i.j)$, the exterior loop will be denoted by $\mathbf{L}_e$ and the number of single-stranded bases in a loop will be denoted by the term $l_s(\mathbf{L})$ in the sequel. Hence, the size of a 1- or 2-loop is defined as $l_s(\mathbf{L})$. In fact, if $\mathbf{L}(i.j)$ is an interior loop with interior base pair $i'.j'$ which is accessible from the exterior base pair $i.j$ of the loop, then its size $l_s(\mathbf{L})$ can be written as $l_s(\mathbf{L}) = l_s^1(\mathbf{L}) + l_s^2(\mathbf{L})$, where $l_s^1(\mathbf{L}) = i' - i - 1$ and $l_s^2(\mathbf{L}) = j - j' - 1$. Due to this fact, there are some special types of interior loops, depending on the combination of the two sizes $l_s^1(\mathbf{L})$ and $l_s^2(\mathbf{L})$:

Definition 2.5 ([ZMT99]) *Let $\mathbf{L}(i.j)$ be an interior loop of size $l_s(\mathbf{L}) = l_s^1(\mathbf{L}) + l_s^2(\mathbf{L})$.*

- *If $l_s^1(\mathbf{L}) = l_s^2(\mathbf{L})$, the loop is called* symmetric*; otherwise, it is* asymmetric*, or* lopsided*.*
- *The asymmetry of the interior loop $\mathbf{L}$, $a(\mathbf{L})$ is defined by:*

$$a(\mathbf{L}) = |l_s^1(\mathbf{L}) - l_s^2(\mathbf{L})|.$$

- *If $l_s^1(\mathbf{L}) = 1$ and $l_s^2(\mathbf{L}) = n$ or $l_s^1(\mathbf{L}) = n$ and $l_s^2(\mathbf{L}) = 1$, $n > 2$, then the interior loop $\mathbf{L}$ is called a* "Grossly Asymmetric Interior Loop" (GAIL)*.*

Finally, note that by including pseudoknots into the definition of an RNA secondary structure $\mathbf{S}$, the k-loop decomposition breaks down.
Since RNA secondary structures are two-dimensional, they can be modeled as planar graphs. Examples are given in Figure 1. Such representations of RNA secondary structures as planar graphs are used universally, as they pictorally represent the structure and the different substructures resp. loops can immediately be determined. Alternatively, RNA secondary structures can be modeled as strings over the alphabet $\Sigma := \{(,), |\}$, where a bar represents an unpaired nucleotide and a pair of corresponding brackets **( )** represents two bases in the RNA molecule that are paired (see, e.g. [VC85]). There is obviously a one-to-one correspondence between planar graph representations and bar-bracket representations of RNA secondary structures, as illustrated by Example 2.1.

Example 2.1 *The bar-bracket representation of the RNA secondary structure which is shown as planar graph in Figure 1 on the left is given by*

||((|||(((|||((((||||)|)))))|||))|(((|(((|||||(((|||||(((||((|||))|||)))))))||((||(((|||)))||)))))||)))|||((||||||||||(((|(||||))))||)).

[2] Note that this decomposition was first introduced in [SKMC83] and was later redefined. In the original definition, the closing pair belongs to the k-loop, but in the redefinition given here, the closing base pair is no longer contained in the k-loop.

[3] In the sequel, such an interior loop will sometimes be called $(i' - i - 1) \times (j - j' - 1)$ interior loop to specify the number of unpaired bases between the paired bases i and i', as well as j and j', respectively.

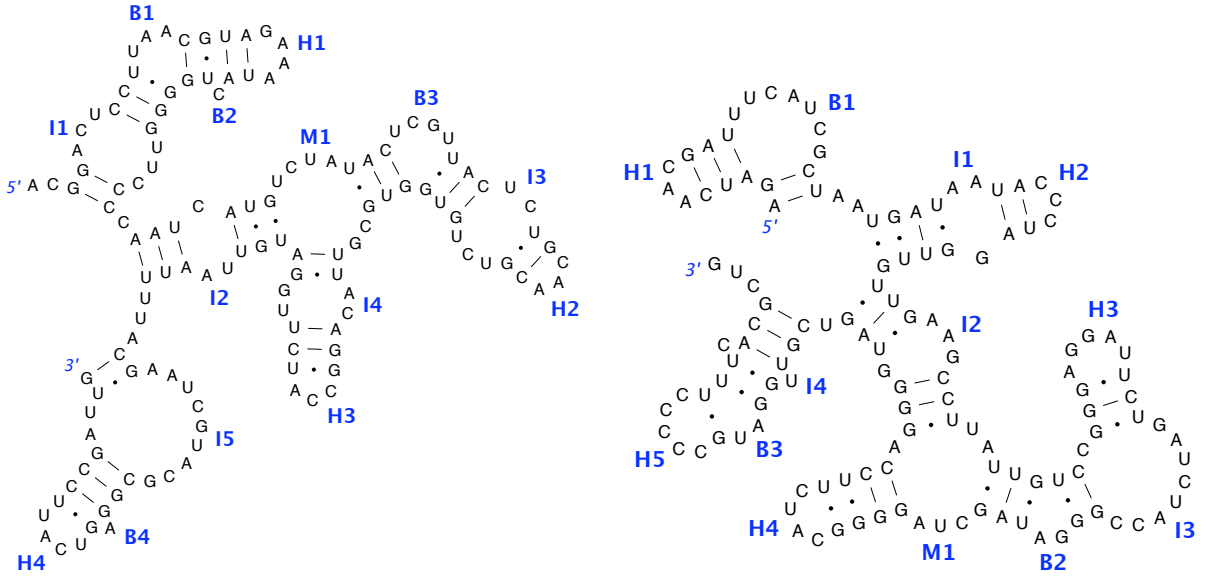

Figure 1: Planar graph representations of RNA secondary structures.

Equally, the secondary structure whose planar graph representation is shown in Figure 1 on the right can be represented as follows:

((((((|||))))|||||||))|||((((||((|||))|))))((((|||((((||((((((|(((((|||||))))|||||||||))||)))))|||((((|||||))))))|))))|))))|(((|(((||(|||||))))))|)))))|||.

It should be clear that these bar-bracket representations abstract from the RNA sequence, as they only consider the number of base pairs and their positions. This means that given a bar-bracket representation of a secondary structure **S**, we do not know the corresponding RNA sequence **R**, i.e. we do not know if a pair of corresponding brackets **( )** represents a non-canonical or a canonical (Watson-Crick or wobble GU) base pair.

3 Computational Prediction of RNA Secondary Structures

Given an RNA sequence, we are usually interested in its secondary structure, as the function of non-coding RNA is often determined by its three-dimensional structure and much of the final structure is determined by the intramolecular base-pairing interactions of the molecule. But the experimental determination of RNA secondary structures is usually time-consuming and expensive and therefore, much effort has been made to create approaches for the computational prediction of RNA secondary structures over the last decades. The problem is that there are lots of different possibilities for a single-stranded RNA molecule to fold into a two-dimensional conformation and we have to compute those foldings that are most realistic or most stable. Hence, the usual method is to predict a secondary structure that is optimal in some sense. The most common approach for predicting the secondary structure of an RNA molecule is free energy minimization. In the context of RNA folding, free energy means the change of the *Gibbs free energy* in the chemical process of folding the RNA molecule. As in nature every RNA molecule seeks to achieve a minimum of free energy by folding into a higher-dimensional conformation, it is assumed that the correct structure is the one with the lowest free energy. Hence, many prediction methods use free energy as their metric and try to compute a conformation of minimum free energy.
The most successful and popular method for energy minimization over the last 30 years has been the use of dynamic programming algorithms. The pioneering work in this domain was published in 1978 by Nussinov et al. [NPGK78]. In this work, the authors introduced an efficient dynamic programming algorithm which used a simple free energy function **E** that is minimized when the secondary structure contains the maximum number of complementary base pairs. More precisely, in this model each pase pair $i.j$ in a given secondary structure **S** is assigned an energy $e(i.j)$, such that the overall energy of the secondary structure **S** is given by

$$\mathbf{E}(\mathbf{S}) = \sum_{i.j \in \mathbf{S}} e(i.j),$$

where the stability of GC pairs is considered to be equal to that of AU pairs. Additionally, contributions due to stacking of base pairs and destabilizing effects for loop formation were ignored. By utilizing a simple method for estimating the free energy of loops found in single-stranded RNA based on their sequence, which was derived earlier [TUL71], the folding rules of this dynamic programming algorithm for maximal matching were modified to allow an estimate of the free energy of loop structures based on sequence data [NJ80]. This means that hydrogen bond potential energies $e(i.j)$ are computed for each base pair $i.j$, such that the algorithm computes one structure with the lowest free energy $\mathbf{E}$. But as these energy rules are only base pair-dependent, stacking and destabilizing energies were not incorporated into this algorithm.
Therefore, in 1981, Zuker and Stiegler presented a new dynamic programming algorithm for folding an RNA molecule that finds a conformation of minimum free energy using thermodynamics and auxiliary information [ZS81]. This algorithm uses loop-dependent energy rules to compute the free energy of each loop, such that the overall energy of a secondary structure $\mathbf{S}$ is given by

$$\mathbf{E}(\mathbf{S}) = e(\mathbf{L}_e) + \sum_{i.j \in \mathbf{S}} e(\mathbf{L}(i.j)).$$

During the following years, this dynamic programming algorithm based on thermodynamic parameters has been improved several times [SKMC83, ZS84, Zuk89a]. However, due to imprecisions in the energy rules and the thermodynamic parameters, as well as the fact that certain chemical aspects, like for example the influence of enzymes or the effect of cotranscriptional folding, have not been incorporated into dynamic programming algorithms, the predicted optimal (minimum free energy) structure was often not the native one. Therefore, there was an urgent need to additionally predict suboptimal foldings.
For this reason, in 1989, an algorithm for determining RNA secondary structures within any prescribed increment of the computed global minimum free energy was introduced [Zuk89b]. This algorithm was implemented in the MFOLD software, which has become a widely used program to predict RNA secondary structures. The description and use of the MFOLD package has appeared in a number of articles [JTZ89, JTZ90, Zuk94, ZMT99]. MFOLD is also available as an online web server [Zuk03]. The portal for the RNA MFOLD web server is `http://www.bioinfo.rpi.edu/applications/mfold` linked to Michael Zuker's homepage.
Finally, it remains to mention that all these algorithms only work for secondary structures without pseudoknots, as they cannot predict crossing base pairs. But as pseudoknots are important to the function of severals kinds of RNAs, much effort has been made to develop algorithms for predicting RNA secondary structures that contain pseudoknots. The most general algorithm for predicting structures with pseudoknots, which is capable of predicting nearly all known classes of pseudoknots, was presented in 1999 by Rivas and Eddy [RE99]. But this algorithm has a theoretical worst-case complexity of $\mathcal{O}(N^6)$ in time and $\mathcal{O}(N^4)$ in storage and is thus only practical for very short RNA sequences. Recently, another algorithm (which has a better theoretical worst-case complexity) has been presented by Metzler and Nebel [MN08].
A recent review on how RNA folding algorithms work and why they can't deal with pseudoknots is given in [Edd04].

4 Thermodynamic Models for RNA Secondary Structures

In the early mid-1970s, biochemists hypothesized that each base pair in a helix contributes to the stability of that helix and that the contribution of a base pair depends on its adjacent base pairs [GC73, BDTU74]. This yielded a new model in which the thermodynamic stability of a given base pair is dependent on the identity of its *nearest neighbor*, the so-called *individual nearest-neighbor (INN) model*. In this model, it was assumed that a newly formed base pair is stacked on an existing pair and the free energy is assigned to their stacking interaction. This means that the formation of a helix includes the concentration-dependent formation of the first base pair, called *initiation*, which includes hydrogen bonding and brings strands together. This initiation is followed by a closing of subsequent base pairs, called *propagation* of the helix by base pairing, which includes stacking interactions as well as hydrogen bonding.
In 1998, an expanded nearest-neighbor model for formation of RNA helices with Watson-Crick base pairs was presented, which was termed the *individual nearest-neighbor hydrogen bond (INN-HB) model* [XSB$^+$98]. This model also includes a penalty term for terminal AU pairs, since it was noticed that helices with the same nearest neighbors but different terminal ends consistently have different stabilites. The name of this model is due to the fact that there are different numbers of hydrogen bonds in AU (two hydrogen bonds) and GC (three hydrogen bonds) base pairs, and by changing a GC base pair to an AU base pair,

the number of hydrogen bonds in the helix is decreased by one.
In 1999, nearest-neighbor free energy parameters for the stacking of wobble GU pairs in RNA helices were derived for the INN-HB model [MSZT99]. The authors suggested that terminal GU pairs are treated like terminal AU pairs in the INN-HB model, because they have the same number of hydrogen bonds. This means that the same penalty term is added for each terminal AU and each terminal GU pair in the INN-HB model. Thermodynamics for RNA secondary structures have also been studies for all other common substructures. These studies led to a number of different thermodynamic parameters for certain (special) types of loops along with corresponding loop-dependent free energy rules. These results are summarized in [ST95] (for the INN-model), as well as in [MSZT99] and in [ZMT99] (for the INN-HB model).
In this work, we will use the INN-HB model with loop-dependent energy rules [XSB$^+$98, MSZT99] to compute the free energy of a given RNA secondary structure $\mathbf{S}$. Note that only Watson-Crick and wobble GU pairs are allowed in this INN-HB model, as nearest neighbor rules break down for non-canonical (i.e. non-Watson-Crick and non-GU) base pairs. This means that non-canonical base pairs in helices must instead be treated as mismatched pairs.
The thermodynamic parameters that will be used in this work are the free energy data from Mathews et al. [MSZT99], which were used for version 3.0 of the MFOLD software [Zuk03]. The corresponding thermodynamic model for RNA secondary structures that will be used in this work is derived from [MSZT99] and [ZMT99]. It includes all but coaxial stacking (which is a favorable interaction of two helices stacked end to end, in multi- and exterior loops) from the latest free energy parameters. This thermodynamic model distinguishes between the following (special) types of loops:

- hairpin loops of size 3, called *triloops*,
- hairpin loops of size 4, called *tetraloops*,
- hairpin loops of size > 4,
- stacked pairs,
- bulge loops of size 1, called *single bulges*,
- bulge loops of size > 1,
- 1×1 interior loops, called *single mismatches*,
- 2×2 interior loops, called *tandem mismatches*,
- 1×2 (resp. 2×1) interior loops,
- non-grossly asymmetric interior loops of size > 4,
- grossly asymmetric interior loops (GAILs),
- multiloops and
- exterior loops.

In particular, for hairpin loops, the thermodynamic parameters and free energy rules include a length-dependent loop destabilizing free energy (which depends on the size, i.e. number of unpaired nucleotides in the loop) and a terminal mismatch stacking energy (for loops of size ≥ 4) resp. the terminal AU/GU penalty (for loops of size 3). Additionally, a GGG loop bonus (applies only to GU closed hairpins in which a $5'$ closing G is preceded by two G residues) and a penalty term for poly-C hairpin loops (i.e. for hairpin loops in which all unpaired nucleotides are C), as well as a tetraloop bonus (for hairpin loops of size 4) are included.
For bulge loops, a length-dependent loop destabilizing free energy, as well as the terminal AU/GU penalty for both the interior and exterior base pair (for loops of size > 1 only) are included in the model. For single bulges and for stacked pairs, a stacking energy for the stacking interaction of the interior and exterior base pair is added.
Small symmetric interior loops and almost symmetric interior loops, particularly 1×1, 2×2 and 1×2 interior loops are treated in a special way, since for these loops, individual sets of free energy values are consulted that contain values for every possible sequence variation. For all other interior loops, the thermodynamic parameters include a length-dependent loop destabilizing free energy and a free energy contribution that penalizes asymmetry in the loop. Additionally, a terminal mismatch stacking energy

(for loops of size > 4 that are no GAIL) resp. two free energy changes associated with the terminal base pairs of the two helices in which the loop ends (for GAILs) is added to the stability of the loop.
Finally, for multi- and exterior loops, the terminal AU/GU penalty and a free energy contribution for the stacking interaction of a base pair with (0, 1 or 2) single-stranded bases adjacent to that base pair are explicitly applied to all the terminal base pairs of the helices that are radiating out from this loop[4].
Additionally, for multiloops, a destabilizing initiation free energy is added, which depends on the number of single-stranded bases and on the number of base pairs accessible from the closing base pair of the loop.
Note that in this model, the terminal AU/GU penalty term for a terminal AU or GU base pair at the end of a helix is added to the free energy of a given secondary structure $\mathbf{S}$ along with the free energy of the loop $\mathbf{L}(i.j)$ closed by a base pair $i.j \in \mathbf{S}$ in which the helix terminates. This means that the terminal AU/GU penalty, if necessary, is formally assigned the loop $\mathbf{L}(i.j)$ closed by the pair $i.j \in \mathbf{S}$, although it really belongs to the helix in which the loop ends.
As the change of the Gibbs free energy G in the chemical process of folding the RNA molecule depends on the temperature and the thermodynamic parameters used here are all for 37°C, we will write $\Delta G^{\circ}_{37}(\mathbf{S})$ in the sequel to denote the free energy change of a secondary structure $\mathbf{S}$ at 37°C.
Finally, in this model, the free energy $\Delta G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S}$ is assumed to be given by the sum of the free energy changes of all its substructures, formally

$$\Delta G^{\circ}_{37}(\mathbf{S}) = \Delta G^{\circ}_{37}(\mathbf{L}_e) + \sum_{i.j \in \mathbf{S}} \Delta G^{\circ}_{37}(\mathbf{L}(i.j)).$$

5 Analysis of the Free Energy in a Stochastic Model for RNA Secondary Structures

Numerous results have been published that deal with the expected shape of secondary structure of RNA molecules. In fact, after the first formal definition of RNA secondary structures was given in [Wat78] (where the RNA molecule is modeled as a certain kind of planar graph), many authors considered the combinatorial model for RNA secondary structures in order to solve enumeration problems related to the combinatorics of these structures (see for example [SW78, VC85, Neb02a]). In the combinatorial model for RNA secondary structures, a uniform distribution of those structures is assumed, which means that all secondary structures are equiprobable. In fact, in the combinatorial model, it is assumed that base pairing is possible between arbitrary pairs of nucleotides, as only the topolgy of the planar secondary structure is considered. Thus, the combinatorial model completely abstracts from the primary structure of which these secondary structures could have been formed.
For this reason, some authors decided to consider a more realistic model for RNA secondary structures, the so-called *Bernoulli-model*, which is capable of incorporating information on the possible RNA sequences for a given secondary structure (see for example [HSS98, Neb04b, ZS84]). This model is obtained by a stochastic approach, where a Bernoulli distribution of the bases is assumed. More precisely, a parameter p is incorporated to specify the probability that two random bases can form a hydrogen bond. In the style of [Les74] and [Neb04b], this base-pairing probability p is often called *stickiness*.
However, in [Neb04b], it was pointed out that both the combinatorial model and the Bernoulli-model for RNA secondary structures are rather unrealistic. As a consequence, in [Neb02b, Neb04a], it was described how to construct a more realisitic model for RNA secondary structures. Like the Bernoulli-model, this model is obtained by a stochastic approach. In particular, a probability distribution of secondary structures is derived from a database of real world RNA secondary structure data. This way, the shape of RNA secondary structures is modeled most realistically. Therefore, we decided to analyse the free energy only in such a stochastic model for RNA secondary structures[5]. In particular, our aim is to determine the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ and the corresponding variance of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of a stochastic model derived from biological data.
To compute the desired results, we will use the methods of *generating functions*. To keep this article mostly self-contained, we will recall the fundamental definitions in Appendix A. For a more comprehensive introduction to generating functions and some of their uses in discrete mathematics, see for example [Wil94]. Several examples for generating functions can be found in [Com74]. Additionally, for

[4]Note that if $i.j$ and $j + 2.l$ are two base pairs, then r_{j+1} can interact with both of them. In this case, the stacking is assigned to only one of the two base pairs, whichever has a lower free energy (usually the $3'$ stack). In fact, the sum of all the free energy contributions for stacking of single-stranded bases to the terminal base pairs has to be minimized.

[5]Obviously, by analysing the free energy in the combinatorial or in the Bernoulli-model, we cannot expect that the corresponding results are realistic.

an introduction to some advanced methods that have to be used for more difficult problems, see for example [GK90].

5.1 Stochastic Context-Free Grammars (SCFGs)

In addition to generating functions, we will consider *stochastic context-free grammars (SCFGs)*, which are an extension of context-free grammars, to obtain our results. It is known that SCFGs can be used to model RNA secondary structures (see, e.g. [SBH+94]). Furthermore, SCFGs have already been used successfully for the prediction of RNA secondary structure [KH99, KH03]. For an introduction on stochastic context-free languages, see for example [HF71].

5.1.1 Basic Concepts

A formal definition is given as follows:

Definition 5.1 ([Neb04a, Neb02b]) *A* stochastic context-free grammar (SCFG) *is a* 5*-tuple* $G = (I, T, R, S, P)$*, where* I *(resp.* T*) is an alphabet (finite set) of intermediate (resp. terminal) symbols (*I *and* T *are disjoint),* $S \in I$ *is a distinguished intermediate symbol called* axiom*,* $R \subset I \times (I \cup T)^*$ *is a finite set of production rules and* P *is a mapping from* R *to* $[0,1]$ *such that each rule* $f \in R$ *is equipped with a probability* $p_f := P(f)$*. The probabilities are chosen in such a way that for all* $A \in I$ *the equaility*

$$\sum_{f \in R} p_f \cdot \delta_{Q(f),A} = 1$$

holds. Here, δ *is Kronecker's delta and* $Q(f)$ *denotes the source of the production* f*, i.e. the first component* A *of a production rule* $(A, \alpha) \in R$*. In the sequel, we will write* $p_f : A \to \alpha$ *instead of* $f = (A, \alpha) \in R$*,* $p_f = P(f)$*.*

Consequently, for a SCFG $G_{\mathrm{st}} := (I_{\mathrm{st}}, T_{\mathrm{st}}, R_{\mathrm{st}}, S_{\mathrm{st}}, P_{\mathrm{st}})$, the mapping $P_{\mathrm{st}} : R_{\mathrm{st}} \to [0,1]$ provides a probability distribution on the production rules that have the same left-hand side.
The concepts of derivation and ambiguity for SCFGs are the same as for usual context-free grammars. This means that each word $w \in \mathcal{L}(G_{\mathrm{st}})$ is generated in exactly the same way as for the corresponding context-free grammar $(I_{\mathrm{st}}, T_{\mathrm{st}}, R_{\mathrm{st}}, S_{\mathrm{st}})$.
It has to be mentioned that in many cases, the probability distribution on the production rules of a SCFG G_{st} implies a probability distribution on the words of the language $\mathcal{L}(G_{\mathrm{st}})$. The SCFG G_{st} is then called *consistent*.
Considering a consistent SCFG G_{st}, the mapping $P_{\mathrm{st}} : R_{\mathrm{st}} \to [0,1]$ assigns a probability $\mathrm{Prob}(d)$ to each derivation d of a word $w \in \mathcal{L}(G_{\mathrm{st}})$. The probability $\mathrm{Prob}(d)$ of a given derivation d is equal to the product of the probabilities of the production rules used in this derivation d. Furthermore, we can use the mapping P_{st} to compute the probability $\mathrm{Prob}(w)$ for each word $w \in \mathcal{L}(G_{\mathrm{st}})$.
As the consistent SCFG G_{st} can be ambiguous, a word $w \in \mathcal{L}(G_{\mathrm{st}})$ may have more than one derivation. In fact, if a word $w \in \mathcal{L}(G_{\mathrm{st}})$ has k different leftmost derivations $d_1, \ldots, d_k$, then the probability $\mathrm{Prob}(w)$ is given by $\sum_{i=1}^{k} \mathrm{Prob}(d_i)$, which means that we must sum up the probabilities of all possible leftmost derivations of this word w. Thus, if the consistent SCFG G_{st} is unambiguous, then the probability $\mathrm{Prob}(w)$ of a word $w \in \mathcal{L}(G_{\mathrm{st}})$ is equal to the product of the probabilities $P_{\mathrm{st}}(f)$ of the production rules $f \in R_{\mathrm{st}}$ that have to be used to generate this word w.

5.1.2 Training of Stochastic Context-Free Grammars

The probabilities of a SCFG G_{st} which generates the language $\mathcal{L}(G_{\mathrm{st}})$ can be trained from a database of words $w \in \mathcal{L}(G_{\mathrm{st}})$. The training of SCFGs is based on the maximum likelihood principle which was invented by R. A. Fisher around 1912. This method works as follows: On a fixed sample from a larger population, the free parameters of the underlying probability model are tuned in such a way that the sample has maximum likelihood. This means that other values for the free parameters make the observation of the sample less likely.
Obviously, in the context of training of a SCFG G_{st} from a database of words $w \in \mathcal{L}(G_{\mathrm{st}})$, the fixed sample is given by the words in the database and the free parameters are the probabilities of the production rules of the SCFG G_{st}. Hence, training the SCFG G_{st} fits the probabilities of the production rules of G_{st} so that words $w \in \mathcal{L}(G_{\mathrm{st}})$ closely match the sample set of words provided for the training. Several methods for the empirical estimation of SCFGs have been proposed in the literature which provide consistent SCFGs. For example, assigning relative frequencies found by counting the production rules used in the

leftmost derivations of a finite sample of words $w \in \mathcal{L}(G_{\mathrm{st}})$ results in a consistent SCFG G_{st} and theses probabilities are then a maximum likelihood estimate [CG98]. For unambiguous SCFGs, the relative frequencies can be counted efficiently, as for every word, there is only one leftmost derivation to consider.

5.1.3 Stochastic Context-Free Grammars and Probability Generating Functions

Translating a consistent SCFG according to the ideas of Schützenberger [CS63] yields a *probability generating function*, which is defined as follows:

Definition 5.2 ([SF01]) *Given a random variable X that takes on only nonnegative integer values, with $p_k := \Pr[X = k]$, the function $P(u) = \sum_{k \geq 0} p_k u^k$ is called the* probability generating function (PGF) *for the random variable.*

When deriving a probability generating function from a SCFG, the kth coefficient is obviously given by the probability for a word of length k in the language generated.
Thus, for a given consistent SCFG G_{st} and the corresponding probability generating function $P(z) = \sum_{k \geq 0} p_k z^k$, the probabilities p_k must provide a probability distribution on the words $w \in \mathcal{L}(G_{\mathrm{st}})$, and therefore they must sum up to 1, i.e. $P(1) = 1$ must hold. Consequently, by evaluating the function $P(z) = \sum_{k \geq 0} \Pr[X = k] z^k$ derived from a SCFG G_{st} for $z = 1$, i.e. by computing $P(1)$, we can check whether the SCFG G_{st} is consistent or not.

5.2 Computation of the Expected Free Energy

As our first goal, we want to determine the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of a stochastic model derived from biological data, more precisely under the assumption of a stochastic model derived from a database of SSU and LSU rRNA secondary structure data. This database contains 1866 SSU and LSU rRNA secondary structure $\mathbf{S} \neq \emptyset$ which we obtained from the databases [WdPWW02] and [WRdP$^+$01]. In particular, this database consists of 1308 SSU rRNA secondary structures $\mathbf{S} \neq \emptyset$ which we obtained from the database [WdPWW02] and 558 LSU rRNA secondary structures $\mathbf{S} \neq \emptyset$ which we obtained from the database [WRdP$^+$01]. Each secondary structure $\mathbf{S} \neq \emptyset$ of size n is given as pair of bar-bracket representation s of length n and corresponding primary structure of length n. Note that for the sake of simplicity, this database of SSU and LSU rRNA secondary structure $\mathbf{S} \neq \emptyset$ will be referred to as biological database in the sequel. Before we start, it should be mentioned that for most of the computations that we had to perform in order to obtain the results that will follow, we have used the Mathematica 6.0 software by Wolfram Research.
Let $\mathcal{S}$ be the combinatorial class of all different bar-bracket representations of secondary structures $\mathbf{S} \neq \emptyset$. Hence, due to the constraint $\mathbf{S} \neq \emptyset$, no bar-bracket representations of completely unpaired structures are contained in this combinatorial class $\mathcal{S}$, as they are assigned no free energy. We model the combinatorial class $\mathcal{S}$ as formal language $\mathcal{L}$. Considering the k-loop decomposition of a secondary structure $\mathbf{S}$ (see Definition 2.4), a formal definition of this language $\mathcal{L}$ can immediately be given as follows:

Definition 5.3 *The language $\mathcal{L}$ containing exactly the elements of the combinatorial class $\mathcal{S}$ is given by $\mathcal{L} := \mathcal{L}_u \mathcal{L}_{lu}^+$, where $\mathcal{L}_{lu} := (\mathcal{L}_l)\mathcal{L}_u$, $\mathcal{L}_u := \{|\}^*$ is the language of all bar-bracket representations of single-stranded regions and $\mathcal{L}_l$ is the language of all bar-bracket representations of k-loops, i.e. is the smallest language satisfying the following conditions:*

1. $\{|\}^+ \setminus \{|, ||\} \subset \mathcal{L}_l$ *(bar-bracket representations of hairpin loops).*
2. *If $w \in \mathcal{L}_l$, then $(w) \in \mathcal{L}_l$ (bar-bracket representation of a stacked pair).*
3. *If $w \in \mathcal{L}_l$, then $\{|\}^+(w) \subset \mathcal{L}_l$ and $(w)\{|\}^+ \subset \mathcal{L}_l$ (bar-bracket representations of bulge loops).*
4. *If $w \in \mathcal{L}_l$, then $\{|\}^+(w)\{|\}^+ \subset \mathcal{L}_l$ (bar-bracket representations of interior loops).*
5. *If $w_1, \ldots, w_n \in \mathcal{L}_l$ and $n \geq 2$, then $\mathcal{L}_u(w_1)\mathcal{L}_u(w_2) \cdots \mathcal{L}_u(w_n)\mathcal{L}_u \subset \mathcal{L}_l$ (bar-bracket representations of multibranched loops).*

A context-free grammar which unambiguously generates exactly the language $\mathcal{L}$ given in Definition 5.3 can obviously be given as follows:

Definition 5.4 *The context-free grammar G generating exactly the language $\mathcal{L}$ is given by $G = (I_G, \Sigma_G, R_G, S)$, where $I_G = \{S, T, C, A, L, G, B, F, H, P, Q, R, J, K, M, N, U\}$, $\Sigma_G = \{(,), |\}$ and R_G contains exactly the following rules:*

$$
\begin{array}{lll}
f_1 = S \to TAC, & f_{15} = G \to (L)B||, & f_{29} = Q \to ||(L)K|||, \\
f_2 = T \to TAC, & f_{16} = G \to |(L), & f_{30} = Q \to |||J(L)K||, \\
f_3 = T \to C, & f_{17} = G \to ||B(L), & f_{31} = R \to |(L)K|||, \\
f_4 = C \to C|, & f_{18} = B \to B|, & f_{32} = R \to |||J(L)|, \\
f_5 = C \to \epsilon, & f_{19} = B \to \epsilon, & f_{33} = J \to J|, \\
f_6 = A \to (L), & f_{20} = F \to |||, & f_{34} = J \to \epsilon, \\
f_7 = L \to (L), & f_{21} = F \to ||||, & f_{35} = K \to K|, \\
f_8 = L \to M, & f_{22} = F \to |||||H, & f_{36} = K \to \epsilon, \\
f_9 = L \to P, & f_{23} = H \to H|, & f_{37} = M \to U(L)U(L)N, \\
f_{10} = L \to Q, & f_{24} = H \to \epsilon, & f_{38} = N \to U(L)N, \\
f_{11} = L \to R, & f_{25} = P \to |(L)|, & f_{39} = N \to U, \\
f_{12} = L \to F, & f_{26} = P \to |(L)||, & f_{40} = U \to U|, \\
f_{13} = L \to G, & f_{27} = P \to ||(L)|, & f_{41} = U \to \epsilon. \\
f_{14} = G \to (L)|, & f_{28} = P \to ||(L)||, &
\end{array}
$$

For the grammar G given in Definition 5.4, different intermediate symbols have been used to distinguish between different substructures. In fact, the grammar distinguishes not only between the different types of k-loops, but also between some special types for hairpin, bulge and interior loops. More precisely, this grammar was constructed to distinguish between all the different classes of substructures for which there are different free energy rules according to the considered thermodynamic model. It should be mentioned that the grammar G given in Definition 5.4 has been constructed by modifying the unambiguous context-free grammar given in [Neb02b, Neb04a].

We now have to transform the unambiguous context-free grammar G given in Definition 5.4 into an unambiguous stochastic context-free grammar G_{sto} with $\mathcal{L}(G_{\text{sto}}) = \mathcal{L}(G)$. Obviously, we can immediately choose $G_{\text{sto}} = (I_G, \Sigma_G, R_G, S, P)$ and hence only have to find the mapping $P : R_G \to [0,1]$ such that each rule $f \in R_G$ is equipped with a probability $p_f := P(f)$, where the probabilities p_f must provide a probability distribution on the production rules having the same left-hand side.

Here, we decided to assign relative frequencies to the production rules in R_G, since such probabilities can be computed efficiently for unambiguous SCFGs. As we already know, by estimating the probabilities by their relative frequencies, the resulting grammar G_{sto} has the consistency property, i.e. the SCFG G_{sto} provides a probability distribution on the language $\mathcal{L}(G_{\text{sto}}) = \mathcal{L}$.

We have trained the probabilities (relative frequencies) of our SCFG G_{sto} from the bar-bracket representations $s \in \mathcal{L}(G_{\text{sto}})$ given in our biological database. The resulting probabilities are given in Table 1, and their floating point approximations, rounded to the fifth decimal place, are given in Table 2 shown in Appendix B.

As each bar-bracket representation $s \in \mathcal{S}$ is contained in the language $\mathcal{L}(G_{\text{sto}})$ and hence is unambiguously generated by the grammar G_{sto}, we can translate the rule set R_G into the following system of equations:

$$
\begin{aligned}
S &= p_1 \cdot T \cdot A \cdot C, \\
T &= p_2 \cdot T \cdot A \cdot C + p_3 \cdot C, \\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1, \\
A &= p_6 \cdot z \cdot L \cdot z, \\
L &= p_7 \cdot z \cdot L \cdot z + p_8 \cdot M + p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + p_{12} \cdot F + p_{13} \cdot G, \\
G &= p_{14} \cdot z \cdot L \cdot z \cdot z + p_{15} \cdot z \cdot L \cdot z \cdot B \cdot z^2 + p_{16} \cdot z \cdot z \cdot L \cdot z + p_{17} \cdot z^2 \cdot B \cdot z \cdot L \cdot z, \\
B &= p_{18} \cdot B \cdot z + p_{19} \cdot 1, \\
F &= p_{20} \cdot z^3 + p_{21} \cdot z^4 + p_{22} \cdot z^5 \cdot H, \\
H &= p_{23} \cdot H \cdot z + p_{24} \cdot 1, \\
P &= p_{25} \cdot z \cdot z \cdot L \cdot z \cdot z + p_{26} \cdot z \cdot z \cdot L \cdot z \cdot z^2 + p_{27} \cdot z^2 \cdot z \cdot L \cdot z \cdot z + p_{28} \cdot z^2 \cdot z \cdot L \cdot z \cdot z^2, \\
Q &= p_{29} \cdot z^2 \cdot z \cdot L \cdot z \cdot K \cdot z^3 + p_{30} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot K \cdot z^2, \\
R &= p_{31} \cdot z \cdot z \cdot L \cdot z \cdot K \cdot z^3 + p_{32} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot z, \\
J &= p_{33} \cdot J \cdot z + p_{34} \cdot 1, \\
K &= p_{35} \cdot K \cdot z + p_{36} \cdot 1, \\
M &= p_{37} \cdot U \cdot z \cdot L \cdot z \cdot U \cdot z \cdot L \cdot z \cdot N, \\
N &= p_{38} \cdot U \cdot z \cdot L \cdot z \cdot N + p_{39} \cdot U,
\end{aligned}
\tag{1}
$$

$$U = p_{40} \cdot U \cdot z + p_{41} \cdot 1.$$

As the SCFG G_{sto} is consistent, by solving this system for the axiom S of the grammar G_{sto}, we obtain a closed form of the probability generating function

$$S_{\text{sto}}(z) = \sum_{s \in \mathcal{S}} \text{Prob}(s) \cdot z^{|s|} = \sum_{n \geq 0} \left(\sum_{s \in \mathcal{S}_n} \text{Prob}(s) \right) \cdot z^n = \sum_{n \geq 0} s_{\text{sto},n} \cdot z^n.$$

Here, $\text{Prob}(s)$ is the probability of the bar-bracket word $s \in \mathcal{S}$ under the assumption of the probability distribution on the words in the combinatorial class $\mathcal{S}$ which is implied by the SCFG G_{sto}. Thus, $\text{Prob}(s)$ is the product of the probabilities of the production rules of the SCFG G_{sto} that have to be used to generate this word $s \in \mathcal{S}$.
Hence, $s_{\text{sto},n}$ is the probability that a bar-bracket representation s of length n is generated by the SCFG G_{sto}, i.e. the probability that a word $s \in \mathcal{S}$ has length n.
To be able to compute the desired expected free energy change, we have to incorporate free energy values into system (1). Therefore, we first have to recall that each factor $z = z^1$ in this system represents a symbol $t \in \Sigma_G = \{(,), |\}$ of length 1, such that in the PGF $S_{\text{sto}}(z)$, the variable z marks length. In addition to that, we now want to use a second variable y marking free energy changes. The resulting generating function is a so-called *bivariate* generating function. A formal definition based on [SF01] is given as follows:

Definition 5.5 *Given a doubly indexed sequence $(a_{nk})_{n \in \mathbb{N}_0, k \in K}$, where $K \subset \mathbb{R}$ is enumerable[6], the function*

$$A(z, u) = \sum_{n \in \mathbb{N}_0} \sum_{k \in K} a_{nk} u^k z^n$$

is called the bivariate generating function *(BGF) of the sequence.*
We use the notation $[u^k z^n]A(z, u)$ to refer to a_{nk}; $[z^n]A(z, u)$ to refer to $\sum_{k \in K} a_{nk} u^k$; and $[u^k]A(z, u)$ to refer to $\sum_{n \in \mathbb{N}_0} a_{nk} z^n$.

Hence, let $g_{\text{sto}}(s)$ denote the free energy change associated with the bar-bracket representation $s \in \mathcal{S}$ under the assumption of the stochastic model under consideration and let K_{sto} be an enumerable[7] subset of $\mathbb{R}$ with the property that for each $s \in \mathcal{S}$, $g_{\text{sto}}(s) \in K_{\text{sto}}$. Furthermore, let X be a random variable (for the length of an element $s \in \mathcal{S}$) that takes on values in $\mathbb{N}$, and let Y be a random variable (for the free energy change $g_{\text{sto}}(s)$ associated with a bar-bracket representation $s \in \mathcal{S}$) that takes on values in K_{sto}.
We thus aim at determining a closed form of the bivariate generating function

$$D_{\text{sto}}(z, y) = \sum_{n \in \mathbb{N}} \sum_{k \in K_{\text{sto}}} \Pr[Y = k \text{ and } X = n] \cdot y^k z^n,$$

where $[y^k z^n]D_{\text{sto}}(z, y) = \Pr[Y = k \text{ and } X = n]$ is the probability that a bar-bracket representation $s \in \mathcal{S}$ has length n and an associated free energy change of k $^{\text{kcal}}/_{\text{mol}}$. The combinatorial form of this bivariate generating function could be written as

$$D_{\text{sto}}(z, y) = \sum_{s \in \mathcal{S}} \left(\text{Prob}(s) \cdot y^{g_{\text{sto}}(s)} \right) \cdot z^{|s|}.$$

Once we have constructed the bivarite generating function $D_{\text{sto}}(z, y)$, the desired expected free energy change $\Delta G^\circ_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ under the assumption of the stochastic model under consideration can immediately be computed, as it is then given by

$$\frac{[z^n] \frac{\partial}{\partial y} D_{\text{sto}}(z, y)\big|_{y=1}}{[z^n] D_{\text{sto}}(z, y)\big|_{y=1}}.$$

Note that by using a consistent SCFG to obtain the corresponding bivarite generating function, the resulting expected value is in fact a conditional expected value, i.e. the expected value with respect to a conditional probability distribution. In particular, by considering the consistent SCFG G_{sto} generating exactly all the elements in $\mathcal{S}$, we obtain the expected free energy change $\Delta G^\circ_{37}(\mathbf{S})$ of a secondary structure

[6] For $K = \mathbb{N}_0$, we obtain the definition given in [SF01].

[7] Note that $K_{\text{sto}} \subset \mathbb{R}$ is enumerable, as the free energy changes are given by $^{\text{kcal}}/_{\text{mol}}$-values with a finite number of decimal places. Thus, by considering a suitable unit which is different to $^{\text{kcal}}/_{\text{mol}}$, we obtain a subset of $\mathbb{N}$.

$\mathbf{S} \neq \emptyset$ under the condition that this secondary structure $\mathbf{S}$ has size n.
The reason why we have to consider conditional probabilities is that the consistent SCFG G_{sto} provides a probability distribution for the generated language $\mathcal{L}(G_{\text{sto}})$ containing exactly all the elements in the combinatorial class $\mathcal{S}$. But to obtain the desired result for a given size n, we must consider a probability distribution on the class $\mathcal{S}_n$ only.
We now want to describe this approach to compute the expected free energy change of a secondary structure of size n formally. First, by computing the partial derivative of $D_{\text{sto}}(z, y)$ with respect to the variable y and then admitting the value 1 to variable y, we obtain a new generating function, which is given as follows:

$$\begin{aligned} E_{\text{sto}}(z) &:= \frac{\partial}{\partial y} D_{\text{sto}}(z,y)\big|_{y=1} \\ &= \frac{\partial}{\partial y}\left(\sum_{n\in\mathbb{N}}\sum_{k\in K_{\text{sto}}} \Pr[Y=k \text{ and } X=n]\cdot y^k z^n\right)\Big|_{y=1} \\ &= \sum_{n\in\mathbb{N}}\left(\sum_{k\in K_{\text{sto}}} \frac{\partial}{\partial y}\Pr[Y=k \text{ and } X=n]\cdot y^k\right)\cdot z^n\Big|_{y=1} \\ &= \sum_{n\in\mathbb{N}}\left(\sum_{k\in K_{\text{sto}}} k\cdot \Pr[Y=k \text{ and } X=n]\right)\cdot z^n. \end{aligned}$$

Consequently,

$$[z^n]\frac{\partial}{\partial y} D_{\text{sto}}(z)\big|_{y=1} = \sum_{k\in K_{\text{sto}}} k\cdot \Pr[Y=k \text{ and } X=n].$$

But obviously, $\Pr[Y=k \text{ and } X=n]$ does not provide a probability measure. However, for $\Pr[X=n]\neq 0$, switching to the conditional probability

$$\Pr[Y=k \mid X=n] = \frac{\Pr[Y=k \text{ and } X=n]}{\Pr[X=n]}$$

yields a probability measure on the elements of size n. Hence, we obviously must divide $[z^n]\frac{\partial}{\partial u}A(z,u)\big|_{u=1}$ by $\Pr[X=n]$ to obtain the desired expected value.
Since X is a random variable for the length of an element $s\in\mathcal{S}$, $\Pr[X=n]$ is the probability that an element $s\in\mathcal{S}$ has length n and is obviously given by the nth coefficient of the PGF for random variable X, which is given by

$$S_{\text{sto}}(z) = D_{\text{sto}}(z,y)\big|_{y=1} = \sum_{n\in\mathbb{N}} \Pr[X=n]\cdot z^n.$$

Thus, we have to divide the nth coefficient of the generating function $E_{\text{sto}}(z)$ by the nth coefficient of $S_{\text{sto}}(z)$, as this yields

$$\begin{aligned} \frac{[z^n]E_{\text{sto}}(z)}{[z^n]S_{\text{sto}}(z)} &= \frac{\sum_{k\in K_{\text{sto}}} k\cdot\Pr[Y=k \text{ and } X=n]}{\Pr[X=n]} \\ &= \sum_{k\in K_{\text{sto}}} k\cdot\frac{\Pr[Y=k \text{ and } X=n]}{\Pr[X=n]} \\ &= \sum_{k\in K_{\text{sto}}} k\cdot\Pr[Y=k\mid X=n] \\ &= \mathbb{E}\left[g_{\text{sto}}(s)\mid |s|=n\right], \end{aligned}$$

which is the expected free energy change associated with a bar-bracket representation $s\in\mathcal{S}$, under the condition that this bar-bracket word s has length n (conditional expectation).
According to the previous discussion, we now have to modify system (1) by multiplying some terms with free energy values, such that solving it for the variable S yields a closed form of the desired bivariate generating function $D_{\text{uni}}(z,y)$. In fact, we have to decide which free energy values will be used for the free energy function g_{sto} and how they should be incorporated into system (1). According to our thermodynamic model, most of the contributions to the free energy change of a secondary structure $\mathbf{S}$ are

sequence-dependent. But for a given bar-bracket representation s of a secondary structure $\mathbf{S}$, we do not know the corresponding RNA sequence $\mathbf{R}$. Therefore, we have to use fixed, sequence-independent values for the different contributions. Similarly, we have to use fixed values for the different length-dependent free energy contributions. Hence, we decided to use expected values for all the different free energy contributions that are considered in the used thermodynamic model.
For each of the different structures that are distinguished, all the expected values of the free energy contributions that are considered to compute the free energy of this structure according to the termodynamic model have to be summed up in the exponent of y each time such a structure is generated by the grammar G_{sto}. This immediately yields the following system of equations:

$$
\begin{aligned}
S &= p_1 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot S \cdot A \cdot C,\\
T &= p_2 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot T \cdot A \cdot C + p_3 \cdot C,\\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1,\\
A &= p_6 \cdot z \cdot L \cdot z,\\
L &= p_7 \cdot y^{(\text{se})} \cdot z \cdot L \cdot z + p_8 \cdot y^{(\text{MBLinitiation+stackingMulti+termAUpenML})} \cdot M +\\
&\quad p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + p_{12} \cdot F + p_{13} \cdot y^{(\text{ldeb})} \cdot G,\\
G &= p_{14} \cdot y^{(\text{seBulge})} \cdot z \cdot L \cdot z \cdot z + p_{15} \cdot y^{(2\cdot\text{termAUpenBL})} \cdot z \cdot L \cdot z \cdot B \cdot z^2 +\\
&\quad p_{16} \cdot y^{(\text{seBulge})} \cdot z \cdot z \cdot L \cdot z + p_{17} \cdot y^{(2\cdot\text{termAUpenBL})} \cdot z^2 \cdot B \cdot z \cdot L \cdot z,\\
B &= p_{18} \cdot B \cdot z + p_{19} \cdot 1,\\
F &= p_{20} \cdot y^{(\text{ldeh+termAUpenHL+GGGLoopBonus+cHairpinOf3})} \cdot z^3 +\\
&\quad p_{21} \cdot y^{(\text{ldeh+tmseh+GGGLoopBonus+cHairpin+tetra})} \cdot z^4 +\\
&\quad p_{22} \cdot y^{(\text{ldeh+tmseh+GGGLoopBonus+cHairpin})} \cdot z^5 \cdot H,\\
H &= p_{23} \cdot H \cdot z + p_{24} \cdot 1,\\
P &= p_{25} \cdot y^{(\text{ile1x1})} \cdot z \cdot z \cdot L \cdot z \cdot z + p_{26} \cdot y^{(\text{ile1x2})} \cdot z \cdot z \cdot L \cdot z \cdot z^2 +\\
&\quad p_{27} \cdot y^{(\text{ile1x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z + p_{28} \cdot y^{(\text{ile2x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z^2,\\
Q &= p_{29} \cdot y^{(2\cdot\text{tmsei+ldei+asym})} \cdot z^2 \cdot z \cdot L \cdot z \cdot K \cdot z^3 + p_{30} \cdot y^{(2\cdot\text{tmsei+ldei+asym})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot K \cdot z^2,\\
R &= p_{31} \cdot y^{(2\cdot\text{tbp1xNil+ldei+asym})} \cdot z \cdot z \cdot L \cdot z \cdot K \cdot z^3 + p_{32} \cdot y^{(2\cdot\text{tbp1xNil+ldei+asym})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot z,\\
J &= p_{33} \cdot J \cdot z + p_{34} \cdot 1,\\
K &= p_{35} \cdot K \cdot z + p_{36} \cdot 1,\\
M &= p_{37} \cdot y^{(2\cdot\text{stackingMulti}+2\cdot\text{termAUpenML})} \cdot U \cdot z \cdot L \cdot z \cdot U \cdot z \cdot L \cdot z \cdot N,\\
N &= p_{38} \cdot y^{(\text{stackingMulti+termAUpenML})} \cdot U \cdot z \cdot L \cdot z \cdot N + p_{39} \cdot U,\\
U &= p_{40} \cdot U \cdot z + p_{41} \cdot 1.
\end{aligned}
\tag{2}
$$

Now, we want to calculate suitable expected values for the parameters used in system (2). Therefore, recall that we have only used the words $s \in \mathcal{S}$ given in our biological database to derive the stochastic model under consideration. Hence, to obtain a free energy model for this stochastic model for RNA secondary structures, i.e. for the elements $s \in \mathcal{S}$ under the assumption of this stochastic model, we obviously have to consider the corresponding primary structures for the bar-bracket words $s \in \mathcal{S}$ given in our database. Consequently, we have to use the same database that we used to derive the probability distribution on the words $s \in \mathcal{S}$ to compute expected values for the different free energy contributions.
Thus, we have to compute the desired expected values for the different free energy contributions by sequence counting using our biological database[8]. To reach this goal, we have to recall that there are no free energy parameters for non-canonical base pairs and hence, according to our thermodynamical model, non-canonical base pairs must be treated as mismatches to obtain appropriate expected values by sequence counting using our biological database. The resulting expected values in floating point representation, as well as their corresponding rational approximations, are given in Table 3 shown in Appendix B. Thus, using the rational approximations[9] given in the fourth column of Table 3, we can

[8]Sequence counting means that for each free energy contribution that has to be considered to compute the free energy of a certain (special) loop type, we have to sum up all the corresponding free energy values for all loops of this (special) type that occur in our database and then divide this sum by the number of loops of this (special) type that occur in the used database, i.e. by the number of values that have been summed up.

[9]Note that we have used the rational approximations instead of the computed floating point values to avoid numerical imprecisions.

solve system (2) for the variable S to obtain a closed form of the desired bivariate generating function $D_{\text{sto}}(z, y)$ and then proceed in the described way.
We now want to use Darboux's theorem [KW89] to determine asymptotics for $[z^n]E_{\text{sto}}(z)$ and $[z^n]S_{\text{sto}}(z)$, respectively, then divide the resulting asymptotics one by the other and compute the series expansion of this fraction about $n \to \infty$ (to eliminate binomial coefficients) to obtain an asymptotic for the expected free energy change of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of the stochastic model under consideration. This expected free energy change will be denoted by $\mu_{\text{sto},n}$ in the sequel.
By applying Darboux's theorem with the choices $m = 1$ and $m = 2$ to the functions $S_{\text{sto}}(z)$ and $E_{\text{sto}}(z)$, respectively, and afterwards computing floating point approximations of the series expansions of the resulting asymptotics about $n \to \infty$, we obtain the following results:

Lemma 5.1 *Under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structures, the expected number of secondary structures $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.000129672^{-n} \left(\frac{26.96760121}{n^{3/2}} - \frac{102833.1842}{n^{5/2}} + \mathcal{O}\left(n^{-7/2}\right) \right), n \to \infty.$$

Lemma 5.2 *Under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structures, the first factorial moment for the free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.000129672^{-n} \left(-\frac{6.683382518}{\sqrt{n}} + \frac{26541.34513}{n^{3/2}} + \mathcal{O}\left(n^{-7/2}\right) \right), n \to \infty.$$

Theorem 5.3 *Under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structures, the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ (in $^{\text{kcal}}/_{\text{mol}}$) of a secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$-0.2478300708n + 39.16513746 + \mathcal{O}\left(\frac{1}{n}\right), n \to \infty.$$

5.2.1 Computation of the Variance of the Free Energy

Now, we would like to compute the variance $\sigma^2_{\text{sto},n}$ of the free energy $\Delta G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of the stochastic model under consideration. To reach this goal, we first consider the second partial derivate of the bivariate generating function $D_{\text{sto}}(z, y)$ with respect to the variable y at the point $y = 1$. This generating function is given by

$$\begin{aligned}
F_{\text{sto}}(z) &:= \frac{\partial^2}{\partial y^2} D_{\text{sto}}(z, y)\big|_{y=1} \\
&= \frac{\partial^2}{\partial y^2} \left(\sum_{n \in \mathbb{N}} \sum_{k \in K_{\text{sto}}} \Pr[Y = k \text{ and } X = n] \cdot y^k z^n \right) \Big|_{y=1} \\
&= \sum_{n \in \mathbb{N}} \left(\sum_{k \in K_{\text{sto}}} \frac{\partial^2}{\partial y^2} \Pr[Y = k \text{ and } X = n] \cdot y^k \right) \cdot z^n \Big|_{y=1} \\
&= \sum_{n \in \mathbb{N}} \left(\sum_{k \in K_{\text{sto}}} k \cdot (k-1) \cdot \Pr[Y = k \text{ and } X = n] \right) \cdot z^n \\
&= \sum_{n \in \mathbb{N}} \left(\sum_{k \in K_{\text{sto}}} (k^2 - k) \cdot \Pr[Y = k \text{ and } X = n] \right) \cdot z^n.
\end{aligned}$$

As

$$\begin{aligned}
\frac{[z^n]F_{\text{sto}}(z)}{[z^n]S_{\text{sto}}(z)} &= \frac{\sum_{k \in K_{\text{sto}}} (k^2 - k) \cdot \Pr[Y = k \text{ and } X = n]}{\Pr[X = n]} \\
&= \frac{1}{\Pr[X = n]} \sum_{k \in K_{\text{sto}}} \left(k^2 \cdot \Pr[Y = k \text{ and } X = n] - k \cdot \Pr[Y = k \text{ and } X = n] \right) \\
&= \frac{1}{\Pr[X = n]} \left(\sum_{k \in K_{\text{sto}}} k^2 \cdot \Pr[Y = k \text{ and } X = n] - \sum_{k \in K_{\text{sto}}} k \cdot \Pr[Y = k \text{ and } X = n] \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k \in K_{\text{sto}}} k^2 \cdot \frac{\Pr[Y = k \text{ and } X = n]}{\Pr[X = n]} - \sum_{k \in K_{\text{sto}}} k \cdot \frac{\Pr[Y = k \text{ and } X = n]}{\Pr[X = n]} \\
&= \sum_{k \in K_{\text{sto}}} k^2 \cdot \Pr[Y = k \mid X = n] - \sum_{k \in K_{\text{sto}}} k \cdot \Pr[Y = k \mid X = n] \\
&= \mathbb{E}\left[g_{\text{sto}}(s)^2 \mid |s| = n\right] - \mathbb{E}\left[g_{\text{sto}}(s) \mid |s| = n\right]
\end{aligned}
$$

holds, the desired variance $\sigma^2_{\text{sto},n}$ is given by

$$
\begin{aligned}
\sigma^2_{\text{sto},n} &= \frac{[z^n]\frac{\partial^2}{\partial y^2} D_{\text{sto}}(z,y)\big|_{y=1}}{[z^n] D_{\text{sto}}(z,y)\big|_{y=1}} + \frac{[z^n]\frac{\partial}{\partial y} D_{\text{sto}}(z,y)\big|_{y=1}}{[z^n] D_{\text{sto}}(z,y)\big|_{y=1}} - \left(\frac{[z^n]\frac{\partial}{\partial y} D_{\text{sto}}(z,y)\big|_{y=1}}{[z^n] D_{\text{sto}}(z,y)\big|_{y=1}}\right)^2 \\
&= \frac{[z^n] F_{\text{sto}}(z)}{[z^n] S_{\text{sto}}(z)} + \frac{[z^n] E_{\text{sto}}(z)}{[z^n] S_{\text{sto}}(z)} - \left(\frac{[z^n] E_{\text{sto}}(z)}{[z^n] S_{\text{sto}}(z)}\right)^2 \\
&= \frac{[z^n] F_{\text{sto}}(z)}{[z^n] S_{\text{sto}}(z)} + \mu_{\text{sto},n} - \mu^2_{\text{sto},n} \\
&= \mathbb{E}\left[g_{\text{sto}}(s)^2 \mid |s| = n\right] - \mathbb{E}\left[g_{\text{sto}}(s) \mid |s| = n\right] + \mathbb{E}\left[g_{\text{sto}}(s) \mid |s| = n\right] - \left(\mathbb{E}\left[g_{\text{sto}}(s) \mid |s| = n\right]\right)^2 \\
&= \mathbb{E}\left[g_{\text{sto}}(s)^2 \mid |s| = n\right] - \left(\mathbb{E}\left[g_{\text{sto}}(s) \mid |s| = n\right]\right)^2 \\
&= \operatorname{Var}\left[g_{\text{sto}}(s) \mid |s| = n\right],
\end{aligned}
$$

which is the variance of the free energy change $g_{\text{sto}}(s)$ associated with a random bar-bracket representation $s \in \mathcal{S}$ in the stochastic model under consideration, under the condition that this bar-bracket word s has length n (conditional variance).
By applying Darboux's theorem with a choice of $m = 3$ to the second partial derivative $F_{\text{sto}}(z)$ and afterwards computing a floating point approximation of the series expansion of the resulting asymptotic about $n \to \infty$, we obtain:

Lemma 5.4 *Under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structures, the second factorial moment for the free energy change $\Delta G^\circ_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$
1.000129672^{-n} \left(1.656343163\sqrt{n} - \frac{6764.547343}{\sqrt{n}} + \mathcal{O}\left(n^{-7/2}\right)\right), n \to \infty.
$$

Using the determined asymptotics for $[z^n]S_{\text{sto}}(z)$, $[z^n]E_{\text{sto}}(z)$ and $[z^n]F_{\text{sto}}(z)$, we immediately obtain the desired asymptotic for the variance $\sigma^2_{\text{sto},n}$. A floating point approximation of this asymptotic is given in the following theorem.

Theorem 5.5 *Under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structures, the variance of the free energy change $\Delta G^\circ_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n (in $\text{kcal}^2/\text{mol}^2$) is asymptotically given by*

$$
2.531493699n + \mathcal{O}(1), n \to \infty.
$$

5.3 Alternative Free Energy Model

We now want to work out a different free energy model for our stochastic model for RNA secondary structures, where the expected values for length-dependent contributions are computed in a different way than before. As any of these length-dependent free energy contributions depends on the number of unpaired nucleotides in loops of a certain type, we could alternatively compute the expected free energy contribution of one unpaired nucleotide in loops of this type and apply the resulting expected value to each unpaired nucleotide.
Using such expected values for each nucleotide in a loop, the length-dependence is modeled better than before, as loops of different lengths are assigned different free energy values, whereas by using expected values for each loop, very small loops are assigned the same free energy as extremely large loops.
By modifiying system (2), we immediately obtain an appropriate system of equations for this new free energy model for our stochastic model for RNA secondary structures. The resulting system is given as follows:

$$
S = p_1 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot T \cdot A \cdot C,
$$

$$
\begin{aligned}
T &= p_2 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot T \cdot A \cdot C + p_3 \cdot C,\\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1,\\
A &= p_6 \cdot z \cdot L \cdot z,\\
L &= p_7 \cdot y^{(\text{se})} \cdot z \cdot L \cdot z +\\
&\quad p_8 \cdot y^{(\text{MBLOffset+stackingMulti+termAUpenML+MBLHelixPenalty})} \cdot M +\\
&\quad p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + p_{12} \cdot F + p_{13} \cdot G,\\
G &= p_{14} \cdot y^{(\text{seBulge})} \cdot y^{(\text{ldebPerNuc})} \cdot z \cdot L \cdot z \cdot z +\\
&\quad p_{15} \cdot y^{(2\cdot\text{termAUpenBL})} \cdot y^{(2\cdot\text{ldebPerNuc})} \cdot z \cdot L \cdot z \cdot B \cdot z^2 +\\
&\quad p_{16} \cdot y^{(\text{seBulge})} \cdot y^{(\text{ldebPerNuc})} \cdot z \cdot z \cdot L \cdot z +\\
&\quad p_{17} \cdot y^{(2\cdot\text{termAUpenBL})} \cdot y^{(2\cdot\text{ldebPerNuc})} \cdot z^2 \cdot B \cdot z \cdot L \cdot z,\\
B &= p_{18} \cdot y^{(\text{ldebPerNuc})} \cdot B \cdot z + p_{19} \cdot 1,\\
F &= p_{20} \cdot y^{(\text{termAUpenHL+GGGLoopBonus+cHairpinOf3})} \cdot y^{(3\cdot\text{ldehPerNuc})} \cdot z^3 +\\
&\quad p_{21} \cdot y^{(\text{tmseh+GGGLoopBonus+tetra})} \cdot y^{(4\cdot\text{ldehPerNuc})} \cdot y^{(4\cdot\text{cHairpinPerNuc})} \cdot z^4 +\\
&\quad p_{22} \cdot y^{(\text{tmseh+GGGLoopBonus})} \cdot y^{(5\cdot\text{ldehPerNuc})} \cdot y^{(5\cdot\text{cHairpinPerNuc})} \cdot z^5 \cdot H,\\
H &= p_{23} \cdot y^{(\text{ldehPerNuc})} \cdot y^{(\text{cHairpinPerNuc})} \cdot H \cdot z + p_{24} \cdot 1,\\
P &= p_{25} \cdot y^{(\text{ile1x1})} \cdot z \cdot z \cdot L \cdot z \cdot z + p_{26} \cdot y^{(\text{ile1x2})} \cdot z \cdot z \cdot L \cdot z \cdot z^2 +\\
&\quad p_{27} \cdot y^{(\text{ile1x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z + p_{28} \cdot y^{(\text{ile2x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z^2,\\
Q &= p_{29} \cdot y^{(2\cdot\text{tmsei+asym})} \cdot y^{(5\cdot\text{ldeiPerNuc})} \cdot z^2 \cdot z \cdot L \cdot z \cdot K \cdot z^3 +\\
&\quad p_{30} \cdot y^{(2\cdot\text{tmsei+asym})} \cdot y^{(5\cdot\text{ldeiPerNuc})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot K \cdot z^2,\\
R &= p_{31} \cdot y^{(2\cdot\text{tbp1xNil+asym})} \cdot y^{(4\cdot\text{ldeiPerNuc})} \cdot z \cdot z \cdot L \cdot z \cdot K \cdot z^3 +\\
&\quad p_{32} \cdot y^{(2\cdot\text{tbp1xNil+asym})} \cdot y^{(4\cdot\text{ldeiPerNuc})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot z,\\
J &= p_{33} \cdot y^{(\text{ldeiPerNuc})} \cdot J \cdot z + p_{34} \cdot 1,\\
K &= p_{35} \cdot y^{(\text{ldeiPerNuc})} \cdot K \cdot z + p_{36} \cdot 1,\\
M &= p_{37} \cdot y^{(2\cdot(\text{stackingMulti+termAUpenML+MBLHelixPenalty}))} \cdot U \cdot z \cdot L \cdot z \cdot U \cdot z \cdot L \cdot z \cdot N,\\
N &= p_{38} \cdot y^{(\text{stackingMulti+termAUpenML+MBLHelixPenalty})} \cdot U \cdot z \cdot L \cdot z \cdot N + p_{39} \cdot U,\\
U &= p_{40} \cdot y^{(\text{MBLFreeBasePenalty})} \cdot U \cdot z + p_{41} \cdot 1.
\end{aligned}
\tag{3}
$$

Again, we can compute suitable values for the free energy parameters used in system (3) by sequence counting using our biological database. The resulting floating point representations of the expected values and their respective rational approximations are given in Table 4 shown in Appendix B. Thus, using the rational approximations given in the fourth column of Table 4, we can solve the system (3) for the variable S to obtain a closed form of a bivariate generating function $\widehat{D}_{\text{sto}}(z, y)$ and then proceed in the same way as we have done for $D_{\text{sto}}(z, y)$.

Obviously, the asymptotic for the expected number of secondary structures $\mathbf{S} \neq \emptyset$ of size n given in Lemma 5.1 does not depend on the free energy function g_{sto} used in the generating function $D_{\text{sto}}(z, y)$. The only difference between the two bivariate generating functions $D_{\text{sto}}(z, y)$ and $\widehat{D}_{\text{sto}}(z, y)$ is in fact the used free energy function. Thus, the asymptotic given in Lemma 5.1 also holds for the currently considered free energy model.

Furthermore, applying Darboux's theorem with the same choice of m to the generating function $\frac{\partial}{\partial y}\widehat{D}_{\text{sto}}(z, y)\big|_{y=1}$ and $\frac{\partial^2}{\partial y^2}\widehat{D}_{\text{sto}}(z, y)\big|_{y=1}$, respectively, as we have done to obtain the corresponding result for the first model, we obtain the following results:

Lemma 5.6 *Under the assumption of our second stochastic model derived from SSU and LSU rRNA secondary structures, the first factorial moment for the free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.000129672^{-n}\left(-\frac{4.967149985}{\sqrt{n}} + \frac{19941.52582}{n^{3/2}} + \mathcal{O}\left(n^{-7/2}\right)\right), n \to \infty.$$

Lemma 5.7 *Under the assumption of our second stochastic model derived from SSU and LSU rRNA secondary structures, the second factorial moment for the free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a random*

secondary structure $\mathbf{S} \neq \emptyset$ *of size* n *is asymptotically given by*

$$1.000129672^{-n}\left(0.9148970567\sqrt{n} - \frac{3745.492293}{\sqrt{n}} + \mathcal{O}\left(n^{-7/2}\right)\right), n \to \infty.$$

Using the three determined asymptotics for $[z^n]\widehat{D}_{\mathrm{sto}}(z,y)\big|_{y=1}$, $[z^n]\frac{\partial}{\partial y}\widehat{D}_{\mathrm{sto}}(z,y)\big|_{y=1}$ and $[z^n]\frac{\partial^2}{\partial y^2}\widehat{D}_{\mathrm{sto}}(z,y)\big|_{y=1}$, we immediately obtain the desired results for the expected free energy and the variance of the free energy. Floating point approximations of their series expansions about $n \to \infty$ are given in the following theorems.

Theorem 5.8 *Under the assumption of our second stochastic model derived from SSU and LSU rRNA secondary structures, the expected free energy change* $\Delta G^{\circ}_{37}(\mathbf{S})$ *(in* $^{\mathrm{kcal}}/_{\mathrm{mol}}$*) of a secondary structure* $\mathbf{S} \neq \emptyset$ *of size* n *is asymptotically given by*

$$-0.1841895371n + 37.10857372 + \mathcal{O}\left(\frac{1}{n}\right), n \to \infty.$$

Theorem 5.9 *Under the assumption of our second stochastic model derived from SSU and LSU rRNA secondary structures, the variance of the free energy change* $\Delta G^{\circ}_{37}(\mathbf{S})$ *of a random secondary structure* $\mathbf{S} \neq \emptyset$ *of size* n *(in* $^{\mathrm{kcal}^2}/_{\mathrm{mol}^2}$*) is asymptotically given by*

$$3.963452967n + \mathcal{O}(1), n \to \infty.$$

5.4 Comparison of the Different Free Energy Models

Finally, we want to compare the results derived for the two different free energy models that we have worked out for this stochastic models for RNA secondary structures to real world RNA secondary structure data in order to judge their quality.
For this reason, we have assigned a point $\{n, \Delta G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n which is given in our biological database. Figure 2 shows a plot of the derived asymptotics for the expected free energy as given in Theorem 5.3 and Theorem 5.8, respectively, as well as the 1866 points corresponding to the free energies of SSU and LSU rRNA secondary structures. Additionally, it also shows the line that best fits these 1866 points. Considering Figure 2, it seems that both models are realistic. Furthermore,

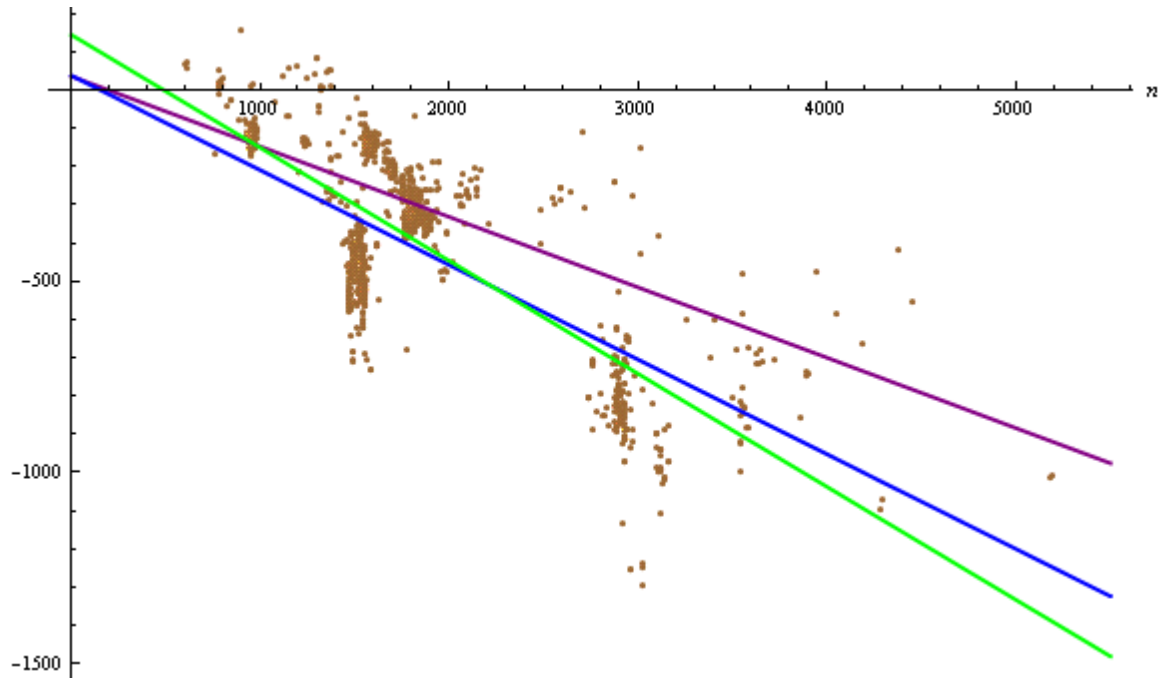

Figure 2: Plots of the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first (blue) and second (purple) model, respectively, together with the 1866 points $\{n, \Delta G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database (brown) and the line that best fits these points (green).

considering the line that best fits the 1866 points corresponding to the free energies of SSU and LSU rRNA secondary structures, it seems that for large values of n, the first model is more realistic than the second.
In addition to that, we observe that the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of the second model is significantly greater than the corresponding expected free energy under the assumption of the first model, and this difference grows with increasing value of n. The reason for this observation is due to the difference between our two free energy models: In the first model, destabilizing free energy contributions for certain (special) types of loops that depend on the number of unpaired bases resp. base pairs in the loop, are added for the whole structure, whereas in the second model, such destabilizing free energy contributions are added for each unpaired base resp. base pair in the loop. As a consequence, in the first model, very small loops are assigned the same free energy as extremely large loops, whereas in the second model, loops of different lengths are assigned different destabilizing free energy values. In fact, in the second model, loops with a greater number of unpaired bases resp. base pairs are assigned greater destabilizing free energies. Consequently, for each loop with a number of unpaired bases resp. base pairs that is large enough, the destabilizing free energy for this loop in the second model is greater than that in the first model. Thus, with increasing n, a secondary structure $\mathbf{S} \neq \emptyset$ of size n may contain more loops for which the corresponding destabilizing free energy in the second model is greater than the corresponding destabilizing free energy in the first model.

6 How to Use the Derived Results to Identify Good Predictions of RNA Secondary Structure

As we have pointed out, both free energy models that we worked out for the stochastic model derived from SSU and LSU rRNA secondary structures are more or less realistic. Therefore, we will now use Chebyshev's inequality to compute probabilities that the free energy of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is less than a given value away from the computed expected free energy of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first and second model, respectively.
The knowledge of these probabilities may help us to identify good predictions of RNA secondary structure. In fact, searching the set of all predicted suboptimal minimum free energy structures for a given RNA sequence of length n and knowing that the free energy of the correct folding is probable to be less than a given value away from the computed expected value for this length n, then we obviously know where we have to start our search for the correct solution.
Hence, we now aim at determining open intervals $I_{\mathrm{sto},n}(k)$ and $\widehat{I}_{\mathrm{sto},n}(k)$ which contain the free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first and second model, respectively, with high probability (i.e. with probability larger than $1 - \frac{1}{k^2}$).

6.1 Computing Intervals $I_{\mathbf{sto},n}(k)$ for the First Model

First, we want to derive the desired results under the assumption of our first model. Considering all bar-bracket words $s \in \mathcal{S}_n$, then according to Chebyshev's inequality,

$$Pr[|g_{\mathrm{sto}}(s) - \mu_{\mathrm{sto},n}| \geq k\sigma_{\mathrm{sto},n}] \leq \frac{1}{k^2}.$$

This means that the probability that the free energy change $g_{\mathrm{sto}}(s)$ associated with a bar-bracket word $s \in \mathcal{S}_n$, i.e. the free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first model, lies not in the open interval

$$(\mu_{\mathrm{sto},n} - k\sigma_{\mathrm{sto},n}, \mu_{\mathrm{sto},n} + k\sigma_{\mathrm{sto},n})$$

is less than or equal to $\frac{1}{k^2}$. Hence, the probability that the free energy change $g_{\mathrm{sto}}(s)$ associated with a bar-bracket word $s \in \mathcal{S}_n$ lies in this interval is greater than $\left(1 - \frac{1}{k^2}\right)$, as for $s \in \mathcal{S}_n$,

$$\Pr[|g_{\mathrm{sto}}(s) - \mu_{\mathrm{sto},n}| < k\sigma_{\mathrm{sto},n}] = 1 - \Pr[|g_{\mathrm{sto}}(s) - \mu_{\mathrm{sto},n}| \geq k\sigma_{\mathrm{sto},n}] > 1 - \frac{1}{k^2}.$$

Thus, considering all $s \in \mathcal{S}_n$, we may assume that at most $\frac{100}{k^2}$ percent of the free energy values $g_{\mathrm{sto}}(s)$ lie not in and at least $\left(100 - \frac{100}{k^2}\right)$ percent of them lie in the interval

$$(\mu_{\mathrm{sto},n} - k\sigma_{\mathrm{sto},n}, \mu_{\mathrm{sto},n} + k\sigma_{\mathrm{sto},n}),$$

respectively.
According to Theorem 5.3 and Theorem 5.5, $\mu_{\text{sto},n}$ and $\sigma_{\text{sto},n}$ are asymptotically given by $(39.16513746 - 0.2478300708n)$ $^{\text{kcal}}/_{\text{mol}}$ and $1.5910668430\sqrt{n}$ $^{\text{kcal}}/_{\text{mol}}$, respectively, as $n \to \infty$. Thus, under the assumption of our first model, we can suppose that at most $\frac{100}{k^2}$ percent of free energy changes $\Delta G^\circ_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n do not lie and that at least $\left(100 - \frac{100}{k^2}\right)$ percent of them do lie in the open interval

$$I_{\text{sto},n}(k) := (a_{\text{sto},n}(k), b_{\text{sto},n}(k)),$$

where

$$a_{\text{sto},n}(k) := (39.16513746 - 1.5910668430k\sqrt{n} - 0.2478300708n) \ ^{\text{kcal}}/_{\text{mol}} \text{ and}$$
$$b_{\text{sto},n}(k) := (39.16513746 + 1.5910668430k\sqrt{n} - 0.2478300708n) \ ^{\text{kcal}}/_{\text{mol}}.$$

Note that this fact must only hold for $n \to \infty$, as the asymptotical representations for $\mu_{\text{sto},n}$ and $\sigma_{\text{sto},n}$ must only hold for $n \to \infty$, according to Darboux's theorem. Furthermore, it has to be mentioned that Chebyshev's inequality does only provide useful information for values of k that are greater than 1, since for $k \leq 1$, the value of $\frac{1}{k^2}$ would be greater than 1 and probabilities must lie in the closed interval $[0, 1]$. Figure 3 shows a plot of the interval $I_{\text{sto},n}(2)$, $I_{\text{sto},n}(\sqrt{10})$, $I_{\text{sto},n}(\sqrt{20})$ and $I_{\text{sto},n}(10)$, respectively. Consequently, the number of free energy changes $\Delta G^\circ_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first model that lie in these intervals are at least 75 percent, at least 90 percent, at least 95 percent and at least 99 percent of all these free energy changes, respectively. In other words, the probability that the free energy change $\Delta G^\circ_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first model lies in the interval $I_{\text{sto},n}(2)$, $I_{\text{sto},n}(\sqrt{10})$, $I_{\text{sto},n}(\sqrt{20})$ and $I_{\text{sto},n}(10)$ is greater than 0.75, 0.9, 0.95 and 0.99, respectively. It should be no surprise that the length of any interval

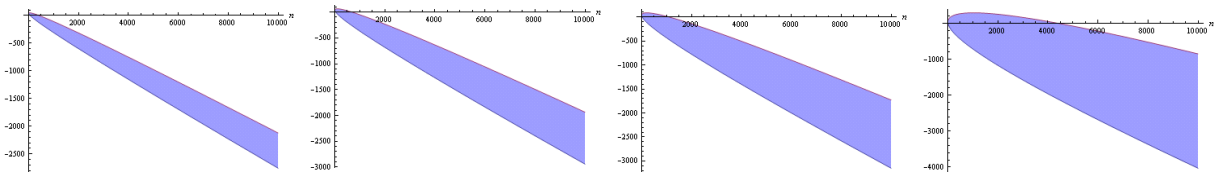

Figure 3: Plots of the intervals $I_{\text{sto},n}(k)$, $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$.

$I_{\text{sto},n}(k)$, $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$, grows with increasing value of n. Furthermore, it should be easy to understand why, for a fixed value of n, the length of the intervals $I_{\text{sto},n}(k)$, $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$, grows with increasing k. The fact that the length of the intervals $I_{\text{sto},n}(k)$, for $k > 1$ and $n > 0$, grows with increasing values of both k and n is illustrated by the three-dimensional plots shown in Figure 4.

6.2 Computing Intervals $\widehat{I}_{\text{sto},n}(k)$ for the Second Model

Equally, under the assumption of our second model, we find out that at most $\frac{100}{k^2}$ percent of free energy changes $\Delta G^\circ_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n lie not and at least $\left(100 - \frac{100}{k^2}\right)$ percent of them do lie in the open interval

$$\widehat{I}_{\text{sto},n}(k) := \left(\widehat{a}_{\text{sto},n}(k), \widehat{b}_{\text{sto},n}(k)\right),$$

where

$$\widehat{a}_{\text{sto},n}(k) := (37.10857372 - 1.9908422758k\sqrt{n} - 0.1841895371n) \ ^{\text{kcal}}/_{\text{mol}} \text{ and}$$
$$\widehat{b}_{\text{sto},n}(k) := (37.10857372 + 1.9908422758k\sqrt{n} - 0.1841895371n) \ ^{\text{kcal}}/_{\text{mol}},$$

$n \to \infty$, according to Theorem 5.8 and Theorem 5.9. Plots of the intervals $\widehat{I}_{\text{sto},n}(2)$, $\widehat{I}_{\text{sto},n}(\sqrt{10})$, $\widehat{I}_{\text{sto},n}(\sqrt{20})$ and $\widehat{I}_{\text{sto},n}(10)$, respectively, are shown in Figure 5 and three-dimensional plots of the endpoints of the open intervals $\widehat{I}_{\text{sto},n}(k)$ for each possible combination of $k \in [\sqrt{2}, 10]$ and $n \in [1, 10000]$, respectively, are shown in Figure 6.

6.3 Discussion

First, comparing Figure 3 to Figure 5, it is easy to see that for fixed values of both n and k, the length of the interval $\widehat{I}_{\text{sto},n}(k)$ is always greater than the length of the corresponding interval $I_{\text{sto},n}(k)$. This is

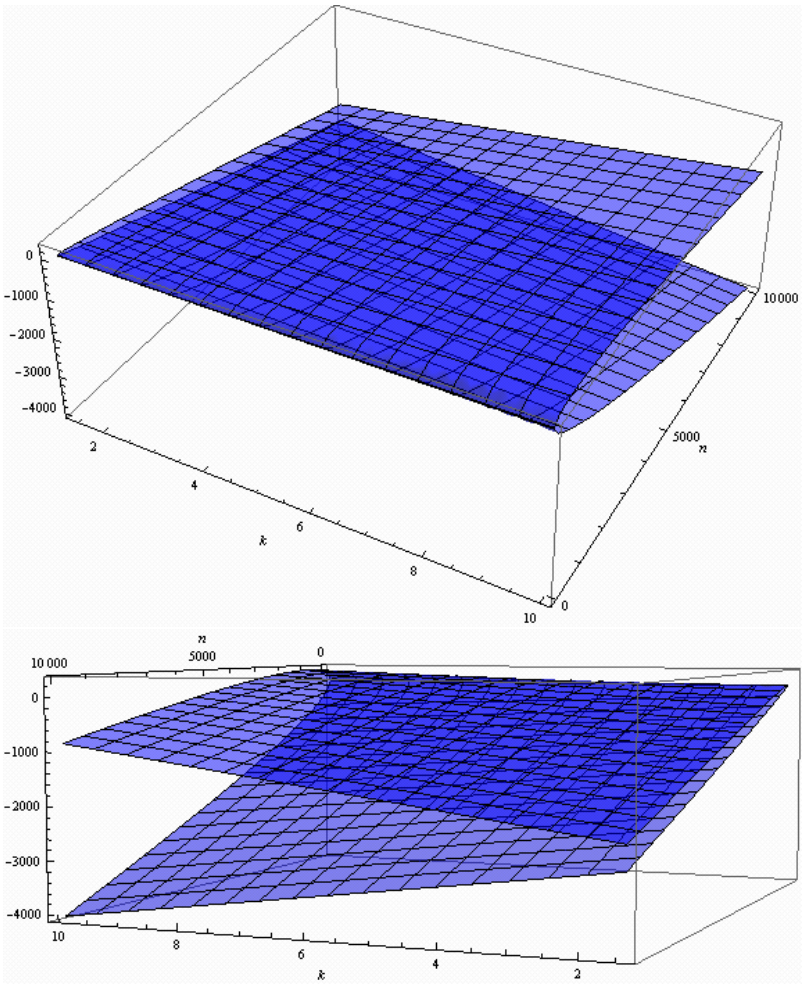

Figure 4: The two endpoints $a_{\mathrm{sto},n}(k)$ and $b_{\mathrm{sto},n}(k)$ of the open interval $I_{\mathrm{sto},n}(k)$, plotted as functions in both k and n, for $\sqrt{2} \leq k \leq 10$ and $1 \leq n \leq 10,000$, respectively. Both three-dimensional plots contain exactly the same information, but they are shown from different points of view.

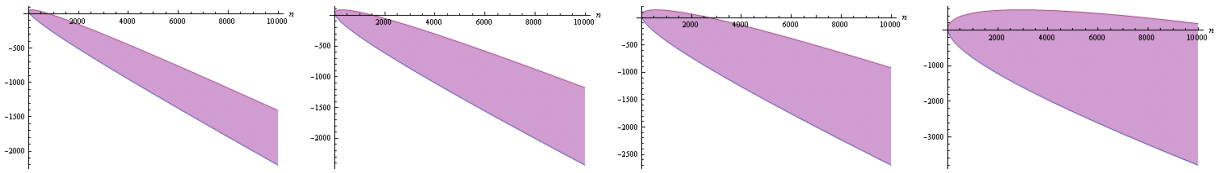

Figure 5: Plots of the intervals $\widehat{I}_{\mathrm{sto},n}(k)$, $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$.

obviously due to the fact that the variance for our second model is always greater than that for our first model.

Moreover, considering Figure 7, we see that for fixed values of both n and k, either the interval $I_{\mathrm{sto},n}(k)$ is completely contained in the interval $\widehat{I}_{\mathrm{sto},n}(k)$ (i.e. $I_{\mathrm{sto},n}(k) \subset \widehat{I}_{\mathrm{sto},n}(k)$ holds) or the intervals $I_{\mathrm{sto},n}(k)$ and $\widehat{I}_{\mathrm{sto},n}(k)$ are disjoint (i.e. $I_{\mathrm{sto},n}(k) \cap \widehat{I}_{\mathrm{sto},n}(k) = \emptyset$ holds) or they are only partially different (i.e. $I_{\mathrm{sto},n}(k) \cap \widehat{I}_{\mathrm{sto},n}(k) \neq \emptyset$ and $0 < |I_{\mathrm{sto},n}(k) \cap \widehat{I}_{\mathrm{sto},n}(k)| < |I_{\mathrm{sto},n}(k)| < |\widehat{I}_{\mathrm{sto},n}(k)|$ holds). Now, we want to consider Figure 8 and Figure 9, where the intervals $I_{\mathrm{sto},n}(k)$ and $\widehat{I}_{\mathrm{sto},n}(k)$ are shown for $k \in \{\sqrt{20}, 10\}$, respectively, all together with the 1866 points corresponding to the free energy changes of the RNA secondary structures given in our biological database. As we can see, for $k = \sqrt{20}$, not all the free energy changes of the RNA secondary structures $\mathbf{S} \neq \emptyset$ of size n given in our database lie in the intervals $I_{\mathrm{sto},n}(k)$ and $\widehat{I}_{\mathrm{sto},n}(k)$, respectively. But for $k = 10$, they do in fact lie in the intervals $I_{\mathrm{sto},n}(k)$ and $\widehat{I}_{\mathrm{sto},n}(k)$, respectively.

Thus, if we had to search the set of all predicted suboptimal minimum free energy structures for an RNA primary structure of length n which is given in our database for the native solution, then we would with

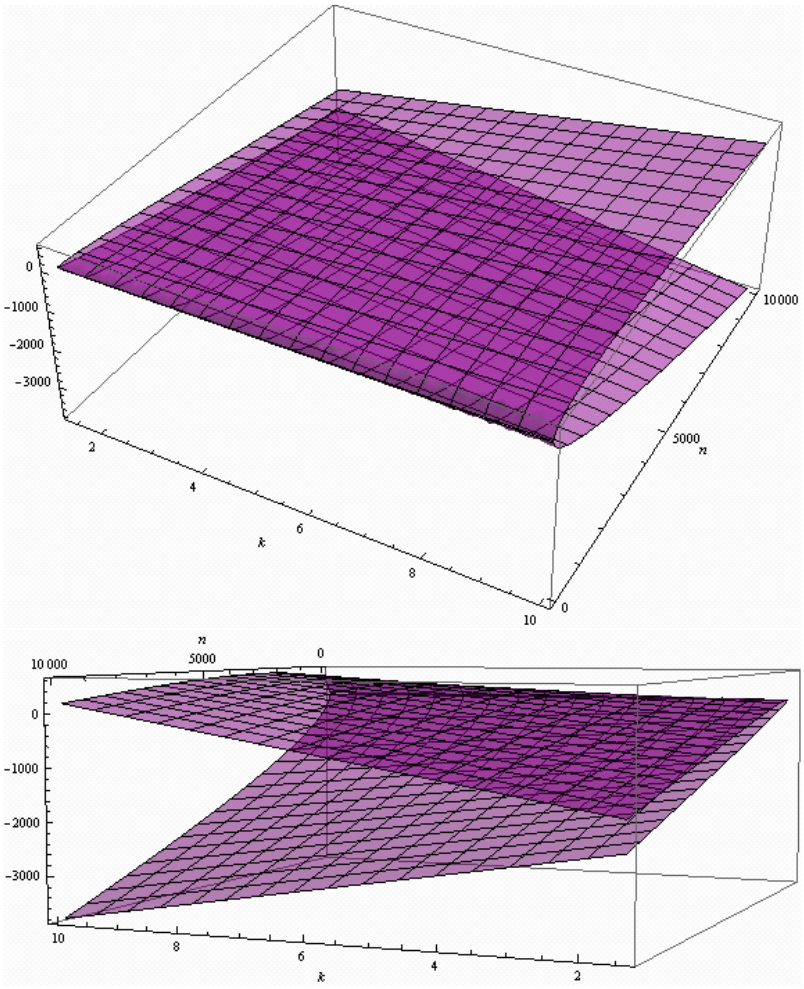

Figure 6: The two endpoints $\widehat{a}_{\mathrm{sto},n}(k)$ and $\widehat{b}_{\mathrm{sto},n}(k)$ of the open interval $\widehat{I}_{\mathrm{sto},n}(k)$, plotted as functions in both k and n, for $\sqrt{2} \leq k \leq 10$ and $1 \leq n \leq 10,000$, respectively. Both three-dimensional plots contain exactly the same information, but they are shown from different points of view.

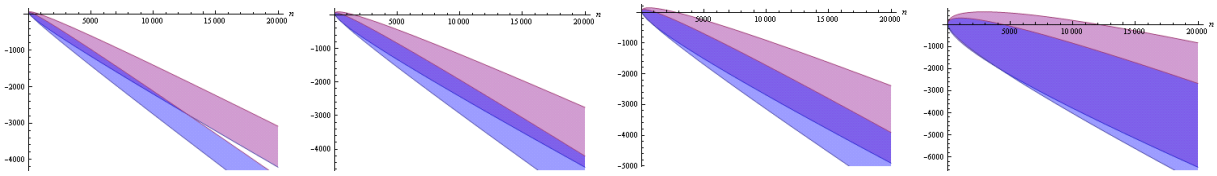

Figure 7: Plots of the intervals $I_{\mathrm{sto},n}(k)$ (blue) and $\widehat{I}_{\mathrm{sto},n}(k)$ (purple), $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$.

high probability find the correct folding by considering only the subset of predicted suboptimal minimum free energy structures that have a free energy change which lies in the open interval $I_{\mathrm{sto},n}(10)$ or $\widehat{I}_{\mathrm{sto},n}(10)$, depending on whether the first or second model is considered.
But if we searched only the subset of predicted suboptimal minimum free energy structures that have a free energy which lies in the open interval $I_{\mathrm{sto},n}(\sqrt{20})$ resp. $\widehat{I}_{\mathrm{sto},n}(\sqrt{20})$, we would possibly not find the correct solution (in case that the considered interval and hence the searched subset is too small). However, it is obviously possible that the free energy change of the correct folding lies in the interval $I_{\mathrm{sto},n}(\sqrt{20})$, but not in $\widehat{I}_{\mathrm{sto},n}(\sqrt{20})$, or the other way round (see Figure 10).

Hence, considering the first model and imagining the case that the free energy change of the correct solution lies in the interval $I_{\mathrm{sto},n}(\sqrt{20})$, then we obviously do not need the search any subset of predicted suboptimal minimum free energy structures that have a stability which lies in $I_{\mathrm{sto},n}(k)$, $k > \sqrt{20}$ to find the native folding. In fact, in this case, it suffices to search for the right solution in the smaller subset of predicted suboptimal minimum free energy structures having a free energy change in $I_{\mathrm{sto},n}(\sqrt{20})$, and perhaps we could even find the right solution considering a shorter interval $I_{\mathrm{sto},n}(k)$, $k < \sqrt{20}$.

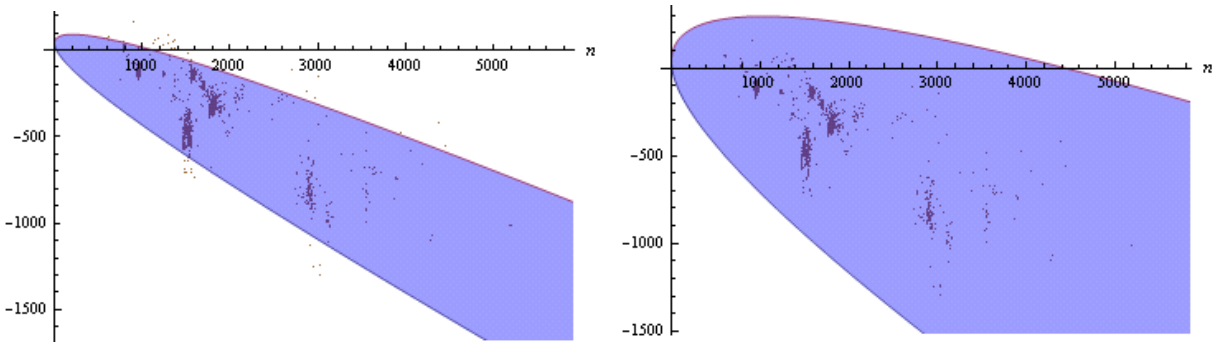

Figure 8: Plots of the intervals $I_{\mathrm{sto},n}(\sqrt{20})$ (left) and $I_{\mathrm{sto},n}(10)$ (right) containing at least 95 percent and at least 99 percent of the free energy changes $\Delta G^{\circ}_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first model, respectively, together with the 1866 points $\{n, \Delta G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database.

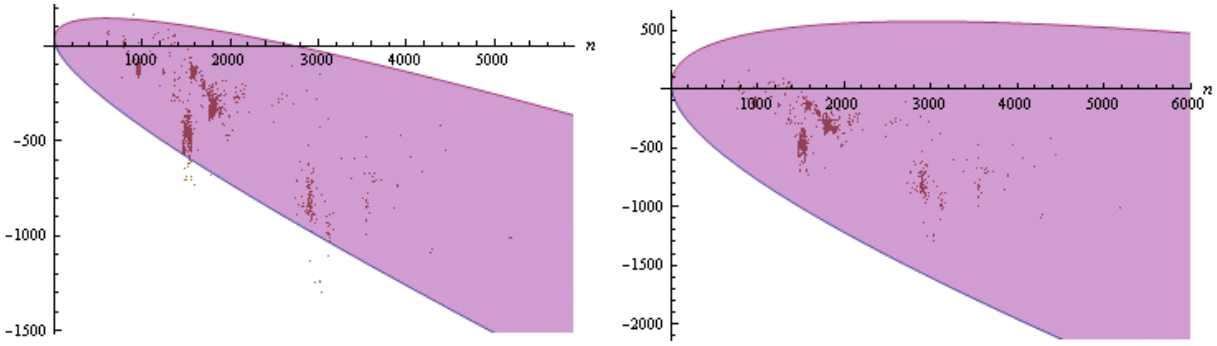

Figure 9: Plots of the intervals $\widehat{I}_{\mathrm{sto},n}(\sqrt{20})$ (left) and $\widehat{I}_{\mathrm{sto},n}(10)$ (right) containing at least 95 percent and at least 99 percent of the free energy changes $\Delta G^{\circ}_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our second model, respectively, together with the 1866 points $\{n, \Delta G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database.

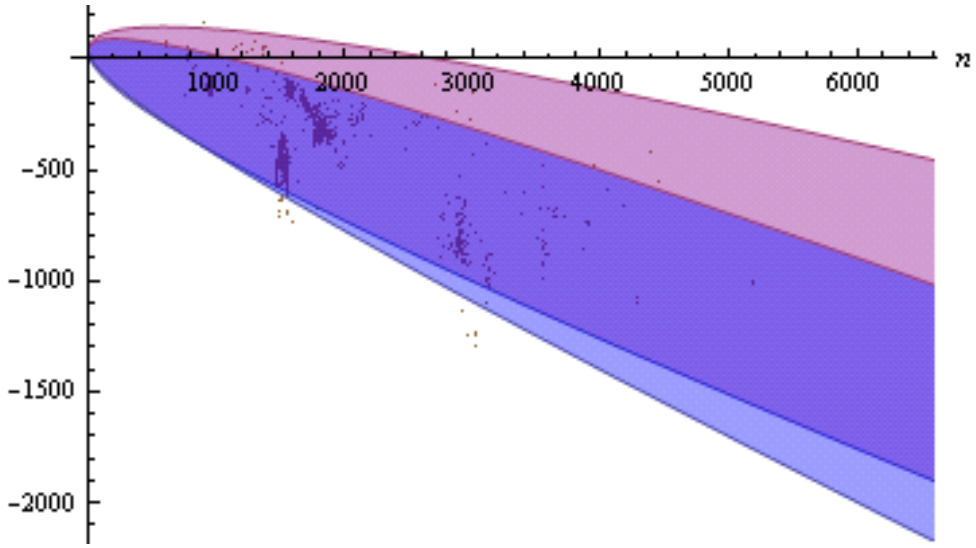

Figure 10: Plots of the intervals $I_{\mathrm{sto},n}(\sqrt{20})$ (blue) and $\widehat{I}_{\mathrm{sto},n}(\sqrt{20})$ (purple) together with the 1866 points $\{n, \Delta G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database.

Consequently, suppose we want to find the correct secondary structure $\mathbf{S} \neq \emptyset$ for any SSU or LSU rRNA sequence $\mathbf{R}$ of length n in the set of all predicted suboptimal minimum free energy structures for $\mathbf{R}$. Then, we can start by choosing a small value k_0 and search in the subset of predicted structures that have a free energy change in $I_{\mathrm{sto},n}(k_0)$ or in $\widehat{I}_{\mathrm{sto},n}(k_0)$, depending on whether we want to consider the first or second free energy model. Obviously, this way, we have to search a smaller set of predicted structures

to find the desired folding. But it is not guaranteed that the correct folding is actually contained in this subset. However, if the correct folding can not be found by considering the open interval $I_{\text{sto},n}(k_0)$ resp. $\widehat{I}_{\text{sto},n}(k_0)$, then we can choose a value $k_1 > k_0$ and search in the subset of all predicted structures that have a free energy in $I_{\text{sto},n}(k_1) \setminus I_{\text{sto},n}(k_0)$ resp. $\widehat{I}_{\text{sto},n}(k_1) \setminus \widehat{I}_{\text{sto},n}(k_0)$. Obviously, if the correct folding can still not be found by considering this new subset, we can choose a value $k_2 > k_1$ and so on, until we eventually are successful.
Note that the results that we have derived in this work under the assumption of our first and second stochastic model derived from SSU and LSU rRNA secondary structures, respectively, could be improved by using a more comprehensive database of SSU and LSU rRNA secondary structures $\mathbf{S} \neq \emptyset$.
Finally, it remains to mention that this approach for identifying good predictions of RNA secondary structure by considering the interval $I_{\text{sto},n}(k)$ resp. $\widehat{I}_{\text{sto},n}(k)$ should only be used for SSU and LSU rRNA sequences. However, by considering a database of known RNA secondary structures $\mathbf{S} \neq \emptyset$ for other types of RNA, we can determine two corresponding intervals that could be used in the described way to help on identifying good predictions of secondary structure for the respective type of RNA. In fact, the corresponding results for secondary structures of any other type of RNA could be determined in the same way as done in this work for SSU and LSU rRNA secondary structures.

7 Conclusions

In this article, we have studied a stochastic model for RNA secondary structures which was derived from a database of SSU and LSU rRNA secondary structures. This database was constructed from the databases given in [WRdP$^+$01] and [WdPWW02]. More precisely, we have worked out two different free energy models for this stochastic model for RNA secondary structures. These free energy models are based on the well-known the INN-HB model with loop-dependent energy rules [XSB$^+$98, MSZT99], where the considered thermodynamic parameters were those given in [MSZT99], which have also been used for version 3.0 of the MFOLD software [Zuk03]. For both models, we have computed asymptotics for the expected free energy change $\Delta G^{\circ}_{37}(\mathbf{S})$ as well as for the corresponding variance of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n. To obtain our results, we have used stochastic context-free grammars and languages, generating functions and Darboux's theorem. In fact, in this article, we have analytically analyzed the free energy in an RNA secondary structure model, which has so far never been done by other authors. As both models have turned out to be realistic, we have finally used Chebyshev's inequality to compute probabilities that the free energy of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is less than a given value away from the computed expected free energy of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first and second model, respectively. More precisely, we have determined two open intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ for $n > 0$ and $k > 1$, which contain at least $\left(100 - \frac{100}{k^2}\right)$ percent of the free energy changes of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our first and second model, respectively. As we have pointed out, the length of both intervals grows with increasing values of n and k. At the end of our investigations, we described an approach on how the consideration of any of these two intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ can help us to identify good predictions of (SSU and LSU r)RNA secondary structure.
Obviously, the usefulness of the described approach, more precisely of the intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$, should be tested in the near future. Moreover, such tests could obviously help us to figure out which one of the two intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$, for a given n and k, is more realistic. Testing the usefulness of the intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ may also result in the detection of some weaknesses of both or any of the two underlying free energy models. In fact, if the intervals $I_{\text{sto},n}(k)$ or the intervals $\widehat{I}_{\text{sto},n}(k)$, for all $n > 0$ and $k > 1$, turn out to be not useful, then there are possibly some weaknesses in the corresponding free energy model that we worked out for our stochastic model for RNA secondary structures. In this case, there is obviously a need to find some improvements for the corresponding free energy model in order to eliminate these weaknesses. Such improvements can also have a positive effect on the quality of the RNA secondary structure prediction methods, which means that they could possibly also be used to improve dynamic programming algorithms for the prediction of RNA secondary structure. As we have already mentioned, the intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ should only be used for identifying good predictions of SSU and LSU rRNA secondary structure in the previously described way. In fact, it would be interesting to compare the corresponding intervals for different types of RNA to see if they are almost equal or completely different. Finally, recall that if we want to find the correct secondary structure $\mathbf{S} \neq \emptyset$ of size n for a given RNA sequence $\mathbf{R}$ of a certain type of RNA and of length n, then we only need to consider the corresponding intervals for this length n and different values of k. Hence, under the assumption that the described approach proves to be useful, it would be another task to find an algorithm that takes

a database of known RNA secondary structures of a certain type of RNA as input and computes the corresponding intervals (or at least one of them) for a given length n efficiently.

References

[AJL+02] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Publishing, fourth edition, 2002.

[AvdBvBP90] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. W. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Res.*, 18(10):3035–3044, 1990.

[BDTU74] P. N. Borer, B. Dengler, I. Tinoco Jr., and O. C. Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*, 86:843–853, 1974.

[CB00] Peter Clote and Rolf Backofen. *Computational Molecular Biology: An Introduction*. John Wiley and Sons, 2000.

[CG98] T. Chi and S. Geman. Estimation of probabilistic context-free grammars. *Computational Linguistics*, 24(2):299–305, 1998.

[Com74] Louis Comtet. *Advanced Combinatorics; The art of finite and infinite expansions*. Reidel Publ. Co., Dordrecht, rev. and enl. edition, 1974.

[CS63] N. Chomsky and M. P. Schützenberger. The algebraic theorey of context-free languages. In P. Braffort and D. Hirschberg, editors, *Computer Programming and Formal Systems*, pages 118–161. North-Holland, Amsterdam, 1963.

[DPD92] E. Dam, K. Pleij, and D. Draper. Structural and functional aspects of RNA pseudoknots. *Biochemistry*, 31:11665–11676, 1992.

[Edd04] Sean R. Eddy. How do RNA folding algorithms work. *Nature Biotechnology*, 22(11):1457–1458, 2004.

[FS07] Philippe Flajolet and Robert Sedgewick. Analytic combinatorics. Preliminary version, March 2007.

[GC73] J. Gralla and D. M. Crothers. Free energy of imperfect nucleic acid helices : II. small hairpin loops. *Journal of Molecular Biology*, 73:497–511, 1973.

[GK90] Daniel H. Greene and Donald E. Knuth. *Mathematics for the Analysis of Algorithms*. Birkhäuser Boston, third edition, 1990.

[GW90] R. R. Gutell and C. R. Woese. Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. USA*, 87:663–667, 1990.

[Har78] Michael A. Harrison. *Introduction to Formal Language Theory*. Addison-Wesley, 1978.

[HF71] T. Huang and K. S. Fu. On stochastic context-free languages. *Information Sciences*, 3:201–224, 1971.

[HMU01] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 2nd edition, 2001.

[Hof95] Micha Hofri. *Analysis of Algorithms: Computational Methods and Mathematical Tools*. Oxford University Press, 1995.

[HSS98] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88:207–237, 1998.

[JTZ89] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[JTZ90] J. A. Jaeger, D. H. Turner, and M. Zuker. Predicting optimal and suboptimal secondary structure for RNA. In R. F. Doolittle, editor, *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, volume 183 of *Methods in Enzymology*, pages 281–306. Academic Pr., San Diego, 1990.

[KH99] B. Knudsen and J. Hein. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–454, 1999.

[KH03] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Research*, 31(13):3423–3428, 2003.

[KW89] Donald E. Knuth and Herbert S. Wilf. A short proof of darboux's lemma. *Applied Mathematics Letters*, 2:139–140, 1989.

[Les74] A. M. Lesk. A combinatorial study of the effects of admitting non-Watson- Crick base pairings and of base compositions on the helix-forming potential of polynucleotides of random sequences. *J. Theor. Biol.*, 44:7–17, 1974.

[MN08] Dirk Metzler and Markus E. Nebel. Predicting RNA secondary structures with pseudoknots by MCMC sampling. *Journal of Mathematical Biology*, 56:161–181, 2008.

[MSZT99] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[Neb02a] Markus E. Nebel. Combinatorial properties of RNA secondary structures. *Journal of Computational Biology*, 9(3):541–574, 2002.

[Neb02b] Markus E. Nebel. On a statistical filter for RNA secondary structures. Technical report, Frankfurter Informatik-Berichte, 5 2002.

[Neb04a] Markus E. Nebel. Identifying good predictions of RNA secondary structure. *Proceedings of the Pacific Symposium on Biocomputing*, pages 423–434, 2004.

[Neb04b] Markus E. Nebel. Investigation of the Bernoulli-model of RNA secondary structures. *Bulletin of Mathematical Biology*, 66:925–964, 2004.

[NJ80] R. Nussinov and A. B. Jacobson. Fast algorithms for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Science of the USA*, 77(11):6309–6313, 1980.

[NPGK78] R. Nussinov, G. Pieczenik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35:68–82, 1978.

[PB89] C. W. Pleij and L. Bosch. RNA pseudoknots: structure, detection, and prediction. *Methods Enzymol.*, 180:289–303, 1989.

[Ple94] C. W. Pleij. RNA pseudoknots. *Curr. Opin. Struct. Biol.*, 4:337–344, 1994.

[RE99] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[SBH+94] Y. Sakakibara, M. Brown, R. Hughey, I. S. Mian, K. Sjölander, R. C. Underwood, , and D. Haussler. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Research*, 22:5112–5120, 1994.

[SF01] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms*. Addison–Wesley Publishing Company, Inc., 2nd edition, September 2001.

[SKMC83] D. Sankoff, J. B. Kruskal, S. Mainville, and R. J. Cedergren. Fast algorithms to determine RNA secondary structures containing multiple loops. In *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*, chapter 3, pages 93–120. Addison-Wesley, Reading, MA, 1983.

[ST95] M. J. Serra and D. H. Turner. Predicting thermodynamic properties of RNA. *Methods in Enzymology*, 259:242–261, 1995.

[SW78] P. R. Stein and M. S. Waterman. On some new sequences generalizing the catalan and motzkin numbers. *Discrete Mathematics*, 26:216–272, 1978.

[TUL71] I. Tinoco, O. Uhlenbeck, and M. D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230:362–367, 1971.

[VC85] G. Viennot and M. Vauchaussade De Chaumont. Enumeration of RNA secondary structures by complexity. *Mathematics in medicine and biology, Lecture Notes in Biomathematics*, 57:360–365, 1985.

[Wat78] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Advances in Mathematics Supplementary Studies*, 1:167–212, 1978.

[WdPWW02] Jan Wuyts, Yves Van de Peer, Tina Winkelmans, and Rupert De Wachter. The european database on small subunit ribosomal RNA. *Nucleic Acids Research*, 30(1):183–185, 2002.

[Wil94] Herbert S. Wilf. *generatingfunctionology*. Academic Press, Inc., second edition, 1994.

[WRdP+01] Jan Wuyts, Peter De Rijk, Yves Van de Peer, Tina Winkelmans, and Rupert De Wachter. The european large subunit ribosomal RNA database. *Nucleic Acids Research*, 29(1):175–177, 2001.

[XSB+98] T. Xia, J. SantaLucia Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with watson-crick base pairs. *Biochemistry*, 37:14719–14735, 1998.

[ZMT99] M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In J. Barciszewski and B. F. C. Clark, editors, *RNA Biochemistry and Biotechnology*, NATO ASI Series, pages 11–43. Kluwer Academic Publishers, Dordrecht, NL, 1999.

[ZS81] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.

[ZS84] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Mathematical Biology*, 46:591–621, 1984.

[Zuk86] M. Zuker. RNA folding prediction: The continued need for interaction between biologists and mathematicians. *Lectures on Mathematics in the Life Sciences*, 17:87–124, 1986.

[Zuk89a] M. Zuker. Computer prediction of RNA structure. In J. E. Dahlberg and J. N. Abelson, editors, *RNA Processing*, volume 180 of *Methods in Enzymology*, pages 262–288. Acad. Pr., San Diego, 1989.

[Zuk89b] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[Zuk94] M. Zuker. Prediction of RNA secondary structure by energy minimization. In A. M. Griffin and H. G. Griffin, editors, *Computer Analysis of Sequence Data*, Methods in Molecular Biology, pages 267–294. Humana Press Inc., 1994.

[Zuk03] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

A Generating Functions

In this section, we will recall some fundamental definitions and methods concerning *generating functions*. The basic definitions are given as follows:

Definition A.1 ([FS07]) *A* combinatorial class*, or simply a* class*, is a finite or denumerable set on which a* size *function is defined, satisfying the following conditions:*

1. *the size of an element is a nonnegative integer;*

2. the number of elements of any given size is finite.

In the sequel, we will use the same notations as in [FS07]. This means that if $\mathcal{A}$ is a class, the size of an element $a \in \mathcal{A}$ is denoted by $|a|$ and given a class $\mathcal{A}$, we consistently let $\mathcal{A}_n$ be the set of objects in $\mathcal{A}$ having size n. Furthermore, we use the same group of letters for the counts $a_n = \text{card}(\mathcal{A}_n)$.

Definition A.2 ([FS07]) *The* counting sequence *of a combinatorial class is the sequence of integers* $(a_n)_{n\geq 0}$ *where* $a_n = \text{card}(\mathcal{A}_n)$ *is the number of objects in class* $\mathcal{A}$ *that have size* n.

Definition A.3 ([FS07]) *The* ordinary generating function *(OGF) of a sequence* $(a_n)_{n\geq 0}$ *is the formal power series*

$$A(z) = \sum_{n=0}^{\infty} a_n z^n.$$

The ordinary generating function *(OGF) of a combinatorial class* $\mathcal{A}$ *is the generating function of the numbers* $a_n = \text{card}(\mathcal{A}_n)$. *Equivalently, the OGF of class* $\mathcal{A}$ *admits the combinatorial form*

$$A(z) = \sum_{a\in\mathcal{A}} z^{|a|}.$$

It is also said that the variable z *marks size in the generating function.*
By $[z^n]A(z)$, *we denote the operation of extracting the coefficient of* z^n *in the formal power series* $A(z) = \sum a_n z^n$, *so that*

$$[z^n]\left(\sum_{n\geq 0} A_n z^n\right) = a_n.$$

(The operator $[z^n]$ *applied to* $f(z)$ *reads as "coefficient of* z^n *in* $f(z)$*".)*

In general, it is not easy to determine the nth coefficient of an ordinary generating function $A(z)$ for a combinatorial class $\mathcal{A}$ of objects. In fact, we usually have to compute a *closed form* of the desired generating function $A(z)$ first and then, we can use this closed form to compute an asymptotic or an exact representation for the nth coefficient $a_n = [z^n]A(z)$ of this generating function $A(z)$.

A.1 Computing Generating Functions

A common way to compute such a closed form of a generating function $A(z)$ is to model the combinatorial class $\mathcal{A}$ of objects as context-free language $\mathcal{L}_A$ containing exactly all the (encodings of the) elements in $\mathcal{A}$. Then, we can construct an unambiguous context-free grammar $G_A = (I_A, \Sigma_A, R_A, S)$ which generates exactly the language $\mathcal{L}_A$. Afterwards, we can translate this grammar G_A into a system of equations, as proposed by Schützenberger [CS63], in order to construct a generating function.
Note that in this article, we assume that the reader has basic knowledge of the notions concerning context-free languages and grammars. An introduction could be found, for example, in [HMU01] or [Har78].
It should be mentioned that translating the grammar G_A into a system of equations means that the production rules contained in the rule set R_A of the grammar G_A are translated into a system of equations. This system then has to be solved for the variable S corresponding to the start symbol (axiom) of the grammar G_A to obtain the desired closed form. More precisely, we first have to eliminate each variable X corresponding to the symbol $X \in I_A \setminus \{S\}$ in this system of equations to obtain a polynomial equation in the variables z and S only and this polynomial equation must then be solved for the variable S. Note that there is a difference between approximating solutions to polynomial equations and finding exact solutions. In fact, for polynomial equations up to a degree of 4, we can compute exact solutions. But for polynomial equations of degree 5 or greater, we can only compute approximate solutions.

A.2 Computing Coefficient Asymptotics

To compute an asymptotic for the nth coefficient of a generating function $A(z)$ (for $n \to \infty$), we can use the methods of *singularity analysis*. To be able to use this method, we now want to recall some definitions and further results. First, it has to be mentioned that in the sequel, we will no longer consider generating functions as formal power series, but as analytic functions that are represented as power series. For details, see for example [FS07]. Then, the functions we consider are defined in certain regions of the complex plane $\mathbb{C}$.

Definition A.4 ([FS07]) *A function $f(z)$ defined over a region $\Omega \subset \mathbb{C}$ is* analytic *at a point $z_0 \in \Omega$ if, for z in some open disk centred at z_0 and contained in Ω, it is representable by a convergent power series expansion*

$$f(z) = \sum_{n \geq 0} c_n (z - z_0)^n.$$

A function is analytic in a region Ω iff it is analytic at every point of Ω.

In addition to the term analytic, we want to introduce the term *regular*. Although these terms have different meanings, in our context we may use them interchangeably.

Definition A.5 ([Hof95]) *If $f(z)$ is analytic and* single-valued *throughout $\Omega \subset \mathbb{C}$ it is said to be* regular *in Ω (or* holomorphic*). The function is* regular at a point *if it is regular in some neighborhood of the point. Such a point is called a* regular point *of $f(z)$. A point which is not regular is* singular.

Singular points are often called singularities and they are essential to coefficient asymptotics. There are different types of singularities:

Definition A.6 (Classification of Singularities [Hof95]) *If z_0 is a singular point of $f(z)$, and the function is regular in a "punctured disk" $0 < |z - z_0| < R \leq \infty$, we say it has an* isolated singularity*. An isolated singularity can be of the following types:*

- removable singularity*, when $\lim_{z \to z_0} f(z)$ exists.*
- pole*, in case $\lim_{z \to z_0} f(z) = \infty$ holds (we say it exists as an improper limit).*
- essential singularity*, when $\lim_{z \to z_0} f(z)$ does not exist, not even improperly.*

A branch point *is a point where branches of a multivalued function coincide (called by some authors, when removable,* weak singularity*).*
An algebraic singularity *is either a pole or a branch point.*

We are only interested in a subset of all the singularites of a generating function, called *dominant singularities*.

Definition A.7 ([FS07]) *For any function $f(z)$ that is analytic at a point z_0, the disk with the property that the series expansion about the point z_0 representing $f(z)$ is convergent for z inside the disk and divergent for z outside the disk is called the* disk of convergence *and its radius is the* radius of convergence *of $f(z)$ at $z = z_0$.*
Singularities of a function $f(z)$ analytic at $z_0 = 0$ which lie on the boundary of the disk of convergence of $f(z)$ at $z_0 = 0$ are called dominant singularities*.*

Theorem A.1 (Boundary singularities [FS07]) *A function $f(z)$ analytic at the origin, whose expansion at the origin has a finite radius of convergence R, necessarily has a singularity on the boundary of its disk of convergence, $|z| = R$.*

The following theorem can help us to determine the dominant singularities of a given generating function.

Theorem A.2 (Pringsheim's Theorem [FS07]) *If $f(z)$ is representable at the origin by a series expansion that has nonnegative coefficients and radius of convergence R, then the point $z = R$ is a singularity of $f(z)$.*

In this work, we will use the following theorem to compute an asymptotical representation for the nth coefficient of a given generating function (for $n \to \infty$):

Theorem A.3 (DARBOUX [KW89]) *Let $v(z)$ be analytic in some disk $|z| < 1 + \eta$, and suppose that in a neighborhood of $z = 1$ it has the expansion $v(z) = \sum v_j (1 - z)^j$. Then for every β and every integer $m \geq 0$ we have*

$$\begin{aligned} [z^n]\{(1-z)^\beta v(z)\} &= [z^n] \left\{ \sum_{j=0}^{m} v_j (1-z)^{\beta+j} \right\} + \mathcal{O}(n^{-m-\beta-2}) \\ &= \sum_{j=0}^{m} v_j \binom{n-\beta-j-1}{n} + \mathcal{O}(n^{-m-\beta-2}), \end{aligned}$$

as $n \to \infty$.

Note that the larger we choose the parameter m for the determination of a coefficient asymptotic according to Darboux's theorem, the more exact the resulting coefficient asymptotic gets. In fact, by choosing $m \to \infty$, the resulting coefficient asymptotic is equal to the exact coefficient.

B Tables

Rule f	Probability p_f	Rule f	Probability p_f	Rule f	Probability p_f
f_1	$p_1 := 1$	f_{15}	$p_{15} := \frac{7235}{38399}$	f_{29}	$p_{29} := \frac{4986}{29105}$
f_2	$p_2 := \frac{5543}{6476}$	f_{16}	$p_{16} := \frac{11831}{38399}$	f_{30}	$p_{30} := \frac{24119}{29105}$
f_3	$p_3 := \frac{933}{6476}$	f_{17}	$p_{17} := \frac{7666}{38399}$	f_{31}	$p_{31} := \frac{2357}{5679}$
f_4	$p_4 := \frac{74489}{81898}$	f_{18}	$p_{18} := \frac{7781}{12748}$	f_{32}	$p_{32} := \frac{3322}{5679}$
f_5	$p_5 := \frac{7409}{81898}$	f_{19}	$p_{19} := \frac{4967}{12748}$	f_{33}	$p_{33} := \frac{57179}{84620}$
f_6	$p_6 := 1$	f_{20}	$p_{20} := \frac{3912}{68075}$	f_{34}	$p_{34} := \frac{27441}{84620}$
f_7	$p_7 := \frac{605069}{792975}$	f_{21}	$p_{21} := \frac{23208}{68075}$	f_{35}	$p_{35} := \frac{37994}{53725}$
f_8	$p_8 := \frac{31912}{792975}$	f_{22}	$p_{22} := \frac{8191}{13615}$	f_{36}	$p_{36} := \frac{15731}{53725}$
f_9	$p_9 := \frac{4912}{264325}$	f_{23}	$p_{23} := \frac{32509}{40700}$	f_{37}	$p_{37} := 1$
f_{10}	$p_{10} := \frac{5821}{158595}$	f_{24}	$p_{24} := \frac{8191}{40700}$	f_{38}	$p_{38} := \frac{23211}{55123}$
f_{11}	$p_{11} := \frac{1893}{264325}$	f_{25}	$p_{25} := \frac{533}{4912}$	f_{39}	$p_{39} := \frac{31912}{55123}$
f_{12}	$p_{12} := \frac{2723}{31719}$	f_{26}	$p_{26} := \frac{1053}{4912}$	f_{40}	$p_{40} := \frac{172939}{212588}$
f_{13}	$p_{13} := \frac{38399}{792975}$	f_{27}	$p_{27} := \frac{2963}{14736}$	f_{41}	$p_{41} := \frac{39649}{212588}$
f_{14}	$p_{14} := \frac{11667}{38399}$	f_{28}	$p_{28} := \frac{7015}{14736}$		

Table 1: The probabilities (relative frequencies) for the production rules of the SCFG G_{sto}, obtained by training it using our biological database.

Rule f	Probability p_f	Rule f	Probability p_f	Rule f	Probability p_f
f_1	$p_1 = 1.00000$	f_{15}	$p_{15} = 0.18842$	f_{29}	$p_{29} = 0.17131$
f_2	$p_2 = 0.85593$	f_{16}	$p_{16} = 0.30811$	f_{30}	$p_{30} = 0.82869$
f_3	$p_3 = 0.14407$	f_{17}	$p_{17} = 0.19964$	f_{31}	$p_{31} = 0.41504$
f_4	$p_4 = 0.90953$	f_{18}	$p_{18} = 0.61037$	f_{32}	$p_{32} = 0.58496$
f_5	$p_5 = 0.09047$	f_{19}	$p_{19} = 0.38963$	f_{33}	$p_{33} = 0.67572$
f_6	$p_6 = 1.00000$	f_{20}	$p_{20} = 0.05747$	f_{34}	$p_{34} = 0.32428$
f_7	$p_7 = 0.76304$	f_{21}	$p_{21} = 0.34092$	f_{35}	$p_{35} = 0.70719$
f_8	$p_8 = 0.04024$	f_{22}	$p_{22} = 0.60161$	f_{36}	$p_{36} = 0.29281$
f_9	$p_9 = 0.01858$	f_{23}	$p_{23} = 0.79875$	f_{37}	$p_{37} = 1.00000$
f_{10}	$p_{10} = 0.03670$	f_{24}	$p_{24} = 0.20125$	f_{38}	$p_{38} = 0.42108$
f_{11}	$p_{11} = 0.00716$	f_{25}	$p_{25} = 0.10851$	f_{39}	$p_{39} = 0.57892$
f_{12}	$p_{12} = 0.08585$	f_{26}	$p_{26} = 0.21437$	f_{40}	$p_{40} = 0.81349$
f_{13}	$p_{13} = 0.04843$	f_{27}	$p_{27} = 0.20107$	f_{41}	$p_{41} = 0.18651$
f_{14}	$p_{14} = 0.30383$	f_{28}	$p_{28} = 0.47605$		

Table 2: Floating point approximations of the probabilities (relative frequencies) for the production rules of the SCFG G_{sto} (rounded to five decimal places).

Loop type	Parameter	Floating point value	Rational approximation
Hairpin loops	ldeh	5.81825	$\frac{146777}{25227}$
	tmseh	−1.32252	$-\frac{44266}{33471}$
	GGGLoopBonus	−0.0117962	$-\frac{653}{55357}$
	cHairpinOf3	0.00787522	$\frac{154}{19555}$
	cHairpin	0.000751223	$\frac{31}{41266}$
	termAUpenHL	0.30248	$\frac{1183}{3911}$
	tetra	−1.39906	$-\frac{38596}{27587}$
Stacked pairs	se	−2.14328	$-\frac{57007}{26598}$
Bulge loops	seBulge	−2.15362	$-\frac{82363}{38244}$
	ldeb	3.57223	$\frac{220453}{61713}$
	termAUpenBL	0.240451	$\frac{3582}{14897}$
Interior loops	ile1x1	0.88689	$\frac{62075}{69991}$
	ile2x2	0.858963	$\frac{29197}{33991}$
	ile1x2	3.20486	$\frac{98181}{30635}$
	ldei	2.25941	$\frac{42321}{18731}$
	asym	0.856416	$\frac{9931}{11596}$
	tmsei	−0.0884185	$-\frac{2953}{33398}$
	tbp1xNil	0.339704	$\frac{551}{1622}$
Multiloops	MBLinitiation	4.89098	$\frac{749107}{153161}$
	stackingMulti	−1.10953	$-\frac{24848}{22395}$
	termAUpenML	0.192775	$\frac{7183}{37261}$
Exterior loops	stackingExterior	−1.04144	$-\frac{27470}{26377}$
	termAUpenEL	0.316206	$\frac{8191}{25904}$

Table 3: Expected free energy contributions used in the first free energy model for the stochastic model for RNA secondary structures.

Loop type	Parameter	Floating point value	Rational approximation
Hairpin loops	ldehPerNuc	1.07399	$\frac{31426}{29261}$
	tmseh	−1.32252	$-\frac{44266}{33471}$
	GGGLoopBonus	−0.0117962	$-\frac{653}{55357}$
	cHairpinOf3	0.00787522	$\frac{154}{19555}$
	cHairpinPerNuc	0.000182974	$\frac{4}{21861}$
	termAUpenHL	0.30248	$\frac{1183}{3911}$
	tetra	−1.39906	$-\frac{38596}{27587}$
Stacked pairs	se	−2.14328	$-\frac{57007}{26598}$
Bulge loops	seBulge	−2.15362	$-\frac{82363}{38244}$
	ldebPerNuc	2.78351	$\frac{53179}{19105}$
	termAUpenBL	0.240451	$\frac{3582}{14897}$
Interior loops	ile1x1	0.8869	$\frac{62075}{69991}$
	ile2x2	0.858963	$\frac{29197}{33991}$
	ile1x2	3.20486	$\frac{98181}{30635}$
	ldeiPerNuc	0.30055	$\frac{9788}{32567}$
	asym	0.856416	$\frac{9931}{11596}$
	tmsei	−0.0884185	$-\frac{2953}{33398}$
	tbp1xNil	0.339704	$\frac{551}{1622}$
Multiloops	MBLOffset	3.4	$\frac{17}{5}$
	MBLFreeBasePenalty	0	0
	MBLHelixPenalty	0.4	$\frac{2}{5}$
	stackingMulti	−1.10953	$-\frac{24848}{22395}$
	termAUpenML	0.192775	$\frac{7183}{37261}$
Exterior loops	stackingExterior	−1.04144	$-\frac{27470}{26377}$
	termAUpenEL	0.316206	$\frac{8191}{25904}$

Table 4: Expected free energy contributions used in the second free energy model for the stochastic model for RNA secondary structures.