Analysis of the Free Energy in a Stochastic RNA Secondary Structure Model

Markus E. Nebel, Anika Scheid[†*]
Department of Computer Science, University of Kaiserslautern, Germany
{nebel,a_scheid}@cs.uni-kl.de

Abstract

There are two custom ways for predicting RNA secondary structures: minimizing the free energy of a conformation according to a thermodynamic model and maximizing the probability of a folding according to a stochastic model. In most cases stochastic grammars are used for the latter alternative applying the maximum likelihood principle for determining a grammar's probabilities. In this paper, building on such a stochastic model, we will analyze the expected minimum free energy of an RNA molecule according to Turner's energy rules. Even if the parameters of our grammar are chosen with respect to structural properties of native molecules only (and therefore independent of molecules' free energy), we prove formulae for the expected minimum free energy and the corresponding variance as functions of the molecule's size which perfectly fit the native behavior of free energies.

This gives proof for a high quality of our stochastic model making it a handy tool for further investigations. In fact, the stochastic model for RNA secondary structures presented in this work has for example been used as the basis of a new algorithm for the (non-uniform) generation of random RNA secondary structures.

1 Introduction

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 this 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 of a fixed size n are assumed to be equiprobable. In fact, in the combinatorial model, it is assumed that base pairing is possible between arbitrary pairs of nucleotides, as only the topology of the planar secondary structure is considered. Thus, the combinatorial model completely abstracts from the RNA sequence 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]). 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], the *stochastic context-free grammar (SCFG) approach* – so far only used for algorithmic purposes – has been applied for analyzing the expected shape of RNA molecules analytically. But do we get a realistic picture of RNA molecules this way? Furthermore, are the shapes considered as typical by such a stochastic model somehow related to the conformation of minimum free energy? To answer these questions, we decided to analyse the expected minimum free energy in such a stochastic model for RNA secondary structures that is based on the SCFG approach.

Considering a single RNA molecule, thermodynamics is responsible for the initially linear structure to fold into a three-dimensional conformation. Here we can assume the resulting structure to (mostly) minimize the free energy. Accordingly, a straight-forward approach to analyze the *expected minimum free energy* of RNA molecules would be to start with random sequences (according to a realistic distribution) and to determine the (minimum) free energies of the corresponding foldings. Unfortunately, it seems impossible to handle such an analysis mathematically. Therefore, in this paper we decided to proceed along the following lines:

[*†] Corresponding author.

- Starting with a database of native RNA molecules (minimum free energy) and a stochastic context-free grammar, a model for typical foldings is determined on grounds of the maximum likelihood (ML) principle for choosing the model's parameters; here we have to expect that secondary structures with a high probability according to the resulting model show a similar behavior to the foldings from the database with respect to the forming of important structural motifs.
- Within the model, the different structural motifs are assigned their contribution to the overall free energy of the entire folding providing a model for the free energy (at its minimum observed for native secondary structures).

If we finally determine the expected (minimum) free energy from this model as a function of the structure's size, we gain knowledge on the behavior of the minimum free energy for native molecules in correspondence with the average minimum free energy determined from the database on statistical grounds. According to this point of view the term *expected (minimum) free energy* should be understood in this paper. However, compared to a mere statistical analysis of the database entries our grammar model is not restricted to fixed sizes of the molecules but allows for the projection of results to sizes not given in the database. Furthermore, comparing the results of our analysis to the averaged minimum free energy values given in the database – in case of a match – provides evidence for a realistic behavior of our model with respect to the different structural motifs (since they all imply different contributions to the overall free energy but the free energies are left unconsidered for the ML approach). As we will see, this is the case indeed and our rather sophisticated stochastic context-free grammar can be used to gather further background information on the space of secondary structures of real RNAs e.g. for validation purposes. Furthermore, it becomes a handy tool for the random generation of secondary structures with a native appearance; a corresponding algorithm has already been implemented (see Section 6 for details). In fact, an algorithm which, for a given structure size n, produces random RNA secondary structures that are – related to the expected free energy[1] of such structures – in most cases realistic is a major improvement over existing approaches which, for example, are only capable of generating secondary structures uniformly for fixed size n (i.e. they deal with the unrealistic combinatorial model). Last but not least, our analysis connects the two most prominent approaches for predicting RNA secondary structure which in our belief is of interest on its own right.

The plan of this paper is given as follows: In Section 2 we will introduce the formal framework of this paper by recalling some basics, definitions and prior results related to RNA secondary structure. Additionally, we will give a short overview on existing literature dealing with computational prediction methods and free energy minimization. Section 3 provides some information on the material and methods that are used in this paper. The main contribution of this article is presented in Section 4 where we will perform an analysis of the expected minimum free energy of a random secondary structure in a stochastic RNA model. In fact, we will start by deriving a stochastic RNA secondary structure model based on a comprehensive SCFG that is appropriate with respect to the free energy rules implied by a common (sequence-dependent) thermodynamic model for computing the free energy of RNA secondary structures. Afterwards, two different (sequence-independent) free energy models based on this common thermodynamic model will be constructed and we describe how to calculate asymptotics for the expected free energy as well as for the corresponding variance of a random secondary structure of size n under the assumption of either model. In Section 5 the respective analytical free energy results will be compared to real world data in order to judge the quality our of models. Finally, in Section 6 we give a short hint at the applicability of our SCFG model for the creation of a non-uniform weighted unranking algorithm that generates random RNA secondary structures according to a realistic distribution. We conclude by presenting and comparing the respective analytically obtained expected free energy results for different types of RNAs.

2 RNA Secondary Structure

Ribonucleic acid (RNA) is a single-stranded *nucleotide polymer*. 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). 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

[1] Of course, the free energy of a secondary structure can not be computed according to one of the common (sequence-dependent) thermodynamic models if the RNA sequence is unknown (since without sequence, there is no thermodynamics). Therefore, we have to consider the corresponding free energy of the secondary structure only under the assumption of a (sequence-independent) energy model that is based on the common thermodynamic model and approximates the energy as good as possible. Details will follow later.

is usually modeled as a string over the alphabet $\Sigma = \{A, C, G, U\}$, i.e. it is represented as a sequence of letters $r_1 r_2 \ldots r_n$, where r_i is either A, C, G or U.
In vivo, single-stranded RNA chains bend and twine about themselves. The reason for this behavior is that the complementary bases A and U resp. C and G form stable base pairs with each other (by creating hydrogen bonds). In addition to these stable *Watson-Crick pairs*, there may occur weaker base pairs, called GU *wobble pairs*, which are formed by the non-complementary bases G and U. All these pairs (Watson-Crick and GU wobble) 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 there are only few restrictions on (complementary) bases to pair, the linear RNA chain is folded into a three-dimensional conformation, called the *tertiary structure* of the RNA molecule, which determines its biochemical activity. It is customary in science to simplify the study of the tertiary structure of an RNA molecule by allowing only non-crossing base pairs such that the resulting structure 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.

2.1 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 to n. This leads to the following definition of a secondary structure of size n:

Definition 2.1. ([ZMT99]) A *secondary structure* $\mathbf{S}$ for an RNA sequence $\mathbf{R}$ of length[2] 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. 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 unstable conformations are considered and any two bases are allowed to pair.
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[3] 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 which, according to Definition 2.2, do not allow isolated base pairs. However, in many models for RNA secondary structures like in our stochastic grammar, isolated base pairs – although being unstable – are allowed. However, for our grammar a large probability for a helix to be extended by additional base pairs excludes isolated pairs from the *typical* structure considered.
For our further investigations, we need to distinguish between different kinds of single-stranded regions. Therefore, we first have to consider the following definition:

Definition 2.3. (k-loop decomposition [ZS84, Zuk86]) If $i.j$ is a base pair in the secondary structure $\mathbf{S}$ and if $i < \kappa < j$, we say that κ is *accessible* from $i.j$ if there is no $i'.j'$ in $\mathbf{S}$ such that $i < i' < \kappa < j' < j$. Similarly, if $\kappa.l$ is also in $\mathbf{S}$, we say that the base pair $\kappa.l$ is *accessible* if both κ and l are accessible. The

[2] For the sake of simplicity, we will say "secondary structure $\mathbf{S}$ of size n" in the sequel, where n is given by the length of the underlying sequence, *not* the cardinality of set $\mathbf{S}$.

[3] A group of $k \geq 2$ 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.

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 (for varying k) $s_0, s_1, s_2, \ldots, s_m$, where s_0 is the null k-loop and $m > 0$ iff $\mathbf{S}$ is nonempty [4].

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*[5] 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 $\mathbf{L}_e$.

Besides many other possible representations, RNA secondary structures can be modeled as strings over the alphabet $\Sigma := \{(,), \bullet\}$, where a dot represents an unpaired nucleotide and a pair of corresponding brackets **( )** represents two bases in the RNA molecule that are paired (see, e.g. [VC85]). However, it should be clear that these dot-bracket representations abstract from the RNA sequence, as they only consider the number of base pairs (and unpaired bases) and their positions.

2.2 Computational Prediction

In bioinformatics, we aim for algorithms predicting the secondary structure of non-coding RNA from its sequence. Due to an exponential growth of the number of possible conformations with respect to the molecule's size, a brute-force attempt is out of reach. As a consequence, more sophisticated methods have been developed. One class of such algorithms builds on SCFG models, which *learn* the typical structural behavior of RNAs from databases of native molecules on stochastic grounds. Then, given an unknown sequence, the most probable folding is computed giving rise to rather accurate predictions (see e.g. [KH99, KH03] for details). However, the most common approach is free energy minimization, i.e. minimizing 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 native structure is the one with lowest free energy.

The most successful and popular method for energy minimization over the last 30 years has been the use of dynamic programming algorithms. In the pioneering work [NPGK78], each base pair $i.j$ in a given secondary structure $\mathbf{S}$ is assigned an energy $e(i.j)$, such that the overall energy of the secondary structure $\mathbf{S}$ is given by $\mathbf{E}(\mathbf{S}) = \sum_{i.j \in \mathbf{S}} e(i.j)$. A corresponding dynamic programming algorithm for folding an RNA molecule that finds a conformation of minimum free energy using thermodynamics and auxiliary information was presented in [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 co-transcriptional folding) have not been incorporated, the predicted optimal (minimum free energy) structure was often not the native one. For these reasons, several efficient algorithms have been developed over the years for generating a set of suboptimal foldings (see, e.g., [WFHS99, Zuk89b]). Implementations of these algorithms are used for example in the MFOLD software [Zuk03] or in the Vienna RNA package [Hof03], which have become widely used tools.

[4] 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.

[5] 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.

3 Methods

In this section, we will provide some information on the material and methods that will be used in the sequel, including thermodynamic models for RNA secondary structure, RNA modeling by stochastic context-free grammars and the considered RNA databases.

3.1 Thermodynamic Models

In the early 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.* Later on, an expanded nearest-neighbor model for formation of RNA helices with canonical base pairs was presented, which was termed the *individual nearest-neighbor hydrogen bond (INN-HB) model* [XSB+98, MSZT99]. Thermodynamics for RNA secondary structures have also been studied 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 paper, 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 $\mathbf{S}$[6]. The thermodynamic parameters that will be used here 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 is derived from [MSZT99] and [ZMT99]. It includes basically all of the latest free energy rules and parameters[7] and will be the foundation of our analysis.
For a more detailled description of this thermodynamic model, i.e., for more information on the distinguished substructures and the corresponding free energy contributions, see Section[8] Sm-I. Finally, it should be mentioned that this model is most commonly called Turner's energy model and therefore, we will also use this name in the sequel.

3.2 Stochastic RNA Models

Stochastic context-free grammars are an extension of context-free grammars and a known concept to model RNA secondary structures (see, e.g. [SBH+94]). For an introduction on stochastic context-free languages, see for example [HF71]. A formal definition is given as follows:

Definition 3.1. ([Neb04a, Neb02b]) A *stochastic context-free grammar (SCFG)* is a 5-tuple $G_{\text{st}} = (I_{\text{st}}, T_{\text{st}}, R_{\text{st}}, S_{\text{st}}, P_{\text{st}})$, where I_{st} (resp. T_{st}) is an alphabet (finite set) of intermediate (resp. terminal) symbols (I_{st} and T_{st} are disjoint), $S_{\text{st}} \in I_{\text{st}}$ is a distinguished intermediate symbol called *axiom* and $R_{\text{st}} \subset I_{\text{st}} \times (I_{\text{st}} \cup T_{\text{st}})^{*}$[9] is a finite set of production rules; in the sequel, we will write $A \to \alpha$ instead of $f = (A, \alpha) \in R_{\text{st}}$. P_{st} is a mapping from R_{st} to $[0,1]$ such that each rule $f \in R_{\text{st}}$ is equipped with a probability $p_f := P_{\text{st}}(f)$. The probabilities are chosen in such a way that for all $A \in I_{\text{st}}$ the equality

$$\sum\nolimits_{f \in R_{\text{st}}, f = A \to \alpha} p_f = 1$$

holds. For $f = A \to \alpha \in R$ with $p_f = P_{\text{st}}(f)$ we will write $p_f : A \to \alpha$ in the sequel.

The concepts of derivation and ambiguity for SCFGs are the same as for usual context-free grammars. This means any word $w \in \mathcal{L}(G_{\text{st}})$ is generated in the same way as by the corresponding context-free grammar $(I_{\text{st}}, T_{\text{st}}, R_{\text{st}}, S_{\text{st}})$. However, we want to stress an important difference with respect to ambiguity between typical applications of SCFGs and ours: When using SCFGs for structure prediction, each sequence has several leftmost derivations representing the different secondary structures possible. Accordingly, a grammar used in this context has to be *syntactically ambiguous.* However, the so-called *semantic ambiguity*, i.e. several derivations representing the same structure, has to be prevented. Since

[6] 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 base pairs. This means that non-canonical base pairs in helices must instead be treated as mismatched pairs for the computation of free energies.

[7] There is only one exception: coaxial stacking (which is a favorable interaction of two helices stacked end to end in multi- and exterior loops) is not considered in our model.

[8] All references starting with Sm are references to the supplementary material.

[9] When A is a set of symbols, A^* denotes the set of all finite strings of symbols of A completed by the empty string ϵ.

we will use an SCFG to generate dot-bracket representations (instead of nucleotide sequences), a syntactically unambiguous grammar is the right choice; each secondary structure can be generated in exactly one way.

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. 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 $\mathcal{L}(G_{\mathrm{st}})$. This especially is always the case when choosing the probabilities according the the maximum likelihood principle [CG98]. 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 $\Pr[d]$ to each derivation d of a word $w \in \mathcal{L}(G_{\mathrm{st}})$. The probability $\Pr[d]$ of a given derivation d is equal to the product of the probabilities of the production rules used in d. Furthermore, we can use the mapping P_{st} to compute the probability $\Pr[w]$ 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 $d_1, \dots, d_k$, then the probability $\Pr[w]$ is given by $\sum_{i=1}^{k} \Pr[d_i]$. Thus, if the consistent SCFG G_{st} is unambiguous, then the probability $\Pr[w]$ 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 w.

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. Generally speaking, the maximum likelihood method is the procedure of finding the value of one or more parameters for a given statistical model. In fact, maximum likelihood estimation is a popular statistical method that is typically used for fitting a statistical model to known sets of data in order to provide estimates for the model's parameters. Particularly, given a fixed set of data (a fixed sample from a larger set) and the corresponding underlying probability model, the maximum likelihood method can be used to determine those values of the considered model parameters that make the data more likely than any other choice of these parameters would make them.

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 considered model parameters are the probabilities of the production rules of G_{st}. Hence, training the SCFG G_{st} fits the probabilities of the production rules of G_{st} so that the resulting probabilities of the 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.

Stochastic Context-Free Grammars and Probability Generating Functions

According to the ideas of Chomsky and Schützenberger [CS63], it is possible to translate a consistent SCFG into a *probability generating function*, defined as follows:

Definition 3.2. ([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 $\sum_{w \in \mathcal{L}(G_{\mathrm{st}}) \cap T_{\mathrm{st}}^k} \Pr[w]$, i.e., the probability that a word of length k is 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 $p_k = \Pr[X = k]$ must provide a probability distribution on the words $w \in \mathcal{L}(G_{\mathrm{st}})$, and therefore coefficients must sum up to 1, i.e. $P(1) = 1$ must hold. Consequently, by evaluating (the closed form[10] of) the function $P(z) = \sum_{k \geq 0} p_k z^k$ 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.

Remark. As indicated by the corresponding probability generating function, a consistent stochastic context-free grammar implies a distribution on the length of its words. In general longer words tend to be generated with smaller probability since we have to apply more grammar rules each implying a

[10] A *closed form* of a generating function P is a representation of P without sums, integrals etc. which allows for the evaluation of $P(z)$ for appropriate values of z.

factor (typically) less than 1 to the probability. However, for our analysis we will switch to conditional expectations fixing the size of the words under consideration. This way we sort of cut out a part of the overall distribution making it a probability distribution on its own by normalizing. Here one could ask for a separated training of the grammar's parameters for different sizes of the input. However, due to the self-similar structure inherent to context-free languages (see pumping-lemma) this would not make a significant difference for inputs like ours which imply the highly iterated application of all the different production rules.

Note that in order to compute the desired analytical free energy results, we will use both stochastic context-free grammars and the methods of (probability) generating functions. To keep our presentation mostly self-contained, we decided to recall the fundamental definitions concerning generating functions in Section Sm-II. For more information, see for example [FS09].

3.3 RNA Data

In order to obtain a realistic RNA secondary structure model, we decided to derive a stochastic model for RNA secondary structures according to biological data. To reach this goal, we will consider a database of known RNA sequences and their corresponding secondary structures (i.e., in this database each structure of size n is given as pair of dot-bracket representation of length n and corresponding primary structure of length n). Since these secondary structures are supposed to be correct foldings of the corresponding sequences, we can assume that any structure contained in the considered database has minimum free energy (or something close to it) among all structures on the same sequence[11].

Moreover, the used database should only contain structures of the same type of RNA or of similar RNA types to ensure the accurray of the resulting RNA structure model (for RNA structures of that type(s)). Therefore, we will consider the following different sets of RNA secondary structures $\mathbf{S} \neq \emptyset$[12] in the sequel:

- tRNA database consisting of 2163 structures (from [SHB$^+$98]),
- 5S rRNA database of 1292 structures (from [SBEB02]),
- SSU rRNA database of 1308 structures (from [WdPWW02]),
- LSU rRNA database of 558 structures (from [WRdP$^+$01]).

In fact, at the end of this paper, we will present analytical expected free energy results for any of these four different types of RNA. However, for the exemplary derivation of a corresponding stochastic model and of the desired results on free energies, we decided to use a database of SSU and LSU rRNA secondary structures. This database is composed of our SSU rRNA database from [WdPWW02] and our LSU rRNA database from [WRdP$^+$01] and thus contains $1308 + 558 = 1866$ RNA secondary structures $\mathbf{S} \neq \emptyset$. For the sake of simplicity, this database of SSU and LSU rRNA structures will be referred to as *biological database* in the following sections.

We decided to exemplarily consider SSU and LSU rRNAs, since they are more interesting than the rather short and hardly variant tRNAs and 5S rRNAs. This is due to the fact that the much longer sequences of these SSU or LSU rRNA molecules imply a significantly larger set of possible structural motifs. This makes us to assume that a corresponding stochastic model for SSU and/or LSU rRNA secondary structures is usually less accurate than for shorter, less invariant RNA structures, since the probabilities of the production rules obtained by training are less explicit due to the larger variety of structure motifs. Moreover, it can be assumed that a corresponding stochastic model derived for two (or more) mixed types of RNAs is probably not as accurate as a corresponding model derived only for a single type (if the mixed types are not similar enough to obtain almost the same rule probabilities for the whole production set of the underlying SCFG). For these reasons, we may assume that if we obtain realistic results when considering SSU and LSU rRNAs (at once), then similar realistic results can also be derived in the same way for any of our four different types of RNA.

[11] Of course, using Turner's energy model to compute the free energy of an RNA secondary structure, this must not always hold, since the thermodynamic parameters are still incomplete. However, since this model is commonly used for free energy minimization algorithms, this assumption seems to be convenient.

[12] Note that due to the constraint $\mathbf{S} \neq \emptyset$, no dot-bracket representations of completely unpaired structures are contained in class $\mathcal{S}$; this is in accordance with MFOLD, where only structures $\mathbf{S} \neq \emptyset$ can be constructed. Since for each sequence length n, there is only one unfolded structure $\mathbf{S} = \emptyset$ and its free energy is equal to 0, this does not change our asymptotic results for a given stochastic model. However, this constraint could improve the quality of the underlying stochastic structure model, as the model's parameters (the production probabilities of the corresponding SCFG) would become less appropriate for a training set containing unfolded secondary structures.

In the following section, we will describe the derivation of a stochastic model for RNA secondary structures from a given database (of known structures of one or more similar types of RNA). Assuming that all secondary structures in the considered database are minimum free energy structures (or something close to them), we obtain a stochastic model for the behavior of different structural motifs in native structures (of the considered RNA types only, not for arbitrary RNAs), from which the expected (near-) minimum free energy of those structures can be derived. It is important to mention that for modeling the different structural motifs, only the secondary structures in the given database are relevant; the corresponding RNA sequences are *not* considered. However, the sequences are in fact used (along with the secondary structures) for the calculation of averaged free energy contributions, since the free energies of structures are strongly sequence-dependent (according to Turner's energy model).

4 Analysis of the Free Energy in a Stochastic RNA Model

Our aim is to determine the expected minimum free energy $G_{37}^{\circ}(\mathbf{S})$ 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. Training this model based on structural features only (leaving the free energy unconsidered), a match of the resulting expected free energy, its variance and corresponding confidence intervals to the free energies of native structures used for training would prove the quality of our model (since all different types of loops yield different contributions to the free energy, only an overall realistic model ought to imply a realistic free energy). As we will prove throughout this paper, this is the case indeed.

4.1 RNA Secondary Structure Model

As our first goal, we want to derive an appropriate SCFG model for RNA secondary structures according to biological data. In particular, we want to exemplarily use our biological database of SSU and LSU rRNA secondary structure data to obtain a corresponding realistic RNA secondary structure model.
Let $\mathcal{S}$ be the combinatorial class of all different dot-bracket representations of secondary structures $\mathbf{S} \neq \emptyset$. This class $\mathcal{S}$ can be modeled by a simple (unambiguous stochastic) context-free grammar with the following production rules:

$$\begin{array}{llll} S \rightarrow CA, & A \rightarrow (L)C, & A \rightarrow (L)CA, & \\ L \rightarrow \bullet\bullet\bullet C, & L \rightarrow CA, & C \rightarrow \epsilon, & C \rightarrow \bullet C. \end{array}$$

Here, non-terminal symbol S produces the whole structure (the exterior loop) which may not be unfolded. Substructures containing at least one base pair (in one or more adjacent helical regions) are produced by symbol A. Moreover, non-terminal symbol L generates loops (of any kind) and C produces single-stranded regions (of arbitrary length).
However, a more sophisticated grammar is needed in order to derive the desired free energy results since this grammar does not distinguish structural motifs with different contributions to the overall free energy. In fact, the basis of our stochastic secondary structure model will be a comprehensive SCFG that has to serve two purposes at the same time: accomodating the energy parameters and still being unambiguous. Accordingly, the desired SCFG for modeling the class $\mathcal{S}$ has to distinguish between different substructures; it must distinguish not only between the different types of k-loops, but also between some special types for hairpin, bulge and interior loops for which there are different free energy rules according to the thermodynamic model as proposed by Turner.
We can thus construct an appropriate grammar by starting with a simple (e.g., the former) unambiguous context-free grammar for class $\mathcal{S}$ and then repeatedly replacing an old rule by a number of new, more specialized productions for the various special cases to be distinguished. Consider, for example, the production rule $L \rightarrow CA$ (of the former simple grammar for $\mathcal{S}$). This rule generates any possible k-loop for $k \geq 2$ (any loop that is not a hairpin loop). By replacing it by the two rules

$$L \rightarrow C(L)C,\ L \rightarrow C(L)CA,$$

it becomes possible to generate any possible 2-loop (i.e., a stacked pair, a bulge (on the left or on the right), or an interior loop) and all kinds of multiloops (i.e., any k-loop with $k \geq 3$) with different productions, which should increase the accuracy of the SCFG model. By additionally replacing the first of these two new rules, $L \rightarrow C(L)C$, by the four productions

$$L \rightarrow (L),\ L \rightarrow \bullet C(L),\ L \rightarrow (L)C\bullet,\ L \rightarrow \bullet C(L)C\bullet,$$

the different types of 2-loops can be generated with distinct rules, yielding a more realistic secondary structure model and later allowing us to assign each type its corresponding loop-dependent free energy parameters. By further repeated replacements of production rules, we can finally obain an appropriate context-free grammar which unambiguously generates the language $\mathcal{S}$ and also matches the thermodynamic model.

In this work, we decided to use the following grammar which has been constructed by modifying the unambiguous context-free grammar given in [Neb02b, Neb04a]:

Definition 4.1. (MoNStER-grammar[13])
A context-free grammar which generates class $\mathcal{S}$ is given by $G = (I_G, \Sigma_G, R_G, S)$, where $I_G = \{S, T, C, A, L, G, B, F, H, P, Q$
$\Sigma_G = \{(,), \bullet\}$ and R_G contains the following rules:

$$\left.\begin{array}{l} f_1 = S \to TAC, \\ f_2 = T \to TAC, \\ f_3 = T \to C, \\ f_4 = C \to C\bullet, \\ f_5 = C \to \epsilon, \end{array}\right\} \text{exterior loop}$$

$$\left.\begin{array}{l} f_6 = A \to (L), \\ f_7 = L \to (L), \end{array}\right\} \text{initiate and extend stem}$$

$$f_8 = L \to M, \quad \text{initiate multiloop}$$

$$\left.\begin{array}{l} f_9 = L \to P, \\ f_{10} = L \to Q, \\ f_{11} = L \to R, \end{array}\right\} \text{initiate interior loop}$$

$$f_{12} = L \to F, \quad \text{initiate hairpin loop}$$

$$f_{13} = L \to G, \quad \text{initiate bulge loop}$$

$$\left.\begin{array}{l} f_{14} = G \to (L)\bullet, \\ f_{15} = G \to (L)B\bullet\bullet, \\ f_{16} = G \to \bullet(L), \\ f_{17} = G \to \bullet\bullet B(L), \\ f_{18} = B \to B\bullet, \\ f_{19} = B \to \epsilon, \end{array}\right\} \text{bulge loops}$$

$$\left.\begin{array}{l} f_{20} = F \to \bullet\bullet\bullet, \\ f_{21} = F \to \bullet\bullet\bullet\bullet, \\ f_{22} = F \to \bullet\bullet\bullet\bullet\bullet H, \\ f_{23} = H \to H\bullet, \\ f_{24} = H \to \epsilon, \end{array}\right\} \text{hairpin loop}$$

$$\left.\begin{array}{l} f_{25} = P \to \bullet(L)\bullet, \\ f_{26} = P \to \bullet(L)\bullet\bullet, \\ f_{27} = P \to \bullet\bullet(L)\bullet, \\ f_{28} = P \to \bullet\bullet(L)\bullet\bullet, \end{array}\right\} \text{small interior loops}$$

$$\left.\begin{array}{l} f_{29} = Q \to \bullet\bullet(L)K\bullet\bullet\bullet, \\ f_{30} = Q \to \bullet\bullet\bullet J(L)K\bullet\bullet, \\ f_{31} = R \to \bullet(L)K\bullet\bullet\bullet, \\ f_{32} = R \to \bullet\bullet\bullet J(L)\bullet, \\ f_{33} = J \to J\bullet, \\ f_{34} = J \to \epsilon, \\ f_{35} = K \to K\bullet, \\ f_{36} = K \to \epsilon, \end{array}\right\} \text{other interior loops}$$

$$\left.\begin{array}{l} f_{37} = M \to U(L)U(L)N, \\ f_{38} = N \to U(L)N, \\ f_{39} = N \to U, \\ f_{40} = U \to U\bullet, \\ f_{41} = U \to \epsilon. \end{array}\right\} \text{multiloop}$$

In order to construct an unambiguous SCFG G_{sto} for class $\mathcal{S}$, 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 $p_f := P(f)$. To ensure that G_{sto} gets consistent, we decided to assign relative frequencies to the production rules in R_G that are derived from our biological database. The resulting probabilities can be found in Table 2 and the last column of Table 5 shown in Section Sm-IV.

[13] MoNStER is sort of an acronym for Modeling Now Stochastically the free Energy of RNA.

4.2 (Static) Free Energy Model

Next, we aim at determining the expected minimum free energy $G_{37}^{\circ}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our stochastic model derived from SSU and LSU rRNA secondary structure data.
To reach this goal, we construct a bivariate generating function that could be written as

$$D_{\text{sto}}(z,y) = \sum\nolimits_{s \in \mathcal{S}} \left(\Pr[s] \cdot y^{g_{\text{sto}}(s)} \right) \cdot z^{|s|},$$

where $g_{\text{sto}}(s)$ denotes the free energy associated with the dot-bracket representation $s \in \mathcal{S}$ under the assumption of a static (basically sequence-independent) free energy model. This energy model is derived from Turner's (strongly sequence-dependent) thermodynamic model for computing the free energy $G_{37}^{\circ}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ (on an RNA sequence $\mathbf{R}$, see above).
In fact, to obtain corresponding fixed, sequence-independent (and also length-independent) values for the different contributions, we decided to use average values for all the different free energy contributions that are considered in Turner's thermodynamic model.

Remark. This approach does not imply a one-to-one correspondence of our generating functions to Turner's model of free energy but – for linearity of expectation – such a correspondence is achieved in expectation. For our purposes, i.e. for computing the expected minimum free energy, this makes no difference at all.

Note that this way, for example, instead of the sequence-dependent free energy contribution for terminal mismatch and/or dangling end stacking[14] (in a multiloop), we will use the corresponding average value (found by sequence counting[15] using our biological database) which will be denoted by stackingMulti. All the resulting average free energy values (i.e., suitable values for the free energy parameters used in the following system (1) obtained by sequence counting using our biological database) are given in Table 3 shown in Section Sm-IV.
Thus, according to the used thermodynamic model and [CS63], this immediately yields the following system of equations, which can be solved for the variable S to obtain the desired bivariate generating function $D_{\text{sto}}(z,y)$:

$$\begin{aligned}
S &= p_1 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot T \cdot A \cdot C,\\
T &= p_2 \cdot y^{(\text{stackingExterior+termAUpenEL})} \cdot T \cdot A \cdot C + p_3 \cdot C,\\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1,\\
A &= p_6 \cdot z \cdot L \cdot z,\\
L &= p_7 \cdot y^{(\text{se})} \cdot z \cdot L \cdot z +\\
&\quad p_8 \cdot y^{(\text{MBLinitiation+stackingMulti+termAUpenML})} \cdot M +\\
&\quad p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + p_{12} \cdot F + p_{13} \cdot y^{(\text{ldeb})} \cdot G,\\
G &= p_{14} \cdot y^{(\text{seBulge})} \cdot z \cdot L \cdot z \cdot z +\\
&\quad p_{15} \cdot y^{(2 \cdot \text{termAUpenBL})} \cdot z \cdot L \cdot z \cdot B \cdot z^2 +\\
&\quad p_{16} \cdot y^{(\text{seBulge})} \cdot z \cdot z \cdot L \cdot z +\\
&\quad p_{17} \cdot y^{(2 \cdot \text{termAUpenBL})} \cdot z^2 \cdot B \cdot z \cdot L \cdot z,\\
B &= p_{18} \cdot B \cdot z + p_{19} \cdot 1,\\
F &= p_{20} \cdot y^{(\text{ldeh+termAUpenHL+GGGLoopBonus+cHairpinOf3})} \cdot z^3 +\\
&\quad p_{21} \cdot y^{(\text{ldeh+tmseh+GGGLoopBonus+cHairpin+tetra})} \cdot z^4 +\\
&\quad p_{22} \cdot y^{(\text{ldeh+tmseh+GGGLoopBonus+cHairpin})} \cdot z^5 \cdot H, \qquad\qquad (1)\\
H &= p_{23} \cdot H \cdot z + p_{24} \cdot 1,\\
P &= p_{25} \cdot y^{(\text{ile1x1})} \cdot z \cdot z \cdot L \cdot z \cdot z +
\end{aligned}$$

[14] This means the sequence-dependent free energy contribution for the stacking interaction of a base pair (i.e., the closing base pair or any accessible base pair in multiloops, and the free base pairs in exterior loops) with its adjacent (i.e., preceding and/or following) unpaired bases in multi- and exterior loops.

[15] This means that the needed average values are computed according to the complete structures (secondary structures and corresponding RNA sequences): for each free energy contribution needed for computing the energy of a certain loop type, we sum up all the corresponding (sequence- and/or length-dependent) energy values for all occurring loops of this type. Then, we divide this sum by the observed number of those loops.

$$
\begin{aligned}
&p_{26} \cdot y^{(\text{ile1x2})} \cdot z \cdot z \cdot L \cdot z \cdot z^2 + \\
&p_{27} \cdot y^{(\text{ile1x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z + \\
&p_{28} \cdot y^{(\text{ile2x2})} \cdot z^2 \cdot z \cdot L \cdot z \cdot z^2, \\
Q = {}&p_{29} \cdot y^{(2\cdot\text{tmsei}+\text{ldei}+\text{asym})} \cdot z^2 \cdot z \cdot L \cdot z \cdot K \cdot z^3 + \\
&p_{30} \cdot y^{(2\cdot\text{tmsei}+\text{ldei}+\text{asym})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot K \cdot z^2, \\
R = {}&p_{31} \cdot y^{(2\cdot\text{tbp1xNil}+\text{ldei}+\text{asym})} \cdot z \cdot z \cdot L \cdot z \cdot K \cdot z^3 + \\
&p_{32} \cdot y^{(2\cdot\text{tbp1xNil}+\text{ldei}+\text{asym})} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot z, \\
J = {}&p_{33} \cdot J \cdot z + p_{34} \cdot 1, \\
K = {}&p_{35} \cdot K \cdot z + p_{36} \cdot 1, \\
M = {}&p_{37} \cdot y^{(2\cdot\text{stackingMulti}+2\cdot\text{termAUpenML})} \cdot \left(U \cdot z^2 \cdot L\right)^2 \cdot N, \\
N = {}&p_{38} \cdot y^{(\text{stackingMulti}+\text{termAUpenML})} \cdot U \cdot z^2 \cdot L \cdot N + p_{39} \cdot U, \\
U = {}&p_{40} \cdot U \cdot z + p_{41} \cdot 1.
\end{aligned}
$$

Remark. Even if we use fixed (sequence-independent) expected energy contributions determined from our database to model the strongly sequence-dependent Turner energy model, the resulting expected minimum free energies – if consistent with the values given in the database – still provide evidence for our model to be realistic. Inspecting system (1) together with Table 3 yields the observation that rather different contributions to the free energy show up for different substructures. Thus, only the *right* behavior of our model with respect to different substructures are likely to introduce the right contributions to the overall free energy.

Using the generating function $D_{\text{sto}}(z, y)$, we can easily obtain the following results:

Theorem 4.1. *Under the assumption of our static free energy model (derived from SSU and LSU rRNAs), the expected minimum free energy $G_{37}^{\circ}(\mathbf{S})$ (in $^{\text{kcal}}/_{\text{mol}}$) of a secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$\mu_{sto,n} := -0.24783007n + 39.16513746 + \mathcal{O}\left(\frac{1}{n}\right), n \to \infty.$$

Theorem 4.2. *Under the assumption of our static model (derived from SSU and LSU rRNAs), the variance of the minimum free energy $G_{37}^{\circ}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n (in $^{\text{kcal}^2}/_{\text{mol}^2}$) is asymptotically given by*

$$\sigma^2_{sto,n} := 2.531493699n + \mathcal{O}(1), n \to \infty.$$

All these analytical (free energy) results are proven in Section Sm-III, by giving a detailled description on how we calculated the expected free energy $\mu_{\text{sto},n}$ and the corresponding variance $\sigma^2_{\text{sto},n}$ of a random secondary structure of size n under the assumption of our model.

Figure 1 shows that the asymptotical representations (asymptotics for coefficients of generating functions) for the expected free energy of a secondary structure of size n and the corresponding variance as presented in Theorems 4.1 and 4.2, respectively, are accurate, since for $n \to \infty$, they converge towards the respective exact values (exact coefficients of generating functions).

However, it might seem inaccurate that for structure sizes up to about $n = 150$, the asymptotical expected free energies are positive, which means that for RNA molecules of these sizes n, the completely unpaired structure (having free energy 0) would be energetically more favorable and thus, they would all remain unfolded. Nevertheless, as the exact values for the expected free energies are all negative[16], this does not imply that the described model is inaccurate; it is only a consequence of the not really fast convergence of the asymptotical values (for given size n) towards the exact values. For sizes appropriate for rRNA our asymptotic is of sufficient precision. Furthermore, in Section 7 we will present corresponding results for different types of RNA (e.g. tRNA) where such small sizes are much more appropriate and where we observe a negative expected minimum free energy even for $n = 0$.

[16] Note that the exact expected free energy values are only positive for structure sizes of $n = 5$ to about $n = 20$. However, the described model only considers SSU and LSU rRNA structures and these sizes n are definitely too short for an SSU or LSU rRNA molecule. Thus, we may ignore these few positive values.

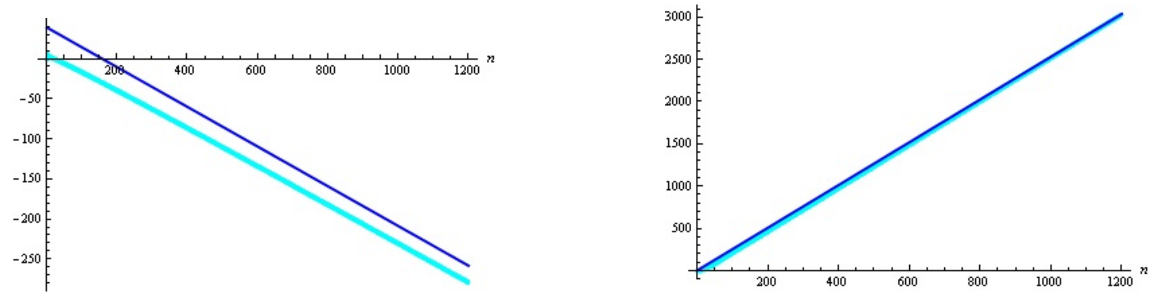

Figure 1: Plots of the coefficient asymptotics (blue) for the expected minimum free energy $G^{\circ}_{37}(\mathbf{S})$ (left) and the corresponding variance (right) of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static free energy model, respectively, as well as points for the respective exact coefficients (cyan, obtained from the corresponding generating functions).

4.3 Alternative (Dynamic) Free Energy Model

The model of the last section considered contributions to the free energy implied by loop lengths not based on the grammar itself but as average values derived from the database. This is sufficient to obtain the realistic behavior of our model with respect to the different loop types and their expected numbers of occurrence but provides no feedback with respect to loop length. Therefore, in this section, we will work out a different free energy model for our stochastic grammar for RNA secondary structures, where the average values for length-dependent contributions are computed in a different way. In fact, given a dot-bracket representation $s \in \mathcal{S}$, we now consider the value of $\widehat{g}_{\text{sto}}(s)$, which denotes the free energy associated with s under the assumption of this alternative dynamic free energy model, instead of the value of $g_{\text{sto}}(s)$.

As any of these length-dependent free energy contributions depends on the number of unpaired nucleotides in loops of a certain type, we will compute the average free energy contribution of a single unpaired nucleotide in loops of the respective type and apply the resulting averages to each unpaired nucleotide. For example, consider a hairpin loop $\mathbf{L}(i.j)$ closed by the base pair $i.j$ in a given secondary structure $s \in \mathcal{S}$. Then (besides other contributions), we have to add the length-dependent free energy contribution for the initiation of this hairpin loop, i.e., a contribution which depends only on the number $j-i-1$ of unpaired bases between the closing base pair $i.j$. In the static model, instead of the correct initiation value (according to Turner's energy parameters), we added the averaged energy value ldeh (for the whole hairpin loop) to the overall energy $g_{\text{sto}}(s)$. However, in the dynamic model, the averaged free energy contribution ldehPerNuc is added for each unpaired nucleotide within $\mathbf{L}(i.j)$, such that for the initiation of this hairpin loop, we have to add the contribution $(j-i-1)\cdot \text{ldehPerNuc}$ to the overall free energy $\widehat{g}_{\text{sto}}(s)$. Again, this provides a correspondence to Turner's model in expectation.

Using such average 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 fixed average values for each loop, very small loops are assigned the same free energy as extremely large loops. By modifiying system (1), 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:

$$\begin{aligned}
S &= p_1 \cdot y^{(\text{stackingExterior}+\text{termAUpenEL})} \cdot T \cdot A \cdot C,\\
T &= p_2 \cdot y^{(\text{stackingExterior}+\text{termAUpenEL})} \cdot T \cdot A \cdot C + p_3 \cdot C,\\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1,\\
A &= p_6 \cdot z \cdot L \cdot z,\\
L &= p_7 \cdot y^{(\text{se})} \cdot z \cdot L \cdot z + p_8 \cdot y^{(\text{MBLOffset})} \cdot\\
&\quad y^{(\text{stackingMulti}+\text{termAUpenML}+\text{MBLHelixPenalty})} \cdot M +\\
&\quad p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + p_{12} \cdot F + p_{13} \cdot G,\\
G &= p_{14} \cdot y^{(\text{seBulge}+\text{ldebPerNuc})} \cdot z \cdot L \cdot z \cdot z +\\
&\quad p_{15} \cdot y^{(2\cdot\text{termAUpenBL}+2\cdot\text{ldebPerNuc})} \cdot z \cdot L \cdot z \cdot B \cdot z^2 +\\
&\quad p_{16} \cdot y^{(\text{seBulge}+\text{ldebPerNuc})} \cdot z \cdot z \cdot L \cdot z +\\
&\quad p_{17} \cdot y^{(2\cdot\text{termAUpenBL}+2\cdot\text{ldebPerNuc})} \cdot z^2 \cdot B \cdot z \cdot L \cdot z,
\end{aligned}$$

$$
\begin{aligned}
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\\
&\quad y^{(3\cdot\text{ldehPerNuc})} \cdot z^3 + p_{21} \cdot y^{(\text{tmseh+GGGLoopBonus+tetra})} \cdot\\
&\quad y^{(4\cdot\text{ldehPerNuc}+4\cdot\text{cHairpinPerNuc})} \cdot z^4 + p_{22} \cdot y^{(\text{tmseh})} \cdot\\
&\quad y^{(\text{GGGLoopBonus}+5\cdot\text{ldehPerNuc}+5\cdot\text{cHairpinPerNuc})} z^5 \cdot H,\\
H &= p_{23} \cdot y^{(\text{ldehPerNuc+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 +\\
&\quad 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 +\\
&\quad 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}+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}+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}+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}+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\\
&\quad 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\\
&\quad 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{2}
$$

To stress the difference of both approaches, we want to consider production $f_{23} = H \to H\,\bullet$ of grammar G_{sto} as one example. Each iteration of this rule produces an additional unpaired nucleotide within a hairpin loop. Nevertheless, within system (1) its corresponding equation $H = p_{23} \cdot H \cdot z + p_{24} \cdot 1$ possesses no variable y since we account for the free energy of the entire hairpin loop assigning an appropriate averaged energy contribution to production f_{22}. Now, we change perspective and use the grammar itself to accumulate the contribution giving rise to a factor $y^{(\text{ldehPerNuc+cHairpinPerNuc})}$ each time the hairpin loop is elongated by the use of production $H \to H\,\bullet$. Therefore, a realistic behavior of the expected free energy derived from the resulting model proves the model's accuracy with respect to loop length.

Again, we can compute suitable values for the free energy parameters used in system (2) by sequence counting using our biological database; results are given in Table 4 of Section Sm-IV. Proceeding as before, system (2) leads to a bivariate generating function $\widehat{D}_{\text{sto}}(z, y)$, where size resp. energies are kept by variable z resp. y. In fact, we obtain the following results for this alternative free energy model (the corresponding intermediate results are given in Section Sm-III):

Theorem 4.3. *Under the assumption of our dynamic energy model (derived from SSU and LSU rRNAs), the expected minimum free energy $G^\circ_{37}(\mathbf{S})$ (in ${}^{\text{kcal}}\!/\!_{\text{mol}}$) of a secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$\widehat{\mu}_{sto,n} := -0.184189537n + 37.1085737 + \mathcal{O}\left(\frac{1}{n}\right), n \to \infty.$$

Theorem 4.4. *Under the assumption of our dynamic model (derived from SSU and LSU rRNAs), the variance of the minimum free energy $G^\circ_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n (in ${}^{\text{kcal}^2}\!/\!_{\text{mol}^2}$) is asymptotically given by*

$$\widehat{\sigma}^2_{sto,n} := 3.963452967n + \mathcal{O}(1), n \to \infty.$$

As before, Figure 2 shows a (not so fast) convergence of the asymptotics presented in Theorems 4.3 and 4.4, respectively, towards the respective exact values.

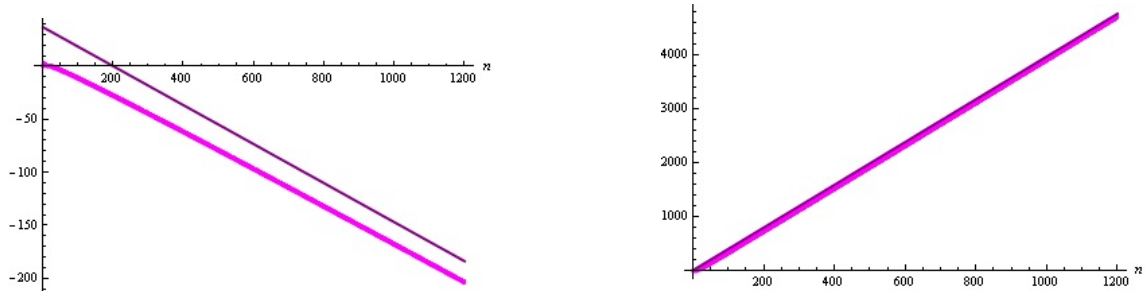

Figure 2: Plots of the coefficient asymptotics (purple) for the expected free energy $G^{\circ}_{37}(\mathbf{S})$ (left) and the corresponding variance (right) of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our dynamic free energy model, respectively, as well as points for the respective exact coefficients (magenta, obtained from the corresponding generating functions).

5 Discussion

In this section, we want to compare our analytic results to real world data in order to judge their quality. For this reason, we first associate a "free energy point" $\{n, G^{\circ}_{37}(\mathbf{S})\}$[17] to each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database. The set of secondary structures given in our biological database will be denoted by bDB in the sequel. Furthermore, the subset of secondary structures in bDB having a size n, with $n_1 \leq n \leq n_2$, i.e. the set $\{\mathbf{S} \in \mathrm{bDB} : n_1 \leq |\mathbf{S}| \leq n_2\}$, will be denoted by bDB_{n_1,n_2} and the subset of elements of size n is given by $\mathrm{bDB}_n := \mathrm{bDB}_{n,n}$.
Additionally, we wanted to derive another set of "average points" $\{n, \mu_n\}$ from our biological database, where

$$\mu_n := \frac{1}{\mathrm{card}(\mathrm{bDB}_n)} \sum\nolimits_{\mathbf{S} \in \mathrm{bDB}_n} G^{\circ}_{37}(\mathbf{S}).$$

However, due to the fact that for many different structure sizes n in the range of structure sizes given in this database, $\mathrm{card}(\mathrm{bDB}_n)$ is not large enough (i.e., there are not enough RNA secondary structures $\mathbf{S}$ of size n), these points are not really appropriate "average points". Therefore, we decided to partition the range of structure sizes into equally large intervals (of size 25 each) and to derive one "average point" for each of these intervals. In fact, we computed the set of "average points" $\left\{n_1 + \frac{n_2-n_1}{2}, \mu_{n_1,n_2}\right\}$, where

$$\mu_{n_1,n_2} := \frac{1}{\mathrm{card}(\mathrm{bDB}_{n_1,n_2})} \sum\nolimits_{\mathbf{S} \in \mathrm{bDB}_{n_1,n_2}} G^{\circ}_{37}(\mathbf{S}),$$

for $(n_2 - n_1) + 1 = 25$ and $n_2 \bmod 25 = 0$. Finally, for the sake of completeness, we determined the corresponding set of "variance points" $\left\{n_1 + \frac{n_2-n_1}{2}, \sigma^2_{n_1,n_2}\right\}$, where

$$\sigma^2_{n_1,n_2} := \frac{\sum_{\mathbf{S} \in \mathrm{bDB}_{n_1,n_2}} \left(\mu_{n_1,n_2} - G^{\circ}_{37}(\mathbf{S})\right)^2}{\mathrm{card}(\mathrm{bDB}_{n_1,n_2}) - 1}.$$

As a start, we plotted our 1866 "free energy points" against the expected free energies as given in Theorems 4.1 (blue line) and 4.3 (purple line). The result as well as a linear regression to the points (green line) is shown in Figure 3.
Considering Figure 3, it seems that both models are realistic but it is not quite clear which should be preferred. Even if the static model is more close to the linear regression, the dynamic *explains* better the sparse points related to large molecules ($n > 4000$). Besides, the linear regression fits nicely for regions where we have many samples but fails to generalize to larger sizes. This shows the necessity of a precise analysis as performed in this paper – a mere inspection of the data at hand is insufficient.
In addition, we observe that the expected free energy $G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of the dynamic model is significantly larger than under the assumption of the static one; the difference grows with n. The reason for this observation is due to the difference between our two energy models: In the static 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 dynamic model, such destabilizing free energy contributions are added for each

[17] Note that for each "free energy point", the corresponding free energy $G^{\circ}_{37}(\mathbf{S})$ of secondary structure $\mathbf{S}$ is computed according to Turner's energy model (and thus neither according to our static energy model nor according to our dynamic energy model, which were both derived from Turner's model).

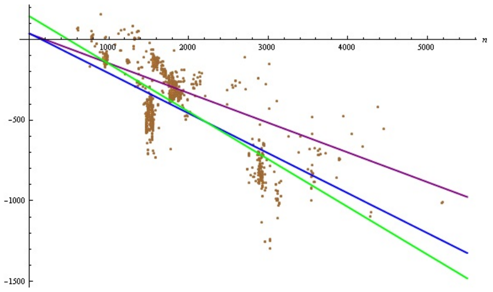

Figure 3: Plots of the expected free energy $G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static (blue) and dynamic (purple) model, respectively, together with the 1866 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database (brown) and a linear regression for these points (green).

unpaired base resp. base pair in the loop. As a consequence, in the static model, very small loops are assigned the same free energy as extremely large loops – thus loop length is no source of variation for the free energy as it should, explaining the smaller variance observed for model one – whereas in the dynamic model, loops of different lengths are assigned different destabilizing (positive) free energy values. In fact, in the dynamic model, loops with a larger number of unpaired bases resp. base pairs are assigned larger 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 dynamic model is greater than that in the static one. Thus, with increasing n, a secondary structure $\mathbf{S} \neq \emptyset$ of size n may contain more loops for which the corresponding destabilizing (positive) free energy in the dynamic model is greater than the corresponding destabilizing free energy in the static model. Since our database contains many structures of size ≤ 2000 and fewer of larger sizes, this gives rise to an underestimated contribution for larger molecules with respect to our static model. For additional evidence on the good quality of our analytical free energy results, see Figures 4 and 5.

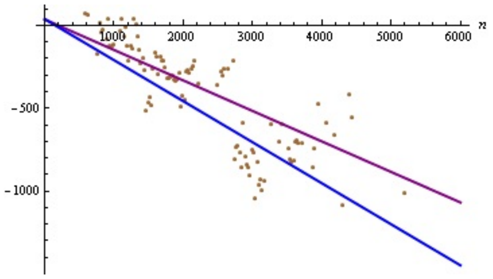

Figure 4: Plots of the expected free energy $G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static (blue) and dynamic (purple) model, respectively, together with the "average points" obtained from our biological database (brown).

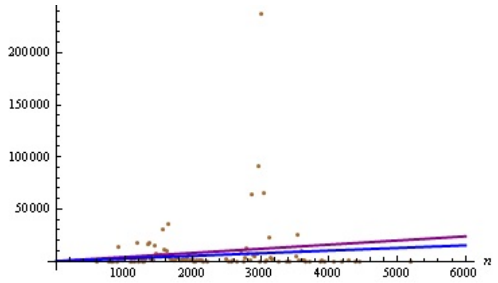

Figure 5: Plots of the variance of the expected free energy $G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static (blue) and dynamic (purple) model, respectively, together with the "variance points" obtained from our biological database (brown).

To further judge our model's accuracy and for their further comparison, we use Chebyshev's inequality to compute probabilities for the free energy of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n to differ at most by a given value from its expectation. First, we consider the desired results under the assumption of our static model:

Theorem 5.1. *Under the assumption of our static model, we can suppose that at most $\frac{100}{k^2}$ percent of free energies $G_{37}^{\circ}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n lie outside and that at least $\left(100 - \frac{100}{k^2}\right)$ percent of them lie inside the open interval $I_{sto,n}(k) := \left(a_{sto,n}(k)\ ^{\text{kcal}}\!/_{\text{mol}}, b_{sto,n}(k)\ ^{\text{kcal}}\!/_{\text{mol}}\right)$, where*

$$a_{sto,n}(k) := 39.16513746 - 1.59106684k\sqrt{n} - 0.24783007n,$$
$$b_{sto,n}(k) := 39.16513746 + 1.59106684k\sqrt{n} - 0.24783007n.$$

Proof. Considering all dot-bracket words $s \in \mathcal{S}_n$, then according to Chebyshev's inequality,

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

Thus the probability for the free energy $g_{\text{sto}}(s)$ associated with a dot-bracket word $s \in \mathcal{S}_n$ to lie outside the open interval

$$I_{\text{sto},n}(k) := (\mu_{\text{sto},n} - k\sigma_{\text{sto},n}, \mu_{\text{sto},n} + k\sigma_{\text{sto},n})$$

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

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

Thus, considering all $s \in \mathcal{S}_n$, we may assume that at most $\frac{100}{k^2}$ percent of the free energy values $g_{\text{sto}}(s)$ lie outside and at least $\left(100 - \frac{100}{k^2}\right)$ percent of them lie inside the interval $I_{\text{sto},n}(k)$, respectively. Finally, consider Theorem 4.1 and Theorem 4.2 to get asymptotical values of $\mu_{\text{sto},n}$ and $\sigma_{\text{sto},n}$ (as $n \to \infty$). □

Note that formally this must only hold for $n \to \infty$, as our theorems only prove asymptotical representations for $\mu_{\text{sto},n}$ and $\sigma_{\text{sto},n}$. However, Figures 3 to 5 provide evidence that even for $n \geq 1000$, our formulae for $\mu_{\text{sto},n}$ and $\sigma_{\text{sto},n}$ are accurate.
Moreover, further evidence can be given by computing two more sets of "interval endpoints"

$$\{n_1 + \frac{n_2 - n_1}{2}, A_{n_1,n_2}(k) := \mu_{n_1,n_2} - k\sigma_{n_1,n_2}\} \text{ and}$$
$$\{n_1 + \frac{n_2 - n_1}{2}, B_{n_1,n_2}(k) := \mu_{n_1,n_2} + k\sigma_{n_1,n_2}\},$$

respectively, where $(n_2 - n_1) + 1 = 25$ and $n_2 \bmod 25 = 0$, and plotting them against the endpoints $a_{\text{sto},n}(k)$ and $b_{\text{sto},n}(k)$ of the open interval $I_{\text{sto},n}(k)$, respectively, as shown in Figure 9 of the supplementary material.
It should be no surprise that the length of any interval $I_{\text{sto},n}(k)$ 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 $I_{\text{sto},n}(k)$ 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 10 of the supplementary material. In the same way, we can derive the corresponding results for the dynamic free energy model. In fact, we immediately obtain:

Theorem 5.2. *Under the assumption of our dynamic model, we find out that at most $\frac{100}{k^2}$ percent of free energies $G_{37}^{\circ}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n lie outside and at least $\left(100 - \frac{100}{k^2}\right)$ percent of them lie inside the open interval $\widehat{I}_{sto,n}(k) := \left(\widehat{a}_{sto,n}(k)\ ^{\text{kcal}}\!/_{\text{mol}}, \widehat{b}_{sto,n}(k)\ ^{\text{kcal}}\!/_{\text{mol}}\right)$, where*

$$\widehat{a}_{sto,n}(k) := 37.1085737 - 1.99084228k\sqrt{n} - 0.184189537n,$$
$$\widehat{b}_{sto,n}(k) := 37.1085737 + 1.99084228k\sqrt{n} - 0.184189537n.$$

The corresponding plots for $\widehat{I}_{\text{sto},n}(k)$ are shown in Figures 11 and 12 of the supplementary material. Comparing Theorems 5.1 and 5.2, it is easy to see that for fixed values of both n and k, the size of the interval $\widehat{I}_{\text{sto},n}(k)$ is always greater than the size of the corresponding interval $I_{\text{sto},n}(k)$, due to the larger variance present in our dynamic model. Figure 6 shows the relative location of both intervals as a function of n for different choices of k.

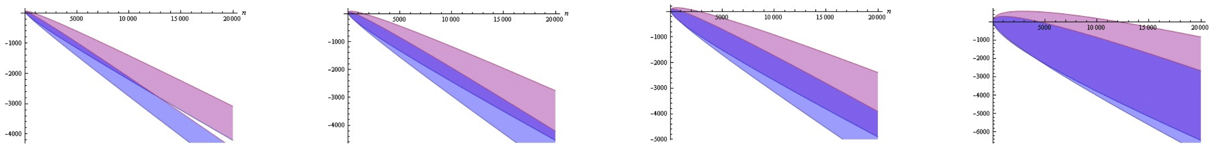

Figure 6: Plots of the intervals $I_{\text{sto},n}(k)$ (blue) and $\widehat{I}_{\text{sto},n}(k)$ (purple), $k \in \{2, \sqrt{10}, \sqrt{20}, 10\}$.

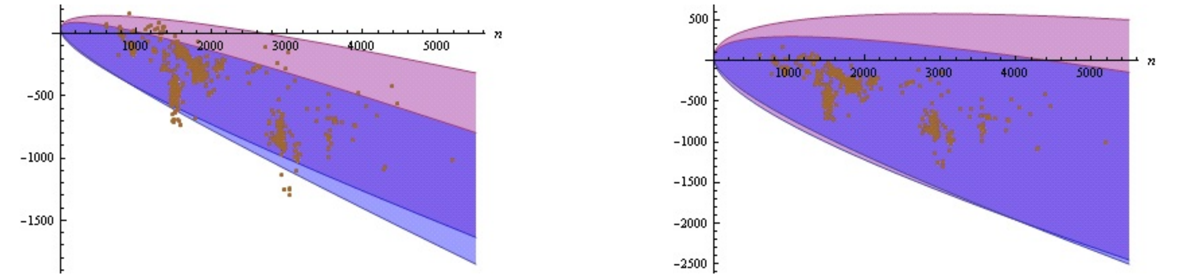

Figure 7: Plots of the intervals $I_{\text{sto},n}(k)$ for the static model (blue) and $\widehat{I}_{\text{sto},n}(k)$ for the dynamic model (purple), for $k = \sqrt{20}$ (left) containing at least 95 percent and $k = 10$ (right) containing at least 99 percent of the free energies $G^{\circ}_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n, respectively. Also displayed are the 1866 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database.

In Figure 7, the intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ are displayed for $k \in \{\sqrt{20}, 10\}$, together with the 1866 "free energy points". As we can see, for $k = \sqrt{20}$, not all the free energies of the RNA secondary structures $\mathbf{S} \neq \emptyset$ of size n given in our biological database lie in the intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$, respectively, but they do for $k = 10$.

Since both models fit nicely with native data, we can conclude that the underlying stochastic RNA secondary structure model based on the comprehensive SCFG $\mathcal{G}_{\text{sto}}$ behaves realistic with respect to free energies and – as the free energy of a given secondary structure is (assumed to be) equal to the sum of the free energies of its substructures – rather likely also with respect to appearance of the different structural motifs of RNA molecules.

6 Applications

Now, having a model at hand which realistically reflects the secondary structure of an RNA molecule and its contributions to free energy, it becomes possible to derive a non-uniform weighted unranking algorithm that generates random RNA secondary structures according to a realistic distribution. In fact, based on the stochastic model for (SSU and LSU r)RNA secondary structures (given by the SCFG $\mathcal{G}_{\text{sto}}$) as presented in this work, the weighted unranking approach of [WN] makes it possible to generate high-quality random RNA secondary structures for a given size n.

Details on the corresponding weighted unranking method are reported elsewhere (see [NS]). However, Figure 8 shows the result of randomly generating secondary structures according to this approach and their realistic behaviour with respect to free energy[18].

7 Conclusions

In this paper, we have studied a stochastic model for RNA secondary structures trained on a database of SSU and LSU rRNA secondary structures derived from [WRdP+01] and [WdPWW02]. Based on the well-known Turner energy model (i.e., the INN-HB model with loop-dependent energy rules [XSB+98, MSZT99]), we have designed two different free energy models for our stochastic model, making use of the thermodynamic parameters given in [MSZT99] (which have also been used for version 3.0 of MFOLD [Zuk03]). For both models, we have computed asymptotics for the expected minimum free energy $G^{\circ}_{37}(\mathbf{S})$ as well as the corresponding variance of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n. To obtain our results, we have used the concept of stochastic context-free grammars and languages and the method of generating functions.

[18] Note that although both energy models have been proven to be realisitic, due to the more realistic variation of free energies connected to varying loop length, we suggest to consider the dynamic model for possible applications.

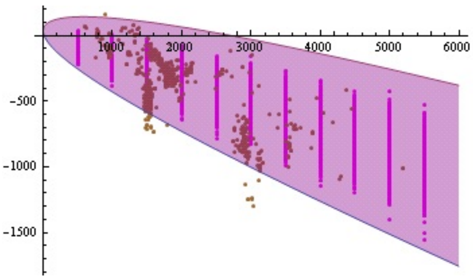

Figure 8: Interval $\widehat{I}_{\mathrm{sto},n}(\sqrt{20})\}$ (purple), corresponding points $\{n, \widehat{g}_{\mathrm{sto}}(s)\}$ (magenta) for each secondary structure s of size n contained in a large set of randomly generated RNA secondary structures and the 1866 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ (brown) for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our biological database.

Even if our grammar was trained only on structural information on native molecules leaving their free energies unconsidered, enriching our stochastic grammar by energy contributions provides a realistic model for free energies. Due to the fact that the RNA secondary structure model induced by our SCFG shows a realistic behaviour (expectation and variance) with respect to minimum free energy and the free energy of a molecule's secondary structure is given by the sum of the energy contributions of all its substructures, it is rather likely that our grammar also shows a realistic picture for all the different structural motifs of a molecule's folding. For that reason, this work marks a stepping stone towards the random generation of RNA secondary structures, and – if a given RNA sequence is also considered – even a stepping stone towards new randomized RNA secondary structure prediction methods.

Table 1: Analytically obtained asymptotical minimum free energy results for secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static and dynamic model, respectively, derived from any of our different databases of RNA structures. Results for the dynamic model are written in italics, respectively.

RNA type	Expected value	Variance
tRNAs	$-0.3014952027n - 2.645002900 + \mathcal{O}(1/n)$	$3.713067703n + \mathcal{O}(1)$
	$\mathit{-0.2961916342n - 2.849523563} + \mathcal{O}(1/n)$	$\mathit{4.169485875n} + \mathcal{O}(\mathit{1})$
5SrRNAs	$-0.3927723339n - 0.3031080166 + \mathcal{O}(1/n)$	$3.048656103n + \mathcal{O}(1)$
	$\mathit{-0.3523192326n + 0.7588376122} + \mathcal{O}(1/n)$	$\mathit{4.649653257n} + \mathcal{O}(\mathit{1})$
SSUrRNAs	$-0.2612433592n + 37.45755167 + \mathcal{O}(1/n)$	$2.50401664n + \mathcal{O}(1)$
	$\mathit{-0.1958643118n + 39.07784261} + \mathcal{O}(1/n)$	$\mathit{4.13009101n} + \mathcal{O}(\mathit{1})$
LSUrRNAs	$-0.2267242639n + 49.45064850 + \mathcal{O}(1/n)$	$2.55752768n + \mathcal{O}(1)$
	$\mathit{-0.1672224228n + 42.13229891} + \mathcal{O}(1/n)$	$\mathit{3.75135111n} + \mathcal{O}(\mathit{1})$
SSU and LSU rRNAs	$-0.2478300708n + 39.16513746 + \mathcal{O}(1/n)$	$2.53149370n + \mathcal{O}(1)$
	$\mathit{-0.1841895371n + 37.10857372} + \mathcal{O}(1/n)$	$\mathit{3.96345297n} + \mathcal{O}(\mathit{1})$

Finally, note that the results that have been derived in this paper under the assumption of our static and dynamic 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 $\mathbf{S} \neq \emptyset$. It remains to mention that the models studied in this work, as well as the presented analytical (free energy) results, should only be used for investigating SSU and LSU rRNA structures; for molecules of other types of RNA (which may have shorter or larger numbers and/or sizes of the different structural motifs, i.e. different expected foldings), the corresponding (free energy) results are more or less different.

This can be observed when comparing the previously presented results (expected values and variances but also probabilities for the grammar rules) to the corresponding results for the four additional sets of RNA data (see Section 3.3). In fact, these analytically obtained free energy results are all presented in Table 1. The respective rule probabilities (relative frequencies) for the productions of the SCFG underlying our stochastic secondary structure model and the corresponding average free energy contributions for the energy parameters used in the static and/or dynamic energy model that were obtained from these databases in order to derive corresponding results, respectively, are tabulated in Tables 5 and 6 of the

supplementary material. In fact, considering the values given there, many structural and energetical information can be extracted for each RNA type and similarities and differences between different types of RNAs can be observed.
Moreover, (Figure 7 and corresponding) Figures 13 to 16 provided in the supplementary material show plots of the corresponding confidence intervals $I_{\text{sto},n}(k)$ and $\widehat{I}_{\text{sto},n}(k)$ under the assumption of the static and the dynamic free energy model for two different suitable values of k, respectively. Note that as expected, the results for RNA types with longer molecules of highly varying structure which imply a significantly larger set of possible structural motifs are not as good as the corresponding ones for rather short and hardly variant types of RNA.
Last but not least, encouraged by one of the referees we implemented a webservice which allows the users to train the MoNStER-grammar with their own data, derive the corresponding energy parameters according to Section 4 and to compute asymptotics related to the free energy. This webservice can be found at `http://wwwagak.cs.uni-kl.de/MoNStER`.

Acknowledgements

The authors wish to thank two anonymous reviewers for their careful and helpful remarks and suggestions made for a previous version of this article.

References

[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.

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

[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.

[FS09] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics.* Cambridge University Press, January 2009.

[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.

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

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

[Hof03] Ivo L. Hofacker. The Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

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

[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.

[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.

[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.

[NS] Markus E. Nebel and Anika Scheid. Random generation of RNA secondary structures according to native distributions. Submitted.

[SBEB02] Maciej Szymanski, Miroslawa Z. Barciszewska, Volker A. Erdmann, and Jan Barciszewski. 5s ribosomal RNA database. *Nucleic Acids Res.*, 30:176–178, 2002.

[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.

[SHB$^+$98] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.

[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.

[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.

[WFHS99] S. Wuchty, W. Fontana, I. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.

[WN] Frank Weinberg and Markus E. Nebel. Non uniform generation of combinatorial objects. Submitted.

[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.

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

Sm-I Details On The Used Thermodynamic Model

The thermodynamic model used in this work basically relys on the k-loop decomposition of a secondary structure $\mathbf{S}$ according to Definition 2.3. Moreover, it additionally distinguishes between some special types of k-loops. Hence, we first have to define all these (special) loop types.

Therefore, let $l_s(\mathbf{L})$ denote the number of single-stranded bases in a loop. 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 Sm-I.1. ([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 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)*.

Our 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 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 preceeded 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 energies 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[19]. 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℃, we use $G^{\circ}_{37}(\mathbf{S})$ to denote the free energy of a secondary structure $\mathbf{S}$ at 37℃.
Finally, in this model, the free energy $G^{\circ}_{37}(\mathbf{S})$ of a secondary structure $\mathbf{S}$ is assumed to be given by the sum of the free energies of all its substructures, formally

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

Sm-II Generating Functions

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

Definition Sm-II.1. ([FS09]) 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 [FS09]. 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.

Definition Sm-II.2. ([FS09]) 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 Sm-II.3. ([FS09]) The *ordinary generating function* (OGF) of a sequence $(a_n)_{n \geq 0}$ is the formal power series

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

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

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

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 $A(z) = \sum_{n \geq 0} a_n z^n$, so that

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

Note that if the elements a_n, $n \geq 0$, of a sequence $(a_n)_{n \geq 0}$ are probabilities, then the corresponding generating function is called *probabilitiy generating function* (PGF), see Definition 3.2.

[19] 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.

Sm-II.1 Computing Generating Functions

A common way to compute a so-called *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 Chomsky and Schützenberger [CS63], in order to derive a generating function.

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 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 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[20].

Sm-II.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 [FS09]. Then, the functions we consider are defined in certain regions of the complex plane $\mathbb{C}$.

Definition Sm-II.4. ([FS09]) 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

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

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 Sm-II.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 Sm-II.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.*

[20] There is one exception: Klein's method of solving polynomial equations of degree 5, which can be extended to solve polynomial equations of degree 6. But for higher order polynomials, only approximate solutions are possible.

Definition Sm-II.7. ([FS09]) 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 Sm-II.1. *(Boundary singularities [FS09]) 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 Sm-II.2. *(Pringsheim's Theorem [FS09]) 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 Sm-II.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*

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

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.

Sm-III Free Energy Analysis – Methods and Proofs

In this section, we prove our analytical (free energy) results by giving a detailled description on how we calculated the expected free energy $G_{37}^{\circ}(\mathbf{S})$ and the corresponding variance of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static free energy model derived from our database of SSU and LSU rRNA secondary structure data. Furthermore, we present the corresponding intermediate results when considering our dynamic free energy model.

Sm-III.1 Computation of the Expected Free Energy

First, we want to describe how to derive the expected free energy $G_{37}^{\circ}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static stochastic model.
According to [CS63], we can translate the rule set R_G (given in Definition 4.1) of the SCFG G_{sto} into the following system of equations for a probability generating function associated to $\mathcal{L}(G_{\mathrm{sto}})$:

$$\begin{aligned}
S &= p_1 \cdot T \cdot A \cdot C, \\
T &= p_2 \cdot T \cdot A \cdot C + p_3 \cdot C, \\
C &= p_4 \cdot C \cdot z + p_5 \cdot 1, \\
A &= p_6 \cdot z \cdot L \cdot z, \\
L &= p_7 \cdot z \cdot L \cdot z + p_8 \cdot M + p_9 \cdot P + p_{10} \cdot Q + p_{11} \cdot R + \\
&\quad p_{12} \cdot F + p_{13} \cdot G, \\
G &= p_{14} \cdot z \cdot L \cdot z \cdot z + p_{15} \cdot z \cdot L \cdot z \cdot B \cdot z^2 + \\
&\quad p_{16} \cdot z \cdot z \cdot L \cdot z + p_{17} \cdot z^2 \cdot B \cdot z \cdot L \cdot z, \\
B &= p_{18} \cdot B \cdot z + p_{19} \cdot 1,
\end{aligned}$$

$$\begin{aligned}
F &= p_{20} \cdot z^3 + p_{21} \cdot z^4 + p_{22} \cdot z^5 \cdot H,\\
H &= p_{23} \cdot H \cdot z + p_{24} \cdot 1, \qquad (3)\\
P &= p_{25} \cdot z \cdot z \cdot L \cdot z \cdot z + p_{26} \cdot z \cdot z \cdot L \cdot z \cdot z^2 +\\
&\quad p_{27} \cdot z^2 \cdot z \cdot L \cdot z \cdot z + p_{28} \cdot z^2 \cdot z \cdot L \cdot z \cdot z^2,\\
Q &= p_{29} \cdot z^2 \cdot z \cdot L \cdot z \cdot K \cdot z^3 +\\
&\quad p_{30} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot K \cdot z^2,\\
R &= p_{31} \cdot z \cdot z \cdot L \cdot z \cdot K \cdot z^3 + p_{32} \cdot z^3 \cdot J \cdot z \cdot L \cdot z \cdot z,\\
J &= p_{33} \cdot J \cdot z + p_{34} \cdot 1,\\
K &= p_{35} \cdot K \cdot z + p_{36} \cdot 1,\\
M &= p_{37} \cdot U \cdot z \cdot L \cdot z \cdot U \cdot z \cdot L \cdot z \cdot N,\\
N &= p_{38} \cdot U \cdot z \cdot L \cdot z \cdot N + p_{39} \cdot U,\\
U &= p_{40} \cdot U \cdot z + p_{41} \cdot 1.
\end{aligned}$$

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

$$\begin{aligned}
S_{\text{sto}}(z) &= \sum\nolimits_{s \in \mathcal{S}} \Pr[s] \cdot z^{|s|}\\
&= \sum\nolimits_{n \geq 0} \left(\sum\nolimits_{s \in \mathcal{S}_n} \Pr[s] \right) \cdot z^n\\
&= \sum\nolimits_{n \geq 0} s_{\text{sto},n} \cdot z^n.
\end{aligned}$$

Here, $\Pr[s]$ is the probability of the dot-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}. Hence, $s_{\text{sto},n}$ is the probability that a dot-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, we have to incorporate free energy values into system (3). Therefore, we first have to recall that each factor $z = z^1$ in this system represents a word of length 1 over $\Sigma_G = \{(,), \bullet\}$, 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 energies. The resulting generating function is a so-called *bivariate* generating function. A formal definition based on [SF01] is given as follows:

Definition Sm-III.1. Given a doubly indexed sequence $(a_{nk})_{n \in \mathbb{N}_0, k \in K}$, where $K \subset \mathbb{R}$ is enumerable[21], the function

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

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 associated with the dot-bracket representation $s \in \mathcal{S}$ under the assumption of the stochastic model under consideration and let K_{sto} be an enumerable[22] 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 $g_{\text{sto}}(s)$ associated with a dot-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

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

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

$$D_{\text{sto}}(z, y) = \sum\nolimits_{s \in \mathcal{S}} \left(\Pr[s] \cdot y^{g_{\text{sto}}(s)} \right) \cdot z^{|s|}.$$

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

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

Once we have constructed the bivariate generating function $D_{\text{sto}}(z,y)$, the desired expected free energy $G_{37}^{\circ}(\mathbf{S})$ 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 bivariate 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 G_{sto} generating exactly all the elements in $\mathcal{S}$, we aim for the expected free energy $G_{37}^{\circ}(\mathbf{S})$ of a secondary structure $\mathbf{S} \neq \emptyset$ under the condition that $\mathbf{S}$ has size n. We have:

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

Consequently,

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

But obviously, for n fix $\Pr[Y=k \text{ and } X=n]$ 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

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

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

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

which is the expected free energy associated with a dot-bracket representation $s\in\mathcal{S}$, under the condition that this dot-bracket word s has length n (conditional expectation).

According to the previous discussion, we now have to modify system (3) 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 (3).

According to our thermodynamic model, most of the contributions to the free energy of a secondary structure $\mathbf{S}$ are sequence-dependent. But for a given dot-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 want to use fixed values for the different length-dependent free energy contributions. Hence, we decided to use average values (derived from a database of known RNA structures) 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 average values of the free energy contributions that are considered to compute the free energy of this structure according to the thermodynamic model have to be summed up in the exponent of y each time such a structure is generated by grammar G_{sto}. This immediately yields system (1).

We can easily calculate suitable average values for the parameters used in system (1) by sequence counting using our biological database. Note that since according to our thermodynamical model, there are no free energy parameters for non-canonical base pairs, those must be treated as mismatches. The resulting average values are given in Table 3 shown in Section Sm-IV.

Using the rational approximations[23] given in the fourth column of Table 3, we can solve system (1) for the variable S to obtain a closed form of the desired bivariate generating function $D_{\text{sto}}(z, y)$ and then proceed the way described before.

In order to compute precise asymptotics for $[z^n]E_{\text{sto}}(z)$ and $[z^n]S_{\text{sto}}(z)$, respectively, we make use of Darboux's theorem as given above. This way, we obtain:

Lemma Sm-III.1. *Under the assumption of our stochastic secondary structure model (derived from SSU and LSU rRNAs), the expected number of secondary structures $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.0001297^{-n} \left(\frac{26.96760121}{n^{3/2}} - \frac{102833.1842}{n^{5/2}} + \mathcal{O}\left(n^{-\frac{7}{2}}\right) \right),$$

$n \to \infty$.

Lemma Sm-III.2. *Under the assumption of our static free energy model (derived from SSU and LSU rRNAs), the first factorial moment for the free energy $G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.0001297^{-n} \left(-\frac{6.68338252}{\sqrt{n}} + \frac{26541.34513}{n^{3/2}} + \mathcal{O}\left(n^{-\frac{5}{2}}\right) \right),$$

$n \to \infty$.

Afterwards, dividing the resulting asymptotics one by the other and computing the series expansion of this fraction about $n \to \infty$ yields an asymptotic for the expected free energy of a secondary structure $\mathbf{S} \neq \emptyset$ of size n under the assumption of the model under consideration. A floating point approximation of this asymptotic is given in Theorem 4.1.

Sm-III.2 Computing the Variance of Free Energies

Now, we describe how to compute the variance $\sigma^2_{\text{sto},n}$ of the free energy $G^{\circ}_{37}(\mathbf{S})$ 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

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

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

As

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

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

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

which is the variance of the free energy $g_{\text{sto}}(s)$ associated with a random secondary structure s conditioned on length n.
By applying Darboux's theorem 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 the following result:

Lemma Sm-III.3. *Under the assumption of our static free energy model (derived from SSU and LSU rRNAs), the second factorial moment for the free energy $G^\circ_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.0001297^{-n}\left(1.65634316\sqrt{n} - \frac{6764.54734}{\sqrt{n}} + \mathcal{O}\left(n^{-\frac{3}{2}}\right)\right),$$

$n \to \infty$.

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 Theorem 4.2.

Sm-III.3 Results for the Dynamic Model

Under the assumption of the alternative dynamic energy model, we use the bivariate generating function $\widehat{D}_{\text{sto}}(z,y)$ (instead of $D_{\text{sto}}(z,y)$ which was used for the static model) to derive to desired results.

Obviously, the asymptotic for the expected number of secondary structures $\mathbf{S} \neq \emptyset$ of size n does not depend on the free energy function $\widehat{g}_{\mathrm{sto}}(s)$ used. Accordingly, the asymptotic given in Lemma Sm-III.1 also holds for the dynamic free energy model and the remaining needed asymptotics can be derived in the same way as for the static free energy model. Thus, we obtain the following results:

Lemma Sm-III.4. *Under the assumption of our dynamic free energy model (derived from SSU and LSU rRNAs), the first factorial moment for the free energy $G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.0001297^{-n}\left(-\frac{4.96714998}{\sqrt{n}}+\frac{19941.5258}{n^{3/2}}+\mathcal{O}\left(n^{-\frac{5}{2}}\right)\right),$$

$n \to \infty$.

Lemma Sm-III.5. *Under the assumption of our dynamic free energy model (derived from SSU and LSU rRNAs), the second factorial moment for the free energy $G^{\circ}_{37}(\mathbf{S})$ of a random secondary structure $\mathbf{S} \neq \emptyset$ of size n is asymptotically given by*

$$1.0001297^{-n}\left(0.91489706\sqrt{n}-\frac{3745.49229}{\sqrt{n}}+\mathcal{O}\left(n^{-\frac{3}{2}}\right)\right),$$

$n \to \infty$.

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 Theorems 4.3 and 4.4.

Sm-IV Tables and Figures

Table 2: 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
f_1	$p_1 := 1$	f_2	$p_2 := \frac{5543}{6476}$
f_3	$p_3 := \frac{933}{6476}$	f_4	$p_4 := \frac{74489}{81898}$
f_5	$p_5 := \frac{7409}{81898}$	f_6	$p_6 := 1$
f_7	$p_7 := \frac{605069}{792975}$	f_8	$p_8 := \frac{31912}{792975}$
f_9	$p_9 := \frac{4912}{264325}$	f_{10}	$p_{10} := \frac{5821}{158595}$
f_{11}	$p_{11} := \frac{1893}{264325}$	f_{12}	$p_{12} := \frac{2723}{31719}$
f_{13}	$p_{13} := \frac{38399}{792975}$	f_{14}	$p_{14} := \frac{11667}{38399}$
f_{15}	$p_{15} := \frac{7235}{38399}$	f_{16}	$p_{16} := \frac{11831}{38399}$
f_{17}	$p_{17} := \frac{7666}{38399}$	f_{18}	$p_{18} := \frac{7781}{12748}$
f_{19}	$p_{19} := \frac{4967}{12748}$	f_{20}	$p_{20} := \frac{3912}{68075}$
f_{21}	$p_{21} := \frac{23208}{68075}$	f_{22}	$p_{22} := \frac{8191}{13615}$
f_{23}	$p_{23} := \frac{32509}{40700}$	f_{24}	$p_{24} := \frac{8191}{40700}$
f_{25}	$p_{25} := \frac{533}{4912}$	f_{26}	$p_{26} := \frac{1053}{4912}$
f_{27}	$p_{27} := \frac{2963}{14736}$	f_{28}	$p_{28} := \frac{7015}{14736}$
f_{29}	$p_{29} := \frac{4986}{29105}$	f_{30}	$p_{30} := \frac{24119}{29105}$
f_{31}	$p_{31} := \frac{2357}{5679}$	f_{32}	$p_{32} := \frac{3322}{5679}$
f_{33}	$p_{33} := \frac{57179}{84620}$	f_{34}	$p_{34} := \frac{27441}{84620}$
f_{35}	$p_{35} := \frac{37994}{53725}$	f_{36}	$p_{36} := \frac{15731}{53725}$
f_{37}	$p_{37} := 1$	f_{38}	$p_{38} := \frac{23211}{55123}$
f_{39}	$p_{39} := \frac{31912}{55123}$	f_{40}	$p_{40} := \frac{172939}{212588}$
f_{41}	$p_{41} := \frac{39649}{212588}$		

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

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 4: Averaged contributions used in the dynamic 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 5: Floating point approximations of the probabilities (relative frequencies) for the production rules of the SCFG G_{sto} (rounded to five decimal places).

Probability p_i for f_i	tRNAs	5S rRNAs	SSU rRNAs	LSU rRNAs	SSU and LSU rRNAs
p_1	1.00000	1.00000	1.00000	1.00000	1.00000
p_2	0.00185	0.00000	0.85703	0.85327	0.85593
p_3	0.99815	1.00000	0.14297	0.14673	0.14407
p_4	0.50807	0.41631	0.88637	0.93924	0.90953
p_5	0.49193	0.58369	0.11363	0.06076	0.09047
p_6	1.00000	1.00000	1.00000	1.00000	1.00000
p_7	0.79898	0.78648	0.76554	0.75885	0.76304
p_8	0.04832	0.02794	0.04003	0.04060	0.04024
p_9	0.00750	0.00028	0.01910	0.01772	0.01858
p_{10}	0.00004	0.05587	0.03471	0.04003	0.03670
p_{11}	0.00000	0.00002	0.00650	0.00827	0.00716
p_{12}	0.14512	0.05587	0.07938	0.09666	0.08585
p_{13}	0.00004	0.07354	0.05474	0.03788	0.04843
p_{14}	0.00000	0.39077	0.29594	0.32290	0.30383
p_{15}	0.00000	0.38254	0.18003	0.20866	0.18842
p_{16}	1.00000	0.22346	0.31951	0.28058	0.30811
p_{17}	0.00000	0.00323	0.20452	0.18786	0.19964
p_{18}	0.00000	0.08508	0.67016	0.32311	0.61037
p_{19}	0.00000	0.91492	0.32984	0.67689	0.38963
p_{20}	0.00000	0.10372	0.06499	0.04715	0.05747
p_{21}	0.00015	0.37577	0.40745	0.24963	0.34092
p_{22}	0.99985	0.52051	0.52756	0.70322	0.60161
p_{23}	0.72677	0.88013	0.79027	0.80679	0.79875
p_{24}	0.27323	0.11987	0.20973	0.19321	0.20125
p_{25}	0.94328	0.69230	0.01815	0.27119	0.10851
p_{26}	0.00000	0.30770	0.18947	0.25922	0.21437
p_{27}	0.00000	0.00000	0.20509	0.19384	0.20107
p_{28}	0.05672	0.00000	0.58729	0.27575	0.47605
p_{29}	0.00000	0.00000	0.13630	0.22202	0.17131
p_{30}	1.00000	1.00000	0.86370	0.77798	0.82869
p_{31}	0.00000	1.00000	0.27535	0.59862	0.41504
p_{32}	0.00000	0.00000	0.72465	0.40138	0.58496
p_{33}	0.92593	0.63784	0.63188	0.72983	0.67572
p_{34}	0.07407	0.36216	0.36812	0.27017	0.32428
p_{35}	0.96000	0.66927	0.69266	0.72484	0.70719
p_{36}	0.04000	0.33073	0.30734	0.27516	0.29281
p_{37}	1.00000	1.00000	1.00000	1.00000	1.00000
p_{38}	0.50000	0.00000	0.34303	0.51585	0.42108
p_{39}	0.50000	1.00000	0.65697	0.48415	0.57892
p_{40}	0.69821	0.72426	0.81335	0.81370	0.81349

Table 6: Averaged contributions used in the static and/or dynamic free energy model for the stochastic model for RNA secondary structures.

Parameter	tRNAs	5S rRNAs	SSU rRNAs	LSU rRNAs	SSU and LSU rRNAs
ldeh	5.98406	6.19129	5.8122	5.82655	5.81825
ldehPerNuc	0.794125	1.00921	1.11209	1.02171	1.07399
tmseh	−1.1645	−1.09879	−1.3771	−1.24903	−1.32252
GGGLoopBonus	−0.00237471	−0.00170279	−0.00290532	−0.0239955	−0.0117962
cHairpinOf3	unknown	0	0.00711215	0.00931953	0.00787522
cHairpin	0	0	0.000312356	0.00134211	0.000751223
cHairpinPerNuc	0	0	0.0000738789	0.000329859	0.000182974
termAUpenHL	unknown	0.0265152	0.329621	0.251109	0.30248
tetra	0	−1.89334	−1.50224	−1.16806	−1.39906
se	−2.39677	−2.29813	−2.11829	−2.18403	−2.14328
seBulge	−2.45	−2.12372	−2.09075	−2.30828	−2.15362
ldeb	3.8	3.41366	3.61455	3.47012	3.57223
ldebPerNuc	3.8	2.83385	2.7694	2.81756	2.78351
termAUpenBL	unknown	0.0201056	0.230861	0.262892	0.240451
ile1x1	0.669725	1.02882	0.786472	1.12073	0.8869
ile2x2	0.93	1.025	0.880492	0.776361	0.858963
ile1x2	unknown	3.11818	2.98851	3.54455	3.20486
ldei	4.05005	2.40294	2.23471	2.29463	2.25941
ldeiPerNuc	0.097602	0.26903	0.302712	0.297467	0.30055
asym	3.0	0.393776	0.76318	0.989365	0.856416
tmsei	0.35	−0.209509	−0.0876916	−0.0894717	−0.0884185
tbp1xNil	unknown	0.0233333	0.415896	0.239564	0.339704
MBLinitiation	5.0	4.6	4.80892	5.02615	4.89098
MBLOffset	3.4	3.4	3.4	3.4	3.4
MBLFreeBasePenalty	0	0	0	0	0
MBLHelixPenalty	0.4	0.4	0.4	0.4	0.4
stackingMulti	−0.689972	−1.11179	−1.11424	−1.10281	−1.10953
termAUpenML	0.227015	0.0823013	0.201926	0.179714	0.192775
stackingExterior	−1.33138	−0.712693	−1.04677	−1.02861	−1.04144
termAUpenEL	0.113982	0.249613	0.318013	0.311859	0.316206

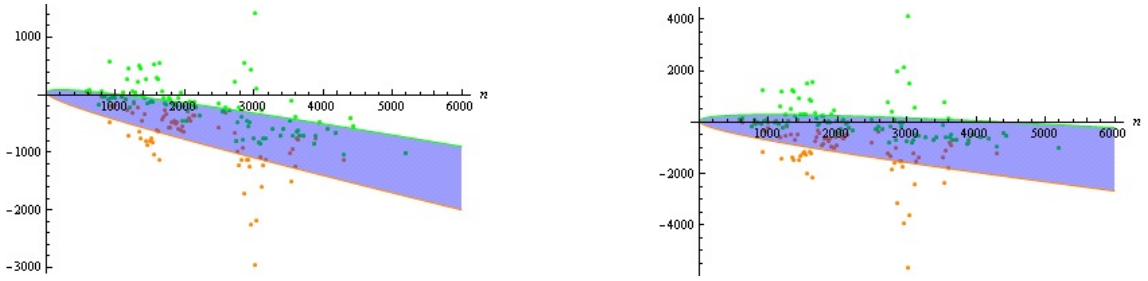

Figure 9: Plots of the endpoints $a_{\mathrm{sto},n}(k)$ (orange line) and $b_{\mathrm{sto},n}(k)$ (green line) of the open intervals $I_{\mathrm{sto},n}(k)$, for $k = \sqrt{20}$ (left) and $k = 10$ (right) containing at least 95 percent and at least 99 percent of the free energies $G^{\circ}_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our static model, respectively, together with the corresponding "interval endpoints" $\{n_1 + \frac{n_2-n_1}{2}, A_{n_1,n_2}(k)\}$ (orange) and $\{n_1 + \frac{n_2-n_1}{2}, B_{n_1,n_2}(k)\}$ (green) obtained from our biological database.

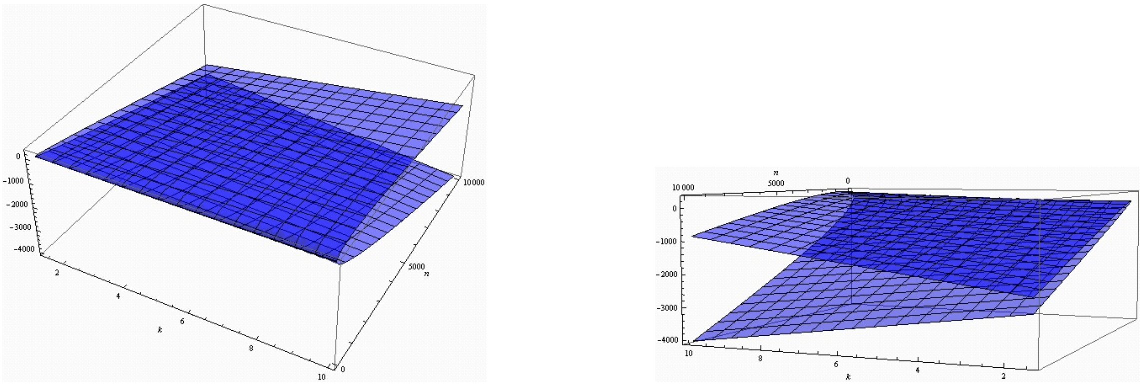

Figure 10: The two endpoints $a_{\mathrm{sto},n}(k)$ and $b_{\mathrm{sto},n}(k)$ of the open interval $I_{\mathrm{sto},n}(k)$, plotted as functions in both k and n, for $\sqrt{2} \leq k \leq 10$ and $1 \leq n \leq 10000$, respectively. Both three-dimensional plots contain exactly the same information, but they are shown from different points of view.

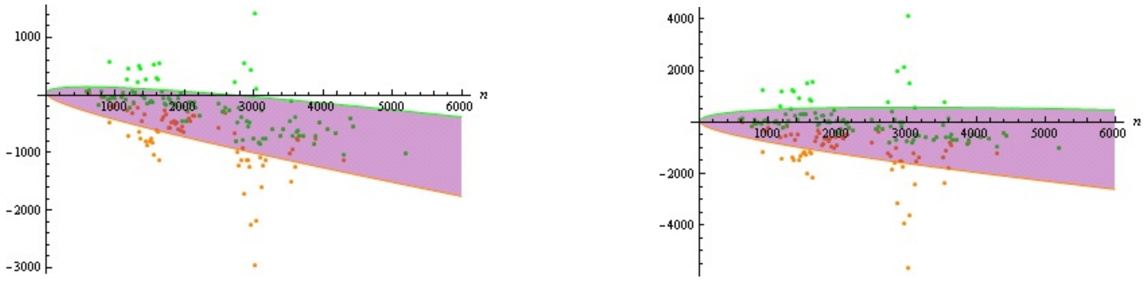

Figure 11: Plots of the endpoints $\widehat{a}_{\mathrm{sto},n}(k)$ (orange line) and $\widehat{b}_{\mathrm{sto},n}(k)$ (green line) of the open intervals $\widehat{I}_{\mathrm{sto},n}(k)$, for $k = \sqrt{20}$ (left) and $k = 10$ (right) containing at least 95 percent and at least 99 percent of the free energies $G^{\circ}_{37}(\mathbf{S})$ of all secondary structures $\mathbf{S} \neq \emptyset$ of size n under the assumption of our dynamic model, respectively, together with the corresponding "interval endpoints" $\{n_1 + \frac{n_2-n_1}{2}, A_{n_1,n_2}(k)\}$ (orange) and $\{n_1 + \frac{n_2-n_1}{2}, B_{n_1,n_2}(k)\}$ (green) obtained from our biological database.

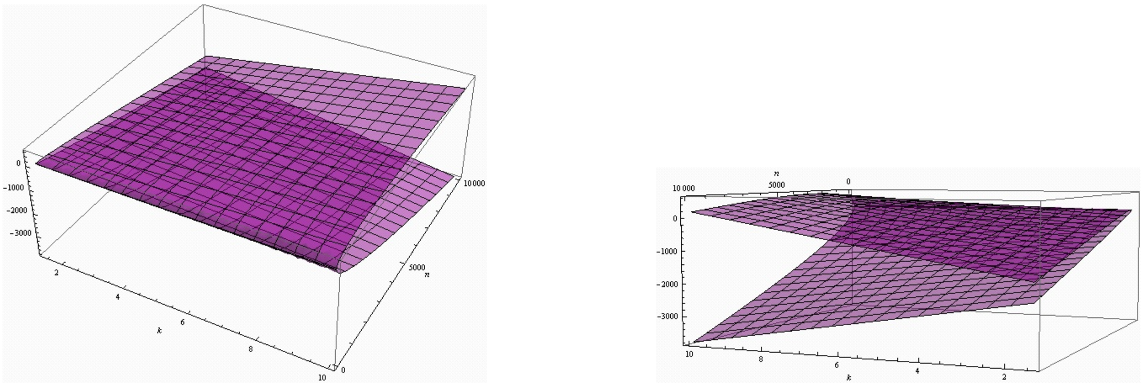

Figure 12: The two endpoints $\widehat{a}_{\text{sto},n}(k)$ and $\widehat{b}_{\text{sto},n}(k)$ of the open interval $\widehat{I}_{\text{sto},n}(k)$, plotted as functions in both k and n, for $\sqrt{2} \leq k \leq 10$ and $1 \leq n \leq 10000$, respectively. Both three-dimensional plots contain exactly the same information, but they are shown from different points of view.

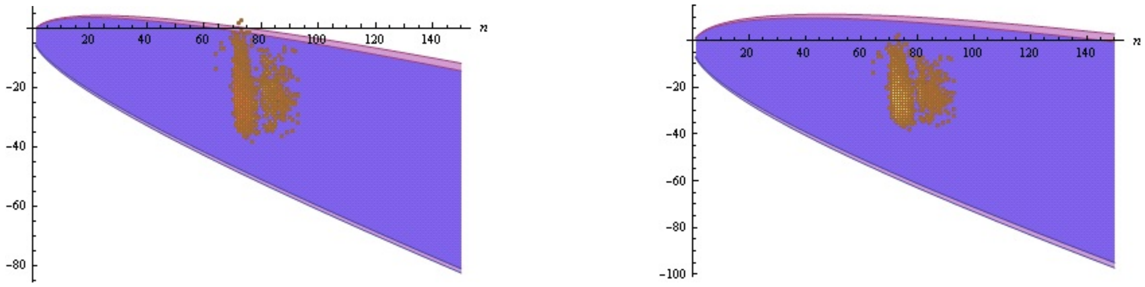

Figure 13: Plots of the intervals $I_{\text{sto},n}(k)$ (blue) and $\widehat{I}_{\text{sto},n}(k)$ (purple) derived from our tRNA database, for $k = \sqrt{2}$ (left) and $k = 2$ (right), respectively, together with the 2163 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our tRNA database.

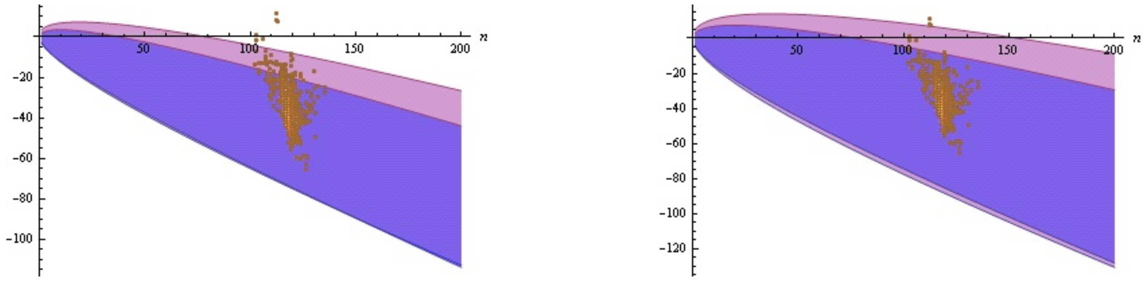

Figure 14: Plots of the intervals $I_{\text{sto},n}(k)$ (blue) and $\widehat{I}_{\text{sto},n}(k)$ (purple) derived from our 5S rRNA database, for $k = \sqrt{2}$ (left) and $k = 2$ (right), respectively, together with the 1292 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our 5S rRNA database.

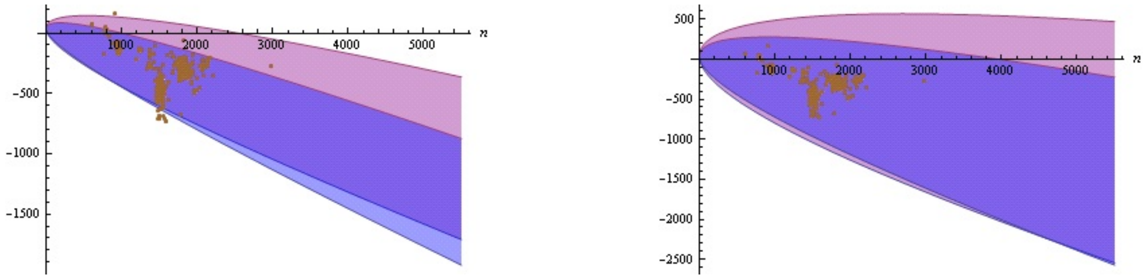

Figure 15: Plots of the intervals $I_{\text{sto},n}(k)$ (blue) and $\widehat{I}_{\text{sto},n}(k)$ (purple) derived from our SSU rRNA database, for $k = \sqrt{20}$ (left) and $k = 10$ (right), respectively, together with the 1308 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our SSU rRNA database.

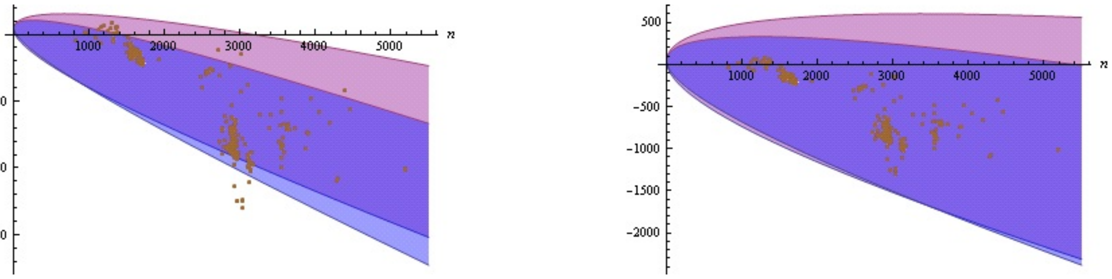

Figure 16: Plots of the intervals $I_{\text{sto},n}(k)$ (blue) and $\widehat{I}_{\text{sto},n}(k)$ (purple) derived from our LSU rRNA database, for $k = \sqrt{20}$ (left) and $k = 10$ (right), respectively, together with the 558 points $\{n, G^{\circ}_{37}(\mathbf{S})\}$ for each secondary structure $\mathbf{S} \neq \emptyset$ of size n given in our LSU rRNA database.