J. Math. Biol. (2014) 69:1743–1772
DOI 10.1007/s00285-013-0750-y

Mathematical Biology

Asymptotic distribution of motifs in a stochastic context-free grammar model of RNA folding

Svetlana Poznanović · Christine E. Heitsch

Received: 16 April 2012 / Revised: 23 August 2013 / Published online: 3 January 2014

Abstract We analyze the distribution of RNA secondary structures given by the Knudsen–Hein stochastic context-free grammar used in the prediction program Pfold. Our main theorem gives relations between the expected number of these motifs—independent of the grammar probabilities. These relations are a consequence of proving that the distribution of base pairs, of helices, and of different types of loops is asymptotically Gaussian in this model of RNA folding. Proof techniques use singularity analysis of probability generating functions. We also demonstrate that these asymptotic results capture well the expected number of RNA base pairs in native ribosomal structures, and certain other aspects of their predicted secondary structures. In particular, we find that the predicted structures largely satisfy the expected relations, although the native structures do not.

Keywords RNA secondary structure · Stochastic context-free grammar · Central limit theorem

Mathematics Subject Classification (2000) 92D20 · 05A16 · 60F05

This work was supported by a BWF CASI grant to C. E. Heitsch. C. E. Heitsch was also supported in part by NIH NIGMS R01 GM083621.

S. Poznanović (✉)
Department of Mathematical Sciences,
Clemson University, Clemson, USA
e-mail: spoznan@clemson.edu

C. E. Heitsch
School of Mathematics, Georgia Institute of Technology,
Atlanta, USA
e-mail: heitsch@math.gatech.edu

1 Introduction

Knowing the base pairings of an RNA sequence can reveal important information about the molecule's function but, unfortunately, experimental determination of the secondary structure is too often nontrivial. For this reason, computational methods have become a standard approach to RNA secondary structure prediction. Most of these prediction methods are based on energy minimization (Mathews and Turner 2006) and depend on the model for the folding free energy change. In order to increase the prediction accuracy, the thermodynamic model has been refined over the years with the inclusion of hundreds of different parameters, most of them experimentally determined (Turner and Mathews 2010). Alas, the prediction accuracy still varies widely (Doshi et al. 2004). As an alternative, methods that use stochastic context-free grammars (SCFGs) have been developed (Eddy and Durbin 1994; Sakakibara et al. 1994). One advantage of these methods over thermodynamic optimization is that phylogenetic information can be incorporated into the prediction, yielding a consensus structure for an aligned family of homologous sequences. One example of such a prediction program is Pfold (Knudsen and Hein 2003).

When developing a prediction method based on a SCFG, several choices need to be made including the SCFG to be used and the set of probabilities for the grammar rules. In designing an SCFG for RNA structure prediction, there is a trade-off between grammar simplicity and prediction accuracy. Simpler grammars are more suitable for computational purposes but are less likely to achieve very high prediction accuracy. Dowell and Eddy (2004) performed an evaluation of the performance of several light-weight SCFGs in the prediction of secondary structures. The Knudsen–Hein grammar (Knudsen and Hein 1999) used in Pfold (Knudsen and Hein 2003) was found to be the most accurate one with prediction accuracy comparable to energy minimization programs, while being significantly simpler than the other SCFGs tested. The authors of (Dowell and Eddy 2004) conclude that "after exploring various alternative SCFG designs, we confirm that the Knudsen/Hein grammar is an excellent, simple framework in which to develop some probabilistic RNA analysis methods". Despite this, to date no rigorous mathematical analysis of this model has been done. This is desirable in order to understand the potential of increasing the prediction accuracy of this grammar by changing the probabilities, since the Sensitivity and the Positive Predicted Value of such predictions are still below 50 % for a lot of sequences (Dowell and Eddy 2004 Table 3).

The goal of this paper is to help clarify the effects of changing the probability parameters for the Knudsen–Hein SCFG. Using tools from analytic combinatorics, we first describe the probability distribution of various RNA features induced by this SCFG over the sets $\{(seq, str) : seq \in \{A, C, G, U\}^*, |seq| = n\}, n \geq 1$, i.e., over all secondary structures of length n. Since these sets contain all possible sequences, the distributions are only affected by the transmission probabilities. As usual for irreducible aperiodic context-free structures, we prove that the distributions of many biologically relevant motifs (helices, hairpins, multibranch loops, etc.) are asymptotically Gaussian for almost all choices of the transmission probabilities for the grammar rules. Moreover, we find explicit formulas for the first two moments of the distributions of these motifs. As an unexpected consequence, we find a set of relations

(Theorem 1) between the expected number of helices and various types of loops in a structure with n nucleotides that are *independent of the grammar probabilities*.

For example, the number of helices is expected to be four times larger than the number of multiloops. Hence, our analysis offers a possible mathematical explanation why the Knuden–Hein grammar predicts the clover-leaf tRNA structure (Knudsen 2005) well.

Of course, these results hold for the homopolymer model of RNA base pairing, without taking into account the base composition of a biological sequence. To assess the confounding effect of the emission probabilities on the predicted structures, we compared the model expectations to observed motifs in the most probable structures for our test set of ribosomal sequences. This comparison is done under the common assumption (for instance, when approximating the inside-outside parameters) that the parse found by the Cocke–Younger–Kasami (CYK) algorithm is the only path with significant probability. According to Durbin et al. (1998), this is "a somewhat startling assumption which however in many cases is surprisingly good."

We find that agreement between the model expectations and the CYK predictions varies for the different types of motifs. For example, the number of base pairs in 91 % of the predictions falls within one standard deviation of the model mean. In contrast, the percentage for helices is only 22 %. Despite this variation in the occurrence of individual motifs, however, *the relations in Theorem 1 for the model expectations largely hold for the CYK predictions* (see Table 4 in Sect. 6). For instance, the ratio of the number of helices to multiloops in the predicted structures is quite close to 4 for the longer sequences.

Importantly, these ratios do not hold for the native ribosomal structures. Thus, Theorem 1, as corroborated by Table 4, indicates that the CYK prediction accuracy for the long 16S and 23S sequences cannot be significantly improved with a simple change of parameters. In particular, this confirms that the strength of Pfold is in coupling the Knudsen–Hein grammar with phylogenetic information from sequence alignments.

The outline of the paper is as follows. In Sect. 2 we state our main results and discuss how they relate to other work on secondary structure analysis. In Sect. 3, we give the formal definitions of secondary structure and the Knudsen–Hein SCFG. In Sect. 4, we illustrate the method of singularity analysis of generating functions on which our proofs are based. In Sect. 5, we derive the central limit theorems for various types of motifs and the asymptotic means as functions of the grammar probabilities. We additionally compute the expected number of multibranch loops of a fixed degree and analyze the structure of the exterior loop. Finally, in Sect. 6, we compare the theoretical results with the secondary structures from the Comparative RNA website (CRW) (Cannone et al. 2002) and the structures predicted for the same sequences using the CYK algorithm with the default Pfold parameters.

2 Main results and discussion of related work

An SCFG induces a probability distribution over all words of fixed length by appropriate normalization of probabilities. Then for a given sequence the predicted structure can be compared to the expected secondary structure with the same number of bases

(e.g. (Nebel 2002b). Here we focus on the distribution of biologically meaningful structural motifs. In particular, we compute the expected number of different loop structures, including base pairs and helices, and compare these with the distribution in native and predicted ribosomal structures. Let $\mathbb{X}_n$ be the number of base pairs, or helices, or loops of a fixed type in a random secondary structure with n nucleotides, as defined by the Knudsen–Hein SCFG. We analyzed the distribution of $\mathbb{X}_n$ and our main result is the following set of relations between the expected number of motifs. Surprisingly, these relations do not depend on the grammar probabilities.

Theorem 1 *For $p_i, q_i > 0$,*

(i) $\mathbb{E}(\mathbb{X}_n^{lb}) = \mathbb{E}(\mathbb{X}_n^{rb})$,
(ii) $\mathbb{E}(\mathbb{X}_n^{m}) = \frac{1}{4}\mathbb{E}(\mathbb{X}_n^{hel})(1 + o(n))$,
(iii) $\mathbb{E}(\mathbb{X}_n^{hp}) = (\mathbb{E}(\mathbb{X}_n^{i}) + \mathbb{E}(\mathbb{X}_n^{m}))(1 + o(n))$,
(iv) $\mathbb{E}(\mathbb{X}_n^{m}) = (\mathbb{E}(\mathbb{X}_n^{lb}) + \mathbb{E}(\mathbb{X}_n^{i}))(1 + o(n))$,
(v) $\mathbb{E}(\mathbb{X}_n^{m,r+1}) < \frac{1}{2}\mathbb{E}(\mathbb{X}_n^{m,r})(1 + o(n))$, $r \geq 2$.

where the superscripts lb, rb, m, hel, hp, and i denote left bulges, right bulges, multibranch loops, helices, hairpins, and internal loops respectively, while $\mathbb{X}_n^{m,r}$ is the number of multibranch loops of degree r in a random secondary structure with n nucleotides.

We find the invariance of these relations under parameter change especially interesting because it illustrates that variation of probability parameters doesn't influence the relative distribution of structural elements in the expected secondary structure. The relations are a consequence of explicit formulas for the corresponding expectations. These, in turn, together with the variances, are a corollary of a central limit law for each of these random variables. More precisely, we have the following result.

Theorem 2 *Let $\mathbb{X}_n$ be the number of base pairs, or helices, or loops of a fixed type in a random secondary structure with n nucleotides. If the probabilities are such that $f(p_1, p_2, p_3) \neq 0$ for a certain function f, then there exist nonzero constants μ and σ such that the normalized random variables*

$$\mathbb{X}_n^* = \frac{\mathbb{X}_n - \mu n}{\sqrt{n\sigma^2}}$$

converge in distribution to a Gaussian variable with a speed of convergence $O\left(\frac{1}{\sqrt{n}}\right)$. That is, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{X}_n^* < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{c^2}{2}}\, dc$$

and

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\mathbb{X}_n^* < x\right) - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{c^2}{2}}\, dc \right| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

The constants μ from Theorem 2 are given as functions of the probabilities in Sect. 5 for all motifs. The function f which appears in the conditions of Theorem 2 is discussed in Sect. 5.6, where we explain why $f(p_1, p_2, p_3) \neq 0$ for all probabilities except for a set of measure zero, so that the result holds for almost all choices of probabilities. Theorem 2 is proved in Sect. 5, where the different types of motifs are considered separately. The proof is based on singularity analysis of bivariate generating functions. In the following section, we illustrate this method by obtaining the asymptotic estimate for the coefficients of $S(z)$.

We analyze here the typical loop composition of a secondary structure generated by the Knudsen–Hein SCFG. This work joins the rich literature dedicated to the analysis of features of RNA secondary structures. In the early work (Waterman 1978; De Chaumont and Viennot 1984) the analysis was done assuming uniform distribution. The growth rate of substructures was first addressed numerically by Hofacker et al. (1998) where various statistical properties were computed for random sequences with lengths up to $n = 100$. These included the mean number of base pairs and of helices/loops as well as the average loop degree, helix length, and loop size. Subsequent work (Fontana et al. 2004) gave exact asymptotics for these and other characteristics of RNA secondary structures. Asymptotics of loops in more general, k-noncrossing, structures which allow pseudoknots were given in (Nebel et al. 2011). Refined asymptotics for the number of substructures in secondary structures of fixed order was computed by Nebel (2002a).

The paper (Nebel 2004) provides related results by considering a Bernoulli model of RNA folding where two bases pair with probability p and computes the asymptotic equivalents for the averaged number of motifs and further parameters, all depending on p. Clote et al. (2012) showed that the expected $5'$–$3'$ distance for structures with n nucleotides both for the uniform and Bernoulli model is bounded by a constant, a result analogous to the property of the Knudsen–Hein SCFG proven in Theorem 9.

Nebel (2003) designed a heavyweight SCFG to create a method for evaluating the reliability of a structure predicted by the NNTM. For that purpose he computed the asymptotic frequency of different structural elements in the SCFG model with numerical, not symbolic probability parameters, obtained by training the grammar on large subunit ribosomal RNA. The energy distribution in a different heavyweight SCFG model designed to mirror the NNTM was studied by Nebel and Scheid (2011). Evaluation of the effects of disturbing the sampling probabilities in this model on the sample was done in (Scheid and Nebel 2012). It was determined that absolute errors drastically affect the sample, while relative errors do not.

3 Preliminaries

A secondary structure of length n is a graph with vertex set $\{1, 2, 3, \ldots, n\}$, whose edge set consists of the edges $\{(k, k+1) : 1 \leq k \leq n-1\}$, together with a collection of edges B called base pairs which satisfies the following conditions. For $(i, j), (k, l) \in B$,

1. $j - i > \theta$ for some threshold $\theta > 0$,
2. $i \neq l$ and $(i = k \Leftrightarrow j = l)$,
3. $i < k < j \Rightarrow i < k < l < j$.

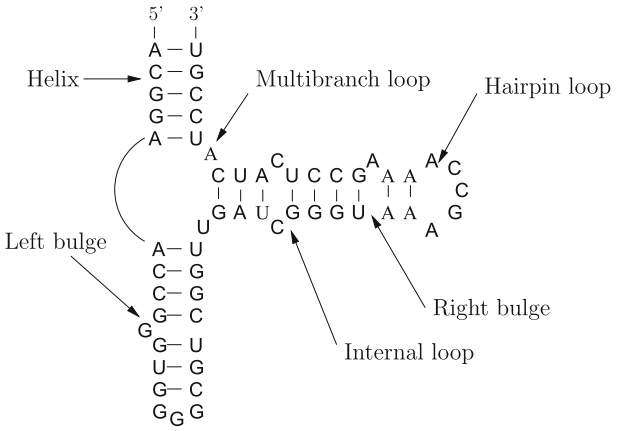

Fig. 1 Helices and different types of loops in RNA secondary structures

The first condition reflects the fact that due to steric constraints, each hairpin in the secondary structure has to contain at least θ unpaired nucleotides. The second condition implies that each vertex (i.e. nucleotide) can belong to at most one base pair. Finally, the third condition excludes pseudoknots which are often considered to be a part of the tertiary structure of the RNA molecule and requires that two edges (i, j) and (k, l) in B with $i < k$, either define separate domains (when $j < k$) or are nested (when $j < l$). All secondary structures consist of the following basic motifs illustrated in Fig. 1. A helix is a set of contiguous nested base pairs. A hairpin is a sequence of consecutive single-stranded nucleotides closed by a single base pair. A bulge loop interrupts helices by having unpaired nucleotides in a single strand. It can be left or right, depending on the side on which the single stranded nucleotides appear. An internal loop separates two helices by having unpaired nucleotides on both strands, while a multibranch loop, or a multiloop, has three or more helices radiating from it. The single stranded nucleotides that are not enclosed by a base pair form an exterior loop.

RNA secondary structures can be modeled using context-free grammars (see Durbin et al. 1998). There are two types of probability parameters: transmission probabilities, which are used for generation of the base pairs in the secondary structure, and emission probabilities, which are used to generate the underlying sequence of nucleotides. The grammar defines a probability measure on the set of all pairs (str, seq) of secondary structures str and RNA sequences seq of same length. Then the basic way to predict a structure for a given RNA sequence seq is to use the CYK algorithm (see Durbin et al. 1998) to compute the most probable pair (str, seq).

The Knudsen–Hein grammar which is used in the RNA secondary structure prediction program Pfold consists of nonterminal symbols $\{S, L, F\}$, terminal symbols $\{d, d', s\}$ and the rules

$$\begin{aligned} S &\to LS \quad (p_1) \quad \text{or} \quad L \quad (q_1) \\ L &\to dFd' \quad (p_2) \quad \text{or} \quad s \quad (q_2) \\ F &\to dFd' \quad (p_3) \quad \text{or} \quad LS \quad (q_3). \end{aligned}$$

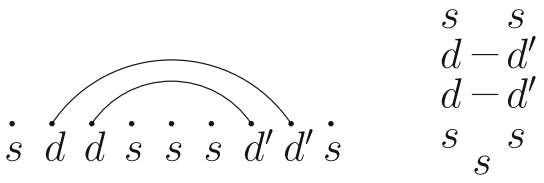

Fig. 2 A simple hairpin loop whose probability is $p_1^3 p_2 p_3 q_1^2 q_2^5 q_3$

The numbers $p_i, q_i, i = 1, 2, 3$ listed in parentheses are the transmission probabilities for the production rules. They satisfy $p_i + q_i = 1$ and $p_i, q_i > 0$ and depend on the structures on which the grammar is trained.

This grammar in non-ambiguous and each derivation corresponds to a unique secondary structure in which $j - i > 2$ for every base pair (i, j). That is why in this paper by a secondary structure we will mean all graphs that satisfy the conditions in the definition of secondary structure for $\theta = 2$. The terminal symbols d and d' correspond to left and right end nucleotides in a base pair, while s corresponds to a single stranded nucleotide. Since secondary structures do not have pseudoknots, specifying the left and right ends of base pairs completely determines the whole structure.

Example 1 The simple hairpin given in Fig. 2 is derived in the following way:

$$S \overset{p_1}{\Rightarrow} LS \overset{q_2}{\Rightarrow} sS \overset{p_1}{\Rightarrow} sLS \overset{p_2}{\Rightarrow} sdFd'S \overset{p_3}{\Rightarrow} sddFd'd'S \overset{q_3}{\Rightarrow} sddLSd'd'S \overset{q_2}{\Rightarrow}$$

$$sddsSd'd'S \overset{p_1}{\Rightarrow} sddsLSd'd'S \overset{q_2}{\Rightarrow} sddssSd'd'S \overset{q_1}{\Rightarrow} sddssLd'd'S \overset{q_2}{\Rightarrow}$$

$$sddsssd'd'S \overset{q_1}{\Rightarrow} sddsssd'd'L \overset{q_2}{\Rightarrow} sddsssd'd's.$$

A stochastic grammar induces a probability distribution on the entire language if the sum of the probabilities of all the derivations is equal to 1. For each nonterminal symbol N, let $N(z)$ be the probability generating function of all secondary structures that can be generated starting from N, where z records the number of nucleotides. In particular, if $n(M)$ is the number of nucleotides in a secondary structure M, we define

$$S(z) = \sum_{S \overset{*}{\Rightarrow} M} p(M) z^{n(M)}$$

where $p(M)$ denotes the probability of the derivation of M and the sum is over all secondary structures. We can determine $S(z)$ by using a technique known as the Delest-Schützenberger-Viennot (DSV) method (Schützenberger 1963). This method has already been applied for studying properties of secondary structures (e.g. Lorenz et al. 2008; Clote et al. 2009). We define $L(z)$ and $F(z)$ to be

$$L(z) = \sum_{L \overset{*}{\Rightarrow} M} p(M) z^{n(M)}, \quad F(z) = \sum_{F \overset{*}{\Rightarrow} M} p(M) z^{n(M)},$$

where the sum is taken over all derivations M that can be obtained starting from the nonterminals L and F, respectively, and $p(M)$ denotes the probability of the derivation M. Through this technique the grammar can be converted into equations involving the generating functions $S(z)$, $L(z)$, $F(z)$. We get

$$\begin{aligned} S(z) &= p_1 L(z)S(z) + q_1 L(z) \\ L(z) &= p_2 z^2 F(z) + q_2 z \\ F(z) &= p_3 z^2 F(z) + q_3 L(z)S(z). \end{aligned} \tag{1}$$

Eliminating $L(z)$ and $F(z)$, we get

$$p_2 q_3 z^2 S(z)^2 - (1 - p_1 q_2 z)(1 - p_3 z^2)S(z) + q_1 q_2 z(1 - p_3 z^2) = 0.$$

Since $S(z)$ is a probabilistic generating function, it has a radius of convergence at least 1. Together with $S(0) = 0$, this implies that

$$S(z) = \frac{(1 - p_1 q_2 z)(1 - p_3 z^2) - \sqrt{(1 - p_1 q_2 z)^2 (1 - p_3 z^2)^2 - 4 p_2 q_1 q_2 q_3 z^3 (1 - p_3 z^2)}}{2 p_2 q_3 z^2} \tag{2}$$

To determine when this grammar generates a probabilistic language, we find when $S(1) = 1$. The condition

$$\frac{(1 - p_1 q_2)(1 - p_3) - \sqrt{(1 - p_1 q_2)^2 (1 - p_3)^2 - 4 p_2 q_1 q_2 q_3 (1 - p_3)}}{2 p_2 q_3} = 1 \tag{3}$$

is equivalent to

$$|p_2 - q_1 q_2| = q_1 q_2 - p_2. \tag{4}$$

Recalling that $p_2 = 1 - q_2$, this reduces to

$$(1 + q_1) q_2 \geq 1. \tag{5}$$

4 Singularity analysis

The total probability of all structures with n nucleotides is given by $[z^n]S(z)$, the coefficient of z^n in $S(z)$. This result will be needed later, so we derive it here as our basic example of asymptotic analysis related to this grammar. We use the following theorem of Flajolet and Odlyzko (1990) to determine the asymptotic growth of the coefficients of $S(z)$.

Theorem 3 (Flajolet and Odlyzko 1990) *Assume that $S(z)$ has a singularity at $z = \rho > 0$, is analytic in the region $\Delta \setminus \{\rho\}$, depicted in Fig. 3, and that as $z \to \rho$ in Δ, $S(z) \sim K(1 - z/\rho)^c$, for some constants $K \neq 0$ and $c \neq 0, 1, 2, \ldots$.*

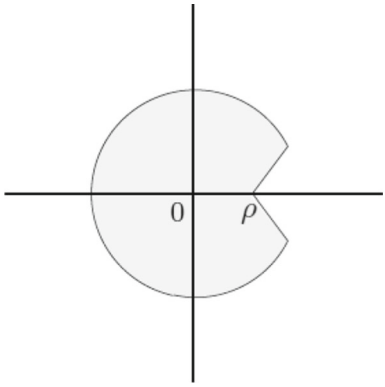

Fig. 3 The function $S(z)$ needs to be analytic at all points in a region Δ of the depicted shape except at ρ

Then, as $n \to \infty$,

$$[z^n]S(z) \sim \frac{K}{\Gamma(-c)} n^{-c-1} \rho^{-n},$$

where $\Gamma(z)$ *denotes the classical gamma function.*

Set

$$R(z) = (1 - p_1 q_2 z)^2 (1 - p_3 z^2) - 4 p_2 q_1 q_2 q_3 z^3. \tag{6}$$

From the explicit formula for $S(z)$ given in (2), we see that the singularities of $S(z)$ are the zeros of the polynomial $P(z) = (1 - p_3 z^2) R(z)$, and in fact the dominant singularity is a root of $R(z)$, which follows from the following two lemmas.

Lemma 1 *If* $p_i, q_i > 0$, *one of the roots of* $R(z)$ *of smallest modulus is a positive real number.*

Proof Since $S(z)$ is a probability generating function, it has a radius of convergence at least 1. By the Pringsheim's theorem, the fact that the coefficients of $S(z)$ are positive implies that it has a positive real singularity equal to its radius of convergence. From (2), we see that this singularity must be a root of the polynomial

$$(1 - p_1 q_2 z)^2 (1 - p_3 z^2)^2 - 4 p_2 q_1 q_2 q_3 z^3 (1 - p_3 z^2) = R(z)(1 - p_3 z^2).$$

Since $R(0) = 1 > 0$ and $R(1/\sqrt{p_3}) < 0$, $R(z)$ has a real zero in the interval $(0, 1/\sqrt{p_3})$. Therefore the smallest positive real singularity of $S(z)$ must come from the zeros of $R(z)$, which implies that among the zeros of smallest modulus of $R(z)$, one is positive and real. □

From now on, let ρ_0 be the root with smallest modulus of $R(z)$ which is a positive real number. The following properties of ρ_0 will be used in the proofs that follow.

Lemma 2 ρ_0 *is the unique root of* $R(z)$ *on the circle* $\{z : |z| = \rho_0\}$. *Moreover,*

$$1 < \rho_0 < \min\{1/p_1 q_2, 1/\sqrt{p_3}\} \quad \textit{and} \quad R'(\rho_0) < 0.$$

Proof This is a corollary of the very general Drmota–Lalley–Woods theorem. For completeness, we present an elementary proof that works for our case. In the proof of Lemma 1, we have already shown that $\rho_0 < 1/\sqrt{p_3}$. The fact that $\rho_0 < 1/p_1q_2$ also follows from $R(0) > 0$ and $R(1/p_1q_2) < 0$. Suppose that $R(z)$ has two complex roots $w, \bar{w}$, with $|w| = \rho_0$. Then, by the triangle inequality,

$$|1 - p_1q_2w| > 1 - p_1q_2\rho_0 > 0 \qquad \text{and} \qquad \left|1 - p_3w^2\right| > 1 - p_3\rho_0^2 > 0,$$

where the inequalities are strict because w is not real. This contradicts

$$\begin{aligned}|1 - p_1q_2w|^2 \left|1 - p_3w^2\right| &= 4p_2q_1q_2q_3|w|^3 \\ &= 4p_2q_1q_2q_3\rho_0^3 = (1 - p_1q_2\rho_0)^2(1 - p_3\rho_0^2).\end{aligned}$$

Similarly, we get a contradiction if we assume that $R(-\rho_0) = 0$. Lastly, we compute

$$R'(z) = -2p_1q_2(1 - p_1q_2z)(1 - p_3z^2) - 2p_3z(1 - p_1q_2z)^2 - 12p_2q_1q_2q_3z^2,$$

from where it is clear that $R'(\rho_0) < 0$. □

As a consequence, if we set

$$Q(z) = \frac{(1 - p_1q_2z)(1 - p_3z^2)}{2p_2q_3z^2}$$

and

$$P(z) = P_1(z)\left(1 - \frac{z}{\rho_0}\right),$$

then

$$S(z) - Q(\rho_0) = \frac{-\sqrt{P_1(\rho_0)}}{2p_2q_3\rho_0^2}\left(1 - \frac{z}{\rho_0}\right)^{1/2} + O\left(1 - \frac{z}{\rho_0}\right)$$

when $z \to \rho_0$.

The coefficients in the expansion of $S(z)$ are the same as in the expansion of $S(z) - Q(\rho)$, except for the first one. From Theorem 3, we get

$$[z^n]S(z) \sim -\frac{\sqrt{(1 - p_1q_2\rho_0)(1-p_3\rho_0^2)(3-p_1q_2\rho_0-p_3\rho_0^2-p_1p_3q_2\rho_0^3)}}{2p_2q_3\rho_0^2\Gamma(-1/2)}n^{-3/2}\rho_0^{-n}. \tag{7}$$

5 Asymptotic distributions of substructures

In this section we prove central limit theorems for various RNA secondary structure motifs for generic choices of the grammar probabilities. We will use the following theorem (Flajolet and Sedgewick 2009 Theorem IX.12) which we state specialized for our purposes.

Theorem 4 (Flajolet and Sedgewick 2009) *Let $G(z,u)$ be a function that is bivariate analytic at $(z,u)=(0,0)$ and has non-negative coefficients and let $\mathbb{X}_n$ be a random variable such that*

$$\mathbb{P}(\mathbb{X}_n = k) = \frac{[z^n u^k]G(z,u)}{[z^n]G(z,1)}.$$

If the technical conditions $(i)-(iii)$ listed below are satisfied, then there exist constants μ and σ such that the normalized random variable

$$\mathbb{X}_n^* = \frac{\mathbb{X}_n - \mu n}{\sqrt{n\sigma^2}}$$

converges in distribution to a Gaussian variable with a speed of convergence $O\left(\frac{1}{\sqrt{n}}\right)$. The technical conditions are

(i) There exist functions A, B, C analytic in a domain $\mathcal{D}=\{|z|<r\}\times\{|u-1|<\epsilon\}$ such that

$$G(z,u) = A(z,u) + B(z,u)C(z,u)^{1/2}$$

for all $(z,u)\in\{|z|<r_0\}\times\{|u-1|<\epsilon\}$ for some $r_0<r$. Furthermore, assume that in $|z|<r$, there exists a unique root z_0 of the equation $C(z,1)=0$ and that $B(z_0,1)\neq 0$,

(ii) $C_{1,0}C_{0,1}\Big|_{z=z_0,u=1} \neq 0$, *where* $C_{i,j} = \frac{\partial^{i+j}}{\partial z^i \partial u^j} C$,

(iii)

$$z_0C_{1,0}^2C_{0,2} - 2z_0C_{1,0}C_{1,1}C_{0,1} + z_0C_{2,0}C_{0,1}^2 + C_{0,1}^2C_{1,0} + z_0C_{0,1}C_{1,0}^2\Big|_{z=z_0,u=1} \neq 0. \tag{8}$$

The constants μ and σ are given by:

$$\mu = \frac{C_{0,1}}{z_0C_{1,0}}\Big|_{z=z_0,u=1} \tag{9}$$

$$\sigma^2 = \frac{z_0C_{1,0}^2C_{0,2} - 2z_0C_{1,0}C_{1,1}C_{0,1} + z_0C_{2,0}C_{0,1}^2 + C_{0,1}^2C_{1,0} + z_0C_{0,1}C_{1,0}^2}{z_0^2C_{1,0}^3}\Big|_{z=z_0,u=1} \tag{10}$$

Remark 1 By analyzing the distribution of each motif separately, we are able to provide proofs which are self-contained. Alternatively, one could consider the joint distribution of all the motifs, and likewise prove that it is asymptotically Gaussian based on a multivariate analog of Theorem 4. Towards this end, the main result of Drmota (1997) gives sufficient conditions for the coefficients of a multivariate function to follow a normal distribution. Unfortunately, those conditions do not apply directly to our analysis since this system is not strongly-connected. Nonetheless, as recognized by Denise et al. (2010), the conditions can often be relaxed when context-free languages are considered and the conclusion still holds, which is the case here. However, because the details of adapting these arguments to our purposes are technical, we present the results individually, rather than as a joint distribution.

5.1 Base pairs

To find the distribution of base pairs, we first find the bivariate generating function $S(z, u)$ where u marks the base pairs. A base pair is added precisely when the rules $L \to dFd'$ and $F \to dFd'$ are used. So, $S(z, u)$ is the solution of the system

$$\begin{aligned} S(z,u) &= p_1 L(z,u)S(z,u) + q_1 L(z,u) \\ L(z,u) &= p_2 z^2 u F(z,u) + q_2 z \\ F(z,u) &= p_3 z^2 u F(z,u) + q_3 L(z,u)S(z,u). \end{aligned} \tag{11}$$

Similarly as before, we can find an explicit formula for $S(z, u)$:

$$S(z,u) = Q(z,u) - \frac{\sqrt{C^{bp}(z,u)}}{2p_2 q_3 z^2 u}$$

where

$$\begin{aligned} Q(z,u) &= \frac{(1-p_1q_2z)(1-p_3z^2u)}{2p_2q_3z^2u}, \\ C^{bp}(z,u) &= (1-p_1q_2z)^2(1-p_3z^2u)^2 - 4p_2q_1q_2q_3z^3u(1-p_3z^2u). \end{aligned} \tag{12}$$

Theorem 5 *Let $\mathbb{X}_n^{bp}$ be a random variable counting the number of basepairs in a secondary structure with n nucleotides. If the probabilities $p_i, q_i > 0$, $1 \le i \le 3$ are such that the polynomial $C^{bp}(z, u)$ given in (12) satisfies the condition (8), then $\mathbb{X}_n^{bp}$ after standardization converges to a Gaussian variable. The mean and standard deviation of $\mathbb{X}_n^{bp}$ are asymptotically linear in n. In particular,*

$$\mathbb{E}(\mathbb{X}_n^{bp}) \sim \frac{\alpha}{\gamma} n,$$

where

$$\alpha = 1 - p_1q_2\rho_0, \;\; \gamma = 3 - p_1q_2\rho_0 - p_3\rho_0^2 - p_1p_3q_2\rho_0^3. \tag{13}$$

The first order approximation of the standard deviation is given by (10) for $C = C^{bp}$ and $z_0 = \rho_0$.

Proof The random variables associated to $G(z,u) = z^2 S(z,u)$ are the same as the ones associated to $S(z,u)$, only shifted in index. So, we will work with the function $G(z,u)$ and we will prove that it satisfies the conditions in Theorem 4. The functions $A(z,u) = z^2 Q(z,u)$, $B(z,u) = -\frac{1}{2p_2q_3u}$, and $C(z,u) = C^{bp}(z,u)$ are clearly analytic in the domain $\mathbb{C} \times \{|u-1| < \epsilon\}$ for small $\epsilon > 0$. Using Lemma 2, we get

$$\begin{aligned} C_{1,0}(\rho_0, 1) &= R'(\rho_0)(1 - p_3\rho_0^2) < 0 \\ C_{0,1}(\rho_0, 1) &= -p_3\rho_0^2(1 - p_3\rho_0^2)(1 - p_1q_2\rho_0)^2 - 4p_2q_1q_2q_3\rho_0^3(1 - p_3\rho_0^2) \\ &= -4p_2q_1q_2q_3\rho_0^3 < 0. \end{aligned}$$

So, condition (ii) is satisfied. By the analytic implicit function theorem, there exists an analytic function $\rho(u)$ defined on some neighborhood of ρ_0 such that $C^{bp}(z,u) = 0$ for (z,u) in a small polydisc $\Delta(\rho_0, 1, \epsilon)$ if and only if $z = \rho(u)$. Since, by Lemma 2, $\rho(1) = \rho_0 > 1$, ϵ can be chosen so that $\rho(u) > 1$.

We claim that if $|u-1| < \epsilon$, $z = \rho(u)$ is the root of smallest modulus of $C^{bp}(z,u)$ as a polynomial in z. There is a neighborhood of $u = 1$ such that $z = \rho(u)$ is the unique zero of smallest modulus of $C^{bp}(z,u)$. Otherwise, there exists a sequence $u_n \to 1$ and $\xi_n \neq \rho(u_n)$ with $|\xi_n| \leq |\rho(u_n)|$ and $C^{bp}(\xi_n, u_n) = 0$. By passing to a subsequence, which we still denote by (ξ_n), we obtain that there exists some ξ_0 such that $\lim_{n\to\infty} \xi_n = \xi_0$. By continuity, $C^{bp}(\xi_0, 1) = 0$ and $|\xi_0| \leq \rho(1)$. Hence, by uniqueness, $\xi_0 = \rho(1)$. This contradicts the uniqueness of the solution $z = \rho(u)$ of $C^{bp}(z,u) = 0$ in a neighborhood of $u = 1$ guaranteed by the implicit function theorem.

Finally, choose ϵ to be small enough so that $R(z)$ has a unique zero in $|z| < \rho_0 + \epsilon$. Setting $r = \rho_0 + \epsilon$ and $r_0 = 1$, such that $R(z)$ has a unique zero make condition (i) satisfied. Indeed, for $\mathcal{D} = \{|z| < r\} \times \{|u-1| < \epsilon\}$, $C(z,1) = R(z)(1 - p_3z^2)$ has a unique zero in $\mathcal{D}$ and clearly $B(\rho_0, 1) \neq 0$. □

5.2 Helices

A helix is started when the rule $L \to dFd'$ is used. If u marks the number of helices in the secondary structure, the relation between the probability generating functions is

$$\begin{aligned} S(z,u) &= p_1L(z,u)S(z,u) + q_1L(z,u), \\ L(z,u) &= p_2z^2uF(z,u) + q_2z, \\ F(z,u) &= p_3z^2F(z,u) + q_3L(z,u)S(z,u), \end{aligned}$$

and therefore,

$$S(z,u) = Q(z,u) - \frac{\sqrt{C^h(z,u)}}{2p_2q_3z^2u},$$

where

$$Q(z,u) = \frac{(1-p_1q_2z)(1-p_3z^2)}{2p_2q_3z^2u},$$
$$C^h(z,u) = (1-p_1q_2z)^2(1-p_3z^2)^2 - 4p_2q_1q_2q_3z^3u(1-p_3z^2). \quad (14)$$

Theorem 6 *Let $\mathbb{X}_n^{hel}$ be the number of helices in a random secondary structure with n nucleotides. If the probabilities $p_i, q_i > 0$, $1 \le i \le 3$ are such that the polynomial $C^h(z,u)$ given in (14) satisfies the condition (8), then $\mathbb{X}_n^h$ after standardization converges to a Gaussian variable. In particular,*

$$\mathbb{E}(\mathbb{X}_n^{hel}) \sim \frac{\alpha\beta}{\gamma}n,$$

where α and γ are given by (13) and

$$\beta = 1 - p_3\rho_0^2. \quad (15)$$

The first order approximation of the standard deviation is given by (10) for $C = C^{hel}$ and $z_0 = \rho_0$.

Proof Similarly as in the proof of Theorem 5, the conditions (i) and (ii) from Theorem 4 are satisfied for the function $G(z,u) = z^2S(z,u)$. Condition (iii) is satisfied by assumption. □

5.3 Loops

In this subsection, let $S(z,x,y,u,v,w)$ be the multivariable probability generating function for RNA structures where x marks hairpin loops, y marks multibranch loops, u marks left bulges, v marks right bulges, and w marks internal loops.

A loop starts exactly when a helix ends, so each application of the rule $F \to LS$ starts one loop. The loop started will be a hairpin loop if this rule is followed by $LS \overset{*}{\Rightarrow} s^n$, $n \ge 2$. To find the probabilities of a hairpin loop of length $n \ge 2$ we note that

$$\begin{aligned} P(LS \overset{*}{\Rightarrow} s^n) &= P(L \Rightarrow s)P(S \overset{*}{\Rightarrow} s^{n-1}) \\ &= q_2P(S \Rightarrow LS)P(L \Rightarrow s)P(S \overset{*}{\Rightarrow} s^{n-2}) \\ &= p_1q_2^2P(S \overset{*}{\Rightarrow} s^{n-2}) \\ &= q_1q_2^2(p_1q_2)^{n-2} \end{aligned}$$

Therefore the probability generating function for the hairpin loops that could be formed is

$$\sum_{n=2}^{\infty} q_1q_2^2(p_1q_2)^{n-2}z^n = \frac{q_1q_2^2z^2}{1-p_1q_2z} =: H_h.$$

Right bulges are formed when the derivation that follows is of the form $LS \stackrel{*}{\Rightarrow} dFd's^l$, $l \geq 1$. Their probability is

$$P(LS \stackrel{*}{\Rightarrow} dFd's^l) = P(L \to dFd')P(S \stackrel{*}{\Rightarrow} s^l) = p_2q_1q_2(p_1q_2)^{l-1}$$

and their contribution to the generating function is

$$\sum_{l=1}^{\infty} z^{l+2} p_2q_1q_2(p_1q_2)^{l-1} F = \frac{p_2q_1q_2z^3}{1 - p_1q_2z} F =: H_bF.$$

Similarly, left bulges are formed by applications of rules that yield $LS \stackrel{*}{\Rightarrow} s^kdFd'$, $k \geq 1$. The probability of the left bulges together with all successive derivations is

$$\begin{aligned} P(LS \stackrel{*}{\Rightarrow} s^kdFd') &= P(LS \stackrel{*}{\Rightarrow} s^kS)P(S \to L)P(L \to dFd') \\ &= p_2q_1P(LS \stackrel{*}{\Rightarrow} s^kS) \\ &= p_2q_1P(LS \stackrel{*}{\Rightarrow} sS)P(S \to LS)P(LS \stackrel{*}{\Rightarrow} s^{k-1}S) \\ &= p_2q_1q_2(p_1q_2)^{k-1}. \end{aligned}$$

The part of the generating function that corresponds to the left bulges is

$$\sum_{k=2}^{\infty} z^{k+2} p_2q_1q_2(p_1q_2)^{k-1} F = \frac{p_2q_1q_2z^3}{1 - p_1q_2z} F = H_bF.$$

Internal loops are created when the rule $F \to LS$ is followed by $LS \stackrel{*}{\Rightarrow} s^kdFd's^l$, for some $k, l \geq 1$.

$$\begin{aligned} P(LS \stackrel{*}{\Rightarrow} s^kdFd's^l) &= P(LS \stackrel{*}{\Rightarrow} s^kS)P(S \to LS)P(LS \stackrel{*}{\Rightarrow} dFd's^l) \\ &= q_2(p_1q_2)^{k-1}p_1p_2q_1q_2(p_1q_2)^{l-1} \\ &= p_1p_2q_1q_2^2(p_1q_2)^{k-1}(p_1q_2)^{l-1} \end{aligned}$$

and their contribution to the generating function is

$$\sum_{k=1}^{\infty}\sum_{l=1}^{\infty} z^{k+l+2} p_1p_2q_1q_2^2(p_1q_2)^{k-1}(p_1q_2)^{l-1} F = \frac{p_1p_2q_1q_2^2z^4}{(1 - p_1q_2z)^2} F =: H_iF.$$

The remaining part of LS corresponds to the substructures that begin with a multi-branch loop. Their contribution is

$$H_m := LS - H_h - 2H_bF - H_iF.$$

Using this, the translation of the grammar rules yields the system

$$\begin{aligned} S &= p_1LS + q_1L, \\ L &= p_2z^2F + q_2z, \\ F &= p_3z^2F + q_3(xH_h + uH_bF + vH_bF + wH_iF + yH_m), \end{aligned}$$

and, by eliminating F and L, we get that $S(z, x, y, u, v, w)$ is a solution to the quadratic equation

$$\begin{aligned} p_2q_3z^2yS^2 &+ (p_1p_2q_3z^2xH_h - p_1p_2q_3z^2yH_h + p_1q_2zH_e - H_e)S \\ &+ (p_2q_1q_3z^2xH_h - p_2q_1q_3z^2yH_h + q_1q_2zH_e) = 0 \end{aligned}$$

where

$$H_e := 1 - p_3z^2 + q_3(y-u)H_b + q_3(y-v)H_b + q_3(y-w)H_i.$$

Theorem 7 *Let $\mathbb{X}_n^{hp}$, $\mathbb{X}_n^{lb}$, $\mathbb{X}_n^{rb}$, $\mathbb{X}_n^{i}$, and $\mathbb{X}_n^{m}$ be the number of hairpin loops, left bulges, right bulges, internal loops, and multibranch loops in a random secondary structure with n nucleotides, respectively. For $\star \in \{hp, lb, rb, i, m\}$, if the probabilities $p_i, q_i > 0$, $1 \le i \le 3$ are such that a certain polynomial $C^\star(z, u)$ satisfies the condition (8), then $\mathbb{X}_n^\star$ after standardization converges to a Gaussian variable. The approximate expectations are explicitly given by*

$$\begin{aligned} \mathbb{E}(\mathbb{X}_n^{hp}) &\sim \frac{(1 + p_1q_2\rho_0)\alpha\beta}{4\gamma}n, \\ \mathbb{E}(\mathbb{X}_n^{lb}) = \mathbb{E}(\mathbb{X}_n^{rb}) &\sim \frac{\alpha^2\beta}{4\gamma}n, \\ \mathbb{E}(\mathbb{X}_n^{i}) &\sim \frac{p_1q_2\rho_0\alpha\beta}{4\gamma}n, \\ \mathbb{E}(\mathbb{X}_n^{m}) &\sim \frac{\alpha\beta}{4\gamma}n, \end{aligned}$$

where α, β, and γ are given by (13) and (15). The first order approximations of the standard deviations are given by (10) for $C = C^\star(z, u)$ and $z_0 = \rho_0$.

Proof By setting $y = u = v = w = 1$, for hairpins we get that

$$S(z, x) = Q(z, x) - \frac{\sqrt{C^{hp}(z, x)}}{2p_2q_3z^2(1 - p_1q_2z)},$$

where

$$\begin{aligned} Q(z, x) &= \frac{(1 - p_1q_2z)^2(1 - p_3z^2) - p_1p_2q_1q_2^2q_3z^4(x - 1)}{2p_2q_3z^2(1 - p_1q_2z)}, \\ C^{hp}(z, x) &= \left(p_1p_2q_1q_2^2q_3z^4(x - 1) - (1 - p_1q_2z)^2(1 - p_3z^2)\right)^2 \\ &\quad - 4p_2q_1q_2q_3z^3(1-p_1q_2z)\left(p_2q_1q_2q_3z^3(x-1)+(1- p_1q_2z)(1-p_3z^2)\right). \end{aligned}$$

To prove the claim for $\mathbb{X}_n^{hp}$, we work with the function $G(z, x) = z^2S(z, x)$. The functions in condition (i) of Theorem 4, are $A(z, x) = Q(z, x)z^2$, $B(z, x) = -\frac{1}{2p_2q_3(1-p_1q_2z)}$, and $C^{hp}(z, x)$. They are all analytic in some polydisc around $(0, 1)$.Since

$$C^{hp}(z,1) = (1-p_1q_2z)^4(1-p_3z^2)^2 - 4p_2q_1q_2q_3z^3(1-p_1q_2z)^2(1-p_3z^2)$$
$$= R(z)(1-p_1q_2z)^2(1-p_3z^2),$$

it follows from Lemma 2 that the smallest zero of $C^{hp}(z,1)$ is ρ_0 and that

$$C^{hp}_{1,0}(\rho_0,1) = R'(\rho_0)(1-p_1q_2\rho_0)^2(1-p_3\rho_0^2),$$
$$C^{hp}_{0,1}(\rho_0,1) = -2p_1p_2q_1q_2^2q_3\rho_0^4(1-p_1q_2\rho_0)^2(1-p_3\rho_0^2)$$
$$-4p_2^2q_1^2q_2^2q_3^2\rho_0^6(1-p_1q_2\rho_0)$$

are both negative.

By setting $x = y = v = w = 1$, for left bulges we get that

$$S(z,u) = Q(z,u) - \frac{\sqrt{C^{lb}(z,u)}}{2p_2q_3z^2(1-p_1q_2z)},$$

where

$$Q(z,u) = \frac{(1-p_1q_2z)(1-p_3z^2) + p_2q_1q_2q_3z^3(1-u)}{2p_2q_3z^2},$$
$$C^{lb}(z,u) = \Big((1-p_1q_2z)^2(1-p_3z^2) + p_2q_1q_2q_3z^3(1-u)(1-p_1q_2z)\Big)^2$$
$$-4p_2q_1q_2q_3z^3(1-p_1q_2z)\Big((1-p_1q_2z)(1-p_3z^2)+p_2q_1q_2q_3z^3(1-u)\Big).$$

To prove the claim for $\mathbb{X}_n^{lb}$, we apply Theorem 4 to $G(z,u) = z^2S(z,u)$. The functions in condition (i) are $A(z,u) = Q(z,u)z^2$, $B(z,u) = -\frac{1}{2p_2q_3(1-p_1q_2z)}$, and $C^{lb}(z,u)$. The conditions of Theorem 4 can be checked as before by using the fact that

$$C^{lb}(z,1) = R(z)(1-p_1q_2z)^2(1-p_3z^2).$$

The proof for right bulges is exactly the same as the one for left bulges and $C^{rb}(z,u) = C^{lb}(z,u)$.

For interior loops, we set $x = y = u = v = 1$ and we get

$$S(z,w) = Q(z,w) - \frac{\sqrt{C^{i}(z,w)}}{2p_2q_3z^2(1-p_1q_2z)},$$

where

$$Q(z,w) = \frac{(1-p_1q_2z)^2(1-p_3z^2) + p_1p_2q_1q_2^2q_3z^4(1-w)}{2p_2q_3z^2(1-p_1q_2z)},$$
$$C^{i}(z,w) = \Big((1-p_1q_2z)^3(1-p_3z^2) + p_1p_2q_1q_2^2q_3z^4(1-w)(1-p_1q_2z)\Big)^2$$
$$-4p_2q_1q_2q_3z^3(1-p_1q_2z)^4(1-p_3z^2)$$
$$-4p_1p_2^2q_1^2q_2^3q_3^2z^7(1-w)(1-p_1q_2z)^2.$$

As in the previous cases one can show that $G(z, w) = z^2 S(z, w)$ satisfies the conditions of Theorem 4 by setting the functions in condition (i) to be $A(z, w) = Q(z, w)z^2$, $B(z, w) = -\frac{1}{2p_2q_3(1-p_1q_2z)}$, and $C^i(z, w)$. Additionally, the factorization

$$C^i(z, 1) = R(z)(1 - p_1q_2z)^4(1 - p_3z^2)$$

is used.

The case for multibranch loops is similar. For completeness, we give the formula for $S(z, y)$.

$$S(z, y) = Q(z, y) - \frac{\sqrt{C^m(z, y)}}{2p_2q_3z^2y(1 - p_1q_2z)^2},$$

where

$$\begin{aligned} Q(z, y) &= \frac{(1 - p_1q_2z)^2(1 - p_3z^2) - 2p_2q_1q_2q_3z^3(1 - y)(1 - p_1q_2z)}{2p_2q_3z^2y}, \\ C^m(z, y) &= \Big((1 - p_1q_2z)^3(1 - p_3z^2) - 2p_2q_1q_2q_3z^3(1 - y)(1 - p_1q_2z)\Big)^2 \\ &\quad - 4p_2q_1q_2q_3z^3y(1 - p_1q_2z)^4(1 - p_3z^2) \\ &\quad + 4p_2^2q_1^2q_2^2q_3^2z^6y(1 - y)(1 - p_1q_2z)^2. \end{aligned}$$

The claim for $\mathbb{X}_n^m$ follows from Theorem 4 for the function $G(z, y) = z^2 S(z, y)$. Then the functions in condition (i) are $A(z, y) = Q(z, y)z^2$, $B(z, y) = -\frac{1}{2p_2q_3y(1-p_1q_2z)^2}$, and $C^m(z, y)$. When checking the conditions, one uses that

$$C^m(z, 1) = R(z)(1 - p_1q_2z)^4(1 - p_3z^2). \qquad \square$$

5.4 Multibranch loops with fixed degree

In this subsection we compute the expected number of multibranch loops of a fixed degree $r \geq 2$. A multibranch loop has a degree r if it contains $r + 1$ base pairs.

Let $r \geq 2$ be fixed. Starting with LS, to get a multibranch loop of degree r with single-stranded segments of lengths $k_0, k_1, \ldots, k_r$ $(k_i \geq 0)$, one needs to apply the rule $S \to LS$ exactly $r - 2 + \sum_{i=0}^r k_i$ times and the rule $S \to L$ exactly once. After this one has $r + \sum_{i=0}^r k_i$ copies of L. Then one applies the rule $L \to s$ exactly $\sum_{i=0}^r k_i$ times to get the single-stranded nucleotides, and the rule $L \to dFd'$ precisely r times to get the r helices. Therefore, if z marks the number of nucleotides and t marks the number of multibranch loops of degree r, the total weight of all substructures with r branches and prescribed lengths of single-stranded segments that can be derived with this process is

$$p_1^{r-2+\sum_{i=0}^r k_i} p_2^r q_1 q_2^{\sum_{i=0}^r k_i} t z^{2r+\sum_{i=0}^r k_i} F^r$$

and the total weight of all substructures starting with a multibranch loops of degree r is

$$\sum_{k_0,k_1,\ldots,k_r\geq 0} p_1^{r-2+\sum_{i=0}^{r} k_i} p_2^r q_1 q_2^{\sum_{i=0}^{r} k_i} t z^{2r+\sum_{i=0}^{r} k_i} F^r = \frac{p_1^{r-2} p_2^r q_1 t z^{2r} F^r}{(1-p_1 q_2 z)^{r+1}}.$$

Translation of the grammar into generating functions yields the system

$$S = p_1 LS + q_1 L, \tag{16}$$
$$L = p_2 z^2 F + q_2 z, \tag{17}$$
$$F = p_3 z^2 F + q_3 \frac{p_1^{r-2} p_2^r q_1 z^{2r} t F^r}{(1-p_1 q_2 z)^{r+1}} + q_3 \left(LS - \frac{p_1^{r-2} p_2^r q_1 z^{2r} F^r}{(1-p_1 q_2 z)^{r+1}} \right). \tag{18}$$

For convenience, set

$$T_r(z) = \frac{p_1^{r-2} p_2^r q_1 q_3 z^{2r}}{(1-p_1 q_2 z)^{r+1}}.$$

Then Eq. (18) can be rewritten as

$$F = p_3 z^2 F + (t-1) T_r F^r + q_3 LS. \tag{19}$$

Multiplying Eqs. (16) and (17) we get

$$LS = p_1 LS(p_2 z^2 F + q_2 z) + q_1 (p_2 z^2 F + q_2 z)^2$$

and hence

$$LS = \frac{q_1 (p_2 z^2 F + q_2 z)^2}{1 - p_1 p_2 z^2 F - p_1 q_2 z}.$$

Substituting back to (19), we get:

$$F = p_3 z^2 F + (t-1) T_r F^r + q_3 \frac{q_1 (p_2 z^2 F + q_2 z)^2}{1 - p_1 p_2 z^2 F - p_1 q_2 z}.$$

which is equivalent to

$$p_1 p_2 z^2 (t-1) T_r F^{r+1} - (t-1)(1-p_1 q_2 z) T_r F^r + \left(p_1 p_2 p_3 z^4 - p_2^2 q_1 q_3 z^4 - p_1 p_2 z^2 \right) F^2 + \left(1 - p_1 q_2 z - p_3 z^2 + p_1 p_3 q_2 z^3 - 2 p_2 q_1 q_2 q_3 z^3 \right) F - q_1 q_2^2 q_3 z^2 = 0.$$

After differentiating with respect to t, we find that $F_t'(z,1)$ is equal to

$$\frac{T_r F^r(z,1)(1-p_1q_2z-p_1p_2z^2F(z,1))}{2p_2z^2[(p_1p_3-p_2q_1q_3)z^2-p_1]F(z,1)+[q_2(p_1p_3-2p_2q_1q_3)z^3-p_3z^2-p_1q_2z+1]} \tag{20}$$

and from here we can easily find the function F at $t=1$, which we will need later. Namely,

$$\begin{aligned} &p_2z^2(p_1-p_1p_3z^2+p_2q_1q_3z^2)F^2(z,1)\\ &-(1-p_1q_2z-p_3z^2+p_1p_3q_2z^3-2p_2q_1q_2q_3z^3)F(z,1)+q_1q_2^2q_3z^2=0 \end{aligned}$$

and hence after simplifications we find that

$$F(z)=\frac{(1-p_1q_2z-p_3z^2+p_1p_3q_2z^3-2p_2q_1q_2q_3z^3)-\sqrt{(1-p_3z^2)R(z)}}{2p_2z^2(p_1-p_1p_3z^2+p_2q_1q_3z^2)}. \tag{21}$$

The solution with the negative sign is chosen because $F(0)=0$. From the explicit formula for $F(z)$, we note that the dominant singularity is again ρ_0. Indeed, $F(z)$ has a positive dominant singularity, since it has positive coefficients and if $z_0<\rho_0$ is a positive solution to the quadratic $p_1-p_1p_3z^2+p_2q_1q_3z^2=0$, we get

$$\begin{aligned} &(1-p_1q_2z+0-p_3z_0^2+p_1p_3q_2z_0^3-2p_2q_1q_2q_3z_0^3)-\sqrt{(1-p_3z_0^2)R(z)}\\ &=(1-p_3z_0^2)(1+p_1q_2z_0)-\sqrt{(1-p_3z_0^2)^2(1+p_1q_2z_0)^2}=0. \end{aligned}$$

Combining (16) and (19) yields

$$S=\frac{p_1}{q_3}(F-p_3z^2F-(t-1)T_rF^r)+q_1(p_2z^2F+q_2z)$$

and hence

$$S_t'(z,1)=\frac{p_1}{q_3}(F_t'(z,1)-p_3z^2F_t'(z,1)-T_rF^r(z,1))+p_2q_1z^2F_t'(z,1). \tag{22}$$

In light of (21), formula (20) simplifies to

$$F_t'(z,1)=\frac{T_rF^r(z,1)(1-p_1q_2z-p_1p_2z^2F(z,1))}{\sqrt{(1-p_3z^2)R(z)}}$$

and plugging this into (22) yields

$$S_t'(z,1)=\frac{T_rF^r(z,1)}{2q_3}\left(\frac{p_1^2p_3q_2z^3-p_1p_3z^2+2p_2q_1q_3z^2-p_1^2q_2z+p_1}{\sqrt{(1-p_3z^2)R(z)}}-p_1\right).$$

Using this expression and Theorem 3, we can estimate the coefficients of $S'_t(z, 1)$:

$$[z^n]S'(z, 1) \sim \frac{K}{\Gamma(1/2)} n^{1/2} \rho_0^{-n}, \tag{23}$$

where

$$K = \frac{p_1^{r-2} q_1 q_2^{r-1} \rho_0^{r-1} (1 - p_3 \rho_0^2)}{4\sqrt{-\rho_0 (1 - p_3 \rho_0^2) R'(\rho_0)} (1 + p_1 q_2 \rho_0)^{r-1}}.$$

Combining this estimate with (7) we get the following theorem.

Theorem 8 *Let $\mathbb{X}_n^{m,r}$ be the number of multibranch loops of degree r in a random secondary structure with n nucleotides. If the probabilities p_i, q_i are all non-zero, then*

$$\mathbb{E}(\mathbb{X}_n^{m,r}) \sim \frac{p_1^{r-2} q_2^{r-2} \rho_0^{r-2} (1 - p_1 q_2 \rho_0)(1 - p_3 \rho_0^2)}{4(1 + p_1 q_2 \rho_0)^{r-1} (3 - p_1 q_2 \rho_0 - p_3 \rho_0^2 - p_1 p_3 q_2 \rho_0^3)} n.$$

Proof The estimate follows from (23), (7), and $\mathbb{E}(\mathbb{X}_n^{m,r}) = \frac{[z^n]S'_t(z,1)}{[z^n]S(z,1)}$. □

5.5 Exterior loop

In this subsection we analyze the branchings of the exterior loop and the 5′–3’ distance. The 5′–3′ distance is defined as the number of nucleotides (paired or single-stranded) enclosed in the exterior loop minus one. Let u be the variable that marks the number of helices in the exterior loop, and let v mark the 5′–3′ distance. The total contribution of all secondary structures with no base pairs in $S(z, u, v)$ is

$$\sum_{n \geq 1} P(S \stackrel{*}{\Rightarrow} s^n) = \sum_{n \geq 1} p_1^{n-1} q_1 q_2^n z^n v^{n-1} = \frac{q_1 q_2 z}{1 - p_1 q_2 z v}.$$

All other structures have $r \geq 1$ helices in the exterior loop. Since

$$P(S \stackrel{*}{\Rightarrow} s^{k_0} d F d' s^{k_1} \cdots d F d' s^{k_r}) = p_1^{r-1+\sum_{i=0}^r k_i} p_2^r q_1 q_2^{\sum_{i=0}^r k_i},$$

the generating function of all structures that have exactly r helices in the exterior loop is given by

$$\sum_{k_0, k_1, \ldots, k_r \geq 0} p_1^{r-1+\sum_{i=0}^r k_i} p_2^r q_1 q_2^{\sum_{i=0}^r k_i} z^{2r+\sum_{i=0}^r k_i} u^r v^{2r-1+\sum_{i=0}^r k_i} F^r(z)$$

which is equal to $\frac{p_1^{r-1}p_2^r q_1 z^{2r} u^r v^{2r-1} F^r(z)}{(1-p_1q_2zv)^{r+1}}$. Therefore $S(z, u, v)$ is given by

$$\begin{aligned} S(z,u,v) &= \frac{q_1q_2z}{1-p_1q_2zv} + \sum_{r\geq 1} \frac{p_1^{r-1}p_2^r q_1 z^{2r} u^r v^{2r-1} F^r(z)}{(1-p_1q_2zv)^{r+1}} \\ &= \frac{q_1q_2z}{1-p_1q_2zv} + \frac{p_2q_1z^2uvF(z)}{(1-p_1q_2zv)(1-p_1q_2zv-p_1p_2z^2uv^2F(z))}. \end{aligned}$$

To compute the expected number of helices in the exterior loops we will need to look at the behavior of $S'_u(z, 1, 1)$ around its dominant singularity. We find that

$$S'_u(z,1,1) = \frac{p_2q_1z^2F(z)}{(1-p_1q_2z-p_1p_2z^2F(z))^2}.$$

Using (21), one can show that $1 - p_1q_2z - p_1p_2z^2F(z) \neq 0$, and so the dominant singularity of $S'_u(z, 1, 1)$ is the same as the dominant singularity of $F(z)$, which was found to be ρ_0. After simplifications of the expansion of $S'_u(z, 1, 1)$, we get that as $z \to \rho_0$,

$$S'_u(z,1,1) \sim -\frac{(1+2p_1q_2\rho_0)\sqrt{-\rho_0R'(\rho_0)(1-p_3\rho_0^2)}}{2p_1q_3\rho_0^2}\left(1-\frac{z}{\rho_0}\right)^{1/2}. \tag{24}$$

Theorem 9 *Let $\mathbb{X}_n^{eh}$ be a random variable counting the number of helices in the exterior loop in a secondary structure with n nucleotides and let $\mathbb{X}_n^{ecd}$ count the 5′–3′ distance. If the probabilities p_i, q_i are all non-zero, then*

$$\begin{aligned} \mathbb{E}(\mathbb{X}_n^{eh}) &\sim 1 + 2p_1q_2\rho_0 \qquad \text{and} \\ \mathbb{E}(\mathbb{X}_n^{ecd}) &\sim \frac{1+5p_1q_2\rho_0-2p_1^2q_2^2\rho_0^2}{1-p_1q_2\rho_0}. \end{aligned}$$

Proof The estimate for $\mathbb{E}(\mathbb{X}_n^{eh})$ follows from (24), (7), Theorem 3, and the fact that $\mathbb{E}(\mathbb{X}_n^{eh}) = \frac{[z^n]S'_u(z,1,1)}{[z^n]S(z,1,1)}$. For $\mathbb{E}(\mathbb{X}_n^{ecd})$, one finds that

$$\begin{aligned} S'_v(z,1,1) = &\frac{p_1q_1q_2^2z^2}{1-p_1q_2z} + \frac{p_2q_1z^2(1-p_1q_2z)(1-p_1q_2z+2p_1q_2z)F(z)}{(1-p_1q_2z)^2(1-p_1q_2z-p_1p_2z^2F(z))^2} \\ &+\frac{p_1p_2^2q_1z^4(1-2p_1q_2z)F(z)^2}{(1-p_1q_2z)^2(1-p_1q_2z-p_1p_2z^2F(z))^2}. \end{aligned}$$

The dominant singularity is again ρ_0, so one proceeds as before to obtain the estimate. □

5.6 The function f in Theorem 2

In this subsection we show that the set of probabilities (p_1, p_2, p_3) for which Theorem 2 does not apply is small in the sense that it has Lebesgue measure zero. Define $V^{bp} = V^{bp}(p_1, p_2, p_3, \rho_0)$ to be

$$\rho_0(C^{bp}_{1,0})^2 C^{bp}_{0,2} - 2\rho_0 C^{bp}_{1,0} C^{bp}_{1,1} C^{bp}_{0,1} + \rho_0 C^{bp}_{2,0}(C^{bp}_{0,1})^2 + (C^{bp}_{0,1})^2 C^{bp}_{1,0} + \rho_0 C^{bp}_{0,1}(C^{bp}_{1,0})^2 \Big|_{\substack{z=\rho_0 \\ u=1}}$$

where C^{bp} is the polynomial that is defined in (12) and appears in the conditions of Theorems 5. Similarly define V^{hel}, V^{hp}, V^{lb}, V^{i}, V^{m} which correspond to the polynomials C^{hel}, C^{hp}, C^{lb}, C^{i}, and C^{m}, which appear in the conditions of Theorems 6 and 7 (since $C^{lb} = C^{rb}$, we do not need to define V^{rb}). Finally, define

$$g(p_1, p_2, p_3, \rho_0) = V^{bp} V^{h} V^{hp} V^{lb} V^{i} V^{m}.$$

Notice that Theorem 2 holds for all $(p_1, p_2, p_3) \in (0, 1)^3$ other than those for which $g(p_1, p_2, p_3, \rho_0) = 0$. Since by Lemma 2 ρ_0 is a root of multiplicity one of the polynomial $R(z)$ for all $(p_1, p_2, p_3) \in (0, 1)^3$ it follows that ρ_0 is an analytic function of (p_1, p_2, p_3) and therefore

$$f(p_1, p_2, p_3) = g(p_1, p_2, p_3, \rho_0(p_1, p_2, p_3))$$

is also analytic on $(0, 1)^3$. This implies that its zero set must be of measure zero and hence the central limit results hold for almost all choices of the grammar probabilities.

6 Discussion

Our analysis of the Knudsen–Hein SCFG yields two sets of results. First, as proved above, the distribution for each of the biological motifs considered is asymptotically Gaussian, with an explicit mean and standard deviation given as a function of the grammar transmission probabilities. Second, and much more significantly, these results imply a set of relations among motifs in expectation which hold *independent of those probabilities.*

Recall that $\mathbb{X}^{lb}_n$, $\mathbb{X}^{rb}_n$, $\mathbb{X}^{m}_n$, $\mathbb{X}^{hel}_n$, $\mathbb{X}^{hp}_n$, and $\mathbb{X}^{i}_n$ are the number of left bulges, right bulges, multibranch loops, helices, hairpins, and internal loops in a random secondary structure on n nucleotides, respectively, while $\mathbb{X}^{m,r}_n$ is the number of multibranch loops of degree r. Since $p_1 q_2 \rho_0 < 1$, based on the calculated expectations, we have the following relations, originally stated as Theorem 1.

Theorem *For $p_i, q_i > 0$,*

(i) $\mathbb{E}(\mathbb{X}^{lb}_n) = \mathbb{E}(\mathbb{X}^{rb}_n)$,
(ii) $\mathbb{E}(\mathbb{X}^{m}_n) = \frac{1}{4}\mathbb{E}(\mathbb{X}^{hel}_n)(1 + o(n))$,
(iii) $\mathbb{E}(\mathbb{X}^{hp}_n) = (\mathbb{E}(\mathbb{X}^{i}_n) + \mathbb{E}(\mathbb{X}^{m}_n))(1 + o(n))$,

Table 1 Five sets of RNA sequences chosen from the CRW database to minimize variance in sequence length

	No. sequences (type)	Av. length	SD length
Set I	122 (5S)	121.17	3.1
Set II	37 (16S)	956.46	6.51
Set III	81 (16S)	1521.33	24.86
Set IV	50 (16S)	1787.1	20.09
Set V	34 (23S)	2912.85	23.08

(iv) $\mathbb{E}(\mathbb{X}_n^m) = (\mathbb{E}(\mathbb{X}_n^{lb}) + \mathbb{E}(\mathbb{X}_n^i))(1 + o(n))$,
(v) $\mathbb{E}(\mathbb{X}_n^{m,r+1}) < \frac{1}{2}\mathbb{E}(\mathbb{X}_n^{m,r})(1 + o(n))$, $r \geq 2$.

Note that these relations hold even for the probabilities for which the function f discussed in Sect. 5.6 is zero. Namely, the means in those cases can be computed using Theorem 3 and calculations similar to the ones in Sects. 5.4 and 5.5. The asymptotic formulas for the expected number of base pairs, helices, and loops remain the same as for the generic probabilities.

There remains, of course, the question of how well this theoretical analysis yields insights into the behaviour of the SCFG in practice, particularly with respect to the distribution of motifs in native RNA secondary structures. We are especially interested in clarifying the effects of changes in the probability parameters, which has the potential for improving the accuracy of the predicted secondary structures. Towards this end, Theorem 1 (restated above) will be critical as we shall see that, although agreement at the motif level between model and prediction is variable, the theoretical ratios are quite accurate in estimating the relationship between motifs in predicted structures. Moreover, these relations do *not* hold in the native secondary structures. From this we conclude that the accuracy of the Knudsen–Hein SCFG cannot be improved simply by adjusting the grammar probabilities.

Our test set for comparing model, prediction, and native distributions of motifs in RNA secondary structures was generated from the CRW database (Cannone et al. 2002). We downloaded all 854 5S, 16S, and 23S ribosomal structures for which the native secondary structure (without pseudoknots) has been determined by covarying sequence analysis and given in a.ct file. Out of those we selected the structures which do not have ambiguous nucleotides and this left us with a final set of 400 structures. From these we selected 5 sets of sequences, each having small variance in length. The five sets consist of sequences of the same type with approximately the same secondary structure. Their composition is given in Table 1.

We folded each test sequence using our implementation of the CYK algorithm, which yielded the most probable predicted secondary structure. As remarked in the introduction, it is common for work in this area (e.g. Dowell and Eddy 2004; Anderson et al. 2012) to use the CYK parse as a proxy for the distribution of possible secondary structures under the assumption that it is the only path with significant probability.

These structures were predicted using the default Pfold probabilities, which were originally obtained by an expectation maximization procedure on a training set of

Table 2 The default paired and unpaired probabilities used in Pfold

	A	U	G	C
A	0.001167	0.177977	0.001058	0.001806
U	0.177977	0.002793	0.049043	0.000763
G	0.001058	0.049043	0.000406	0.266974
C	0.001806	0.000763	0.266974	0.000391
A	0.364097			
U	0.273013			
G	0.211881			
C	0.151009			

tRNA and large subunit ribosomal RNA secondary structures (Knudsen and Hein 1999). The transmission probabilities are

$$p_1 = 0.868534,\ p_2 = 0.105397,\ p_3 = 0.787640,$$
$$q_1 = 0.131466,\ q_2 = 0.894603,\ q_3 = 0.212360$$

and the emission probabilities are given in Table 2.

In considering our results, it is important to recognize the distinction between a CYK prediction under the Knudsen–Hein SCFG and the output of the Pfold program for a single sequence. Pfold predicts the structure with the highest expected number of correctly predicted positions, which may have very few base pairs. For example, the structures obtained by using PPfold (Sukosd et al. 2011) (a parallelized version of Pfold) for each sequence in Set I have on average 15.78 base pairs, while the native structures have 38 and the structures predicted using CYK have 34.05 base pairs on average. The crucial difference is that Pfold has been designed to find a consensus structure for a set of aligned sequences, and this is where its strength lies.

Figures 4 and 5 display our analysis of the distribution of individual motifs in structures predicted by the SCFG in theory and in practice, as well as in the native structures for the 5S, 16S, and 23S ribosomal sequences. For each test set, we give the average number of each motif in the native structures and in the CYK predictions, with variances represented by error bars. We also give the expected number of motifs for sequences of the same length, where n is taken to be the average length of the corresponding set. The default Pfold probabilities were used in calculating the theoretical means and standard deviations according to our results above and Eq. (10).

Figure 4 shows the motifs where we observed agreement between the model expectations and the CYK prediction averages. Since the model is Gaussian, 68.2 % of the distribution should be within one standard deviation of the mean. Table 3 shows the proportion of CYK predictions for which the statistics fall within one standard deviation of the model average. By this criterion, there is a good agreement for the number of base pairs (except Set II) and the number of internal loops (except Set V). We also have reasonable agreement for right and left bulges, the number of which falls within one standard deviation of the model mean for at least 50 % of the predicted structures in most cases and within two standard deviations

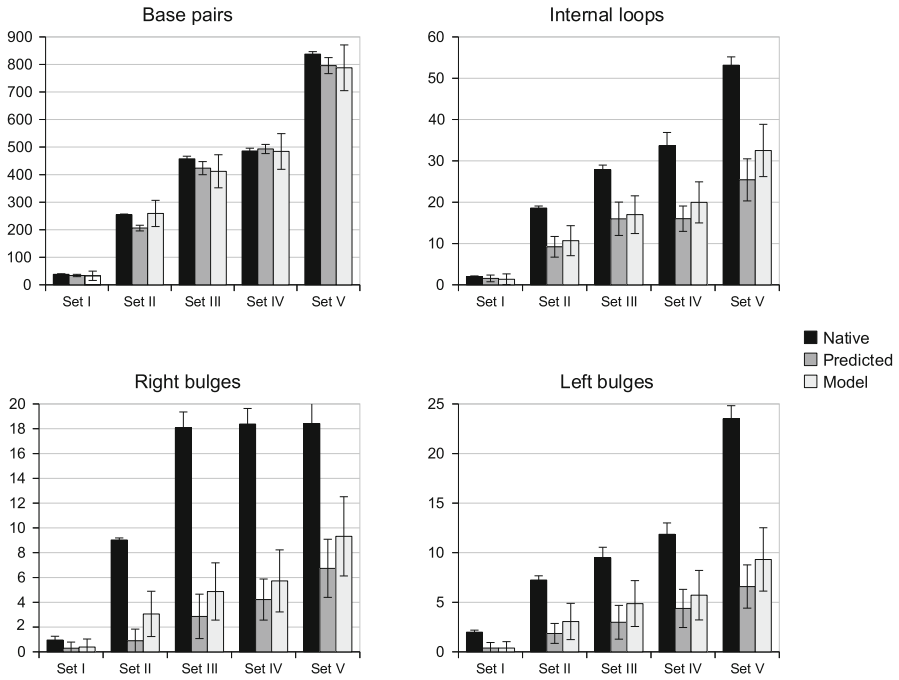

Fig. 4 Motifs where there is agreement between the model expectations and the observed distribution in the most probable predicted structures. The charts also illustrate that, in theory and in practice, the SCFG captures well the number of base pairs in the native structures, but clearly not the number of internal loops and bulges. Note that the scales on the y-axis vary between bar charts

for at least 95 % of the predicted structures for all sets. Hence, although emission probabilities were not taken into account, our model captures well the expected number of base pairs and internal loops, and reasonably well the number of left and right bulges seen in the most probable secondary structures predicted by the SCFG.

It is also interesting to consider these distributions with reference to the native structures. We see that not only are the model expectations and CYK predictions in close agreement for the number of base pairs, but that this agrees well with the native secondary structures. However, we infer from the distribution of the other motifs that the arrangement of base pairs in the predicted and in the native structures is quite different. Based on the higher number of internal loops and left/right bulges, as well as helices, we conclude that the native structures have on average shorter helices interrupted with frequent internal loops and bulges.

When the model expectations and CYK predictions do not agree, the comparison with the native structures is less striking. As seen in Fig. 5 and Table 3, the distribution of helices, multibranch loops, and hairpins in the CYK predictions differs from the model expectations (except for Set I). There are correlations among these motifs, which reflect the branching of the secondary structures, and the differences between model, prediction, and reality may merit further study. In all cases the model expectations are higher than the average for the CYK predictions, suggesting that the for-

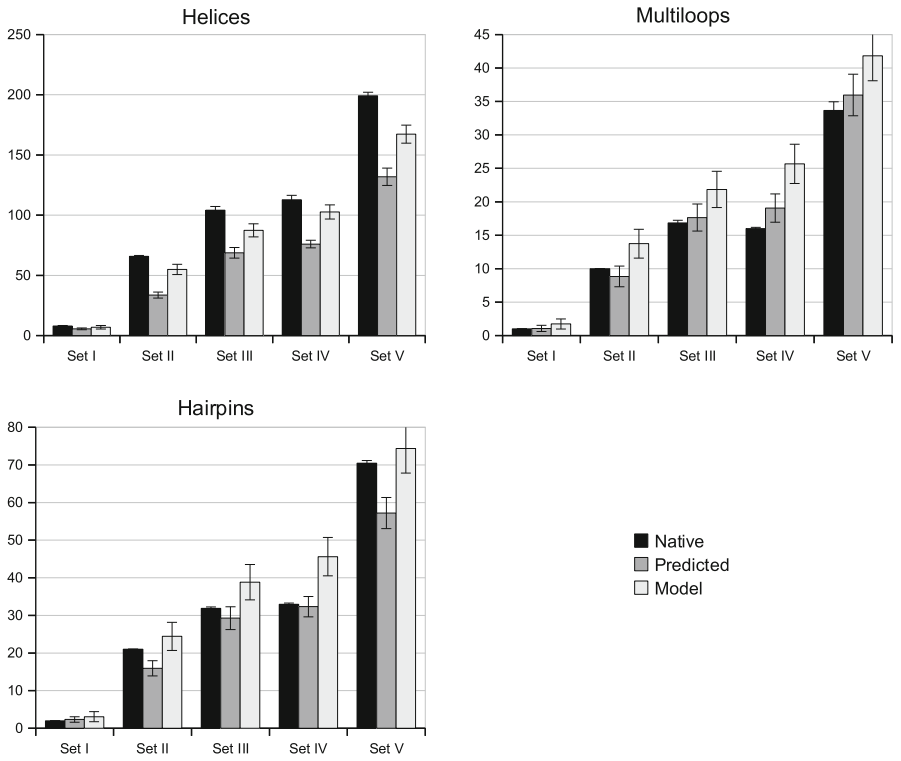

Fig. 5 Motifs for which there is a larger discrepancy between the predicted structures and the model. Note that the scales on the y-axis vary between bar charts

Table 3 Proportion of predicted structures for which the statistics fall within one standard deviation of the model mean

	Base pairs	Internal loops	Right bulges	Left bulges	Helices	Multi-branch loops	Hairpin loops
Set I	1	0.8	0.98	0.98	0.58	0.93	0.93
Set II	0.32	0.76	0.57	0.22	0	0.05	0.03
Set III	0.95	0.7	0.56	0.52	0	0.16	0.05
Set IV	1	0.72	0.62	0.66	0	0.1	0
Set V	1	0.47	0.56	0.44	0	0.24	0

mation of these motifs is less favorable when the emission probabilities are taken into account.

Hence, we see that the model captures well some aspects of the distribution of motifs in the predicted secondary structures, as approximated by the most probable ones, for our test sequences, but certainly not all. It is interesting, then, that despite the numerical differences, the CYK predictions largely satisfy the relations from Theorem 1, as listed in Table 4. Moreover, the difference from the native structures is especially striking

Table 4 Ratios of the average number of occurrences of various motifs for the native and predicted structures from the five sets

	Ratios of averages				
	RB/LB	Hel/ML	IL+ML/HP	IL+LB/ML	IL+RB/ML
Set I					
Native	0.48	7.96	1.51	4.01	2.97
Predicted	0.77	5.27	1.12	1.80	1.72
Set II					
Native	1.25	6.58	1.36	2.58	2.76
Predicted	0.50	3.80	1.13	1.24	1.14
Set III					
Native	2.50	6.18	1.40	2.22	2.72
Predicted	0.96	3.90	1.15	1.07	1.07
Set IV					
Native	1.93	7.06	1.51	2.85	3.25
Predicted	0.96	3.99	1.09	1.07	1.06
Set V					
Native	0.78	5.92	1.23	2.28	2.13
Predicted	1.02	3.67	1.07	0.89	0.89
Model	1	4	1	1	1

The last row contains the asymptotic model averages as given by Theorem 1

when these ratios are considered. For example, the helix-to-multiloop ratio in the predicted structures, which is related to branching, is close to the expected value of 4 for the longer sequences, but not for the native structures.

Table 4 clearly indicates the differences between the native and the predicted structures, as well as the agreement of the predicted structures with the model averages. The agreement between the ratios for the CYK algorithm and the model is particularly apparent for the Sets III–V, which contain longer sequences, and is more expected because our results are asymptotic. This suggests that even though the grammar probabilities can be adjusted to, say, increase the number of helices in the predicted structures, the relative frequencies of the loops in the CYK structures will remain close to the model expectations, which are independent of the parameters.

Since these ratios are constant for the model and do not agree with the corresponding ratios in the native ribosomal structures, we conclude that the CYK prediction for these sequences using the Knudsen–Hein grammar cannot be significantly improved by varying the grammar parameters. Given that the native structures for these sequences are complex, it would be interesting to see whether there are grammars which reflect their branching behavior more closely, while still being simple enough for computational purposes.

Acknowledgments The authors would like to thank Christian Reidys for useful comments on an earlier version of these results, David Esposito for implementing the CYK parsing and running the predictions, and the reviewers for their thoughtful comments which helped improve the presentation in this article.

References

Anderson J, Tataru P, Staines J, Hein J, Lyngsø R (2012) Evolving stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinform 13(1):78

Cannone J, Subramanian S, Schnare M, Collett J, D'Souza L, Du Y, Feng B, Lin N, Madabusi L, Muller K, Pande N, Shang Z, Yu N, Gutell R (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinform 3:2, [Correction: (2002), BMC Bioinformatics 3:15]

Clote P, Kranakis E, Krizanc D, Salvy B (2009) Asymptotics of canonical and saturated RNA secondary structures. J Bioinform Computat Biol 7(05):869–893

Clote P, Ponty Y, Steyaert J-M (2012) Expected distance between terminal nucleotides of RNA secondary structures. J Math Biol 65(3):581–599

De Chaumont M, Viennot G (1984) Polynômes orthogonaux et problemes dénumération en biologie moléculaire. Séminaire Lotharingien de Combinatoire 8

Denise A, Ponty Y, Termier M (2010) Controlled non-uniform random generation of decomposable structures. Theor Comput Sci 411(40):3527–3552

Doshi KJ, Cannone JJ, Cobaugh CW, R GR, (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinform 5:105

Dowell RD, Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinform 5:14

Drmota M (1997) Systems of functional equations. Random Struct Algorithms 10(1–2):103–124

Durbin R, Eddy S, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University Press, Cambridge

Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22:2079–2088

Flajolet P, Odlyzko AM (1990) Singularity analysis of generating functions. SIAM J Discret Math 3:216–240

Flajolet P, Sedgewick R (2009) Analytic combinatorics. Cambridge University Press, Cambridge

Fontana W, Konings D, Stadler P, Schuster P (2004) Statistics of RNA secondary structures. Biopolymers 33(9):1389–1404

Hofacker I, Schuster P, Stadler P (1998) Combinatorics of RNA secondary structures. Discret Appl Math 88(1):207–237

Knudsen B, Hein JJ (1999) Using stochastic context-free grammars and molecular evolution to predict RNA secondary structure. Bioinformatics 15:446–454

Knudsen B, Hein JJ (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31:3423–3428

Knudsen M (2005) Stochastic context-free grammars and RNA secondary structure prediction. PhD thesis, Aarhus Universitet, Datalogisk Institut

Lorenz W, Ponty Y, Clote P (2008) Asymptotics of RNA shapes. J Comput Biol 15(1):31–63

Mathews DH, Turner DH (2006) Prediction of RNA secondary structure by free energy minimization. Curr Opin Struct Biol 16:270–278

Nebel M (2002a) Combinatorial properties of RNA secondary structures. J Comput Biol 9(3):541–573

Nebel M (2002b) On a statistical filter for RNA secondary structures. Johann-Wolfgang-Goethe-Univ., Inst. für Informatik

Nebel M (2003) Identifying good predictions of RNA secondary structure. In: RB Altman, AK Dunker, L. Hunter, TE Klein (eds) Pacific symposium on biocomputing, vol 9, pp 423–434

Nebel M (2004) Investigation of the Bernoulli model for RNA secondary structures. Bull Math Biol 66(5):925–964

Nebel M, Scheid A (2011) Analysis of the free energy in a stochastic RNA secondary structure model. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 8(6):1468–1482

Nebel M, Reidys C, Wang R (2011) Loops in canonical RNA pseudoknot structures. J Comput Biol 18(12):1793–1806

Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, Haussler D (1994) Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res 22:5112–5120

Scheid A, Nebel M (2012) Evaluating the effect of disturbed ensemble distributions on SCFG based statistical sampling of RNA secondary structures. BMC Bioinform 13(1):159

Schützenberger MP (1963) On context-free languages and push-down automata. Inform control 6:246–264
Sukosd Z, Knudsen B, Vaerum M, Kjems J, SAndersen E (2011) Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. BMC Bioinform 12:103
Turner DH, Mathews DH (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res 38:D280–D282
Waterman M (1978) Secondary structure of single-stranded nucleic acids. Adv Math Suppl Stud 1:167–212