Investigation of the Bernoulli-Model for RNA Secondary Structures

Markus E. Nebel
Johann Wolfgang Goethe-Universität
Fachbereich Biologie und Informatik
Institut für Informatik
Frankfurt a. M.
Germany

Abstract

Within this paper we investigate the Bernoulli-model for random secondary structures of RNA molecules. Assuming that two random bases can form a hydrogen bound with probability p we prove asymptotic equivalents for the averaged number of hairpins and bulges, the averaged loop-length, the expected order, the expected number of secondary structures of size n and order k and further parameters all depending on p. In this way we get an insight into the change of shape of a random structure during the process $1 \xrightarrow{p} 0$. Afterwards we compare the computed parameters for random structures in the Bernoulli model to the corresponding quantities of real existing secondary structures of large subunit rRNA molecules found in the database of Wuyts et al. . That's how it becomes possible to identify the mayor weaknesses of the Bernoulli-model for secondary structures.

1 RNA Secondary Structure

A ribonucleic acid (RNA) molecule consists of a chain of four different types of nucleotides. Each nucleotide contains a base, a phosphate group (PO_4^{3-}) and a sugar group (ribose). The different types of nucleotides only differ by the base involved; the chemical structure of the four choices adenine (A), cytosine (C), guanine (G) and uracil (U) is illustrated in Figure 1. The non-planar 5 member ribose ring connects the phosphate to the base. The chain is formed by means of the phosphate groups; the phosphate group of one nucleotide is linked to the ribose ring of its neighbor (see Figure 2). The specific sequence of bases along the chain is called the *primary structure* of the molecule. It is usually modelled as a string over the alphabet $\{A, C, G, U\}$. Through the creation of hydrogen bounds,

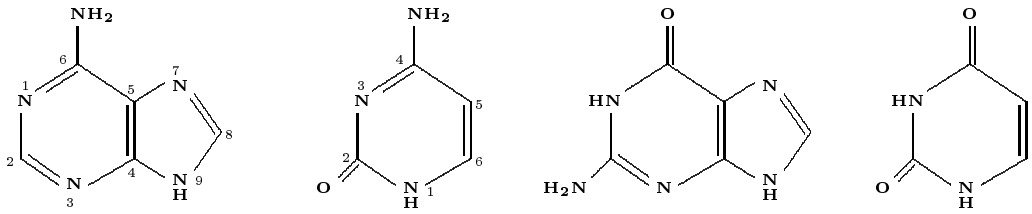


Figure 1: The four bases adenine, cytosine, guanine and uracil (from left to right).

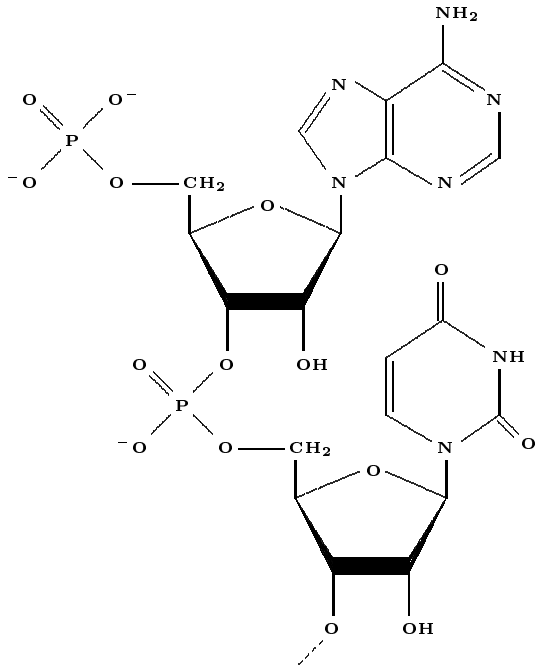


Figure 2: Two nucleotides chained by a link between their phosphate groups.

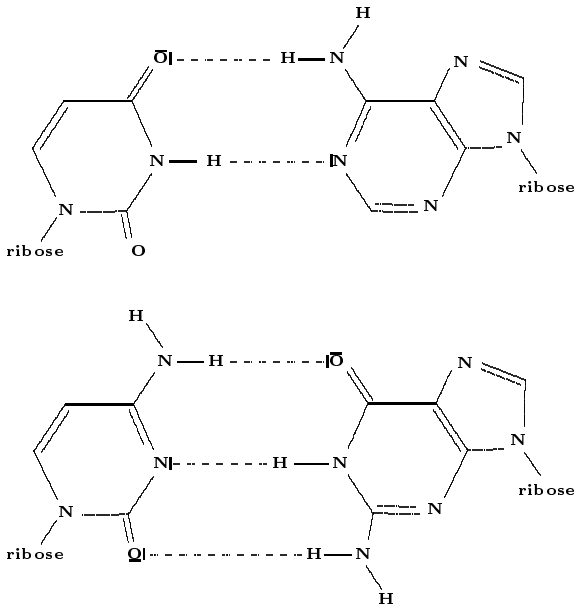


Figure 3: The hydrogen bounds between the complementary bases adenine and uracil (top), resp. cytosine and guanine (bottom).

the complementary bases A and U (resp. C and G) form stable base pairs with each other (see Figure 3). Additionally, there exists the weaker G-U pair, where the bases bind in a skewed fashion. All of these are called *canonical* base pairs. Other *non-canonical* base pairs may occur, some of which are stable. By the creation of base pairs the primary structure is folded into a stable three-dimensional conformation called *tertiary structure* of the molecule. It is customary in sciences to study the simplified *secondary structure* by focusing ones attention just on what bases form pairs and allow the sequence to form helical regions in two dimensions. The secondary structure plays a rôle in the interaction of tRNA with proteins [14], in stabilizing mRNA and in packing RNA into virus particles. Since experimental approaches like X-ray diffraction are quite expensive much effort has been made to deduce the secondary structure from the knowledge of the primary structure. One possible technique is to compute a conformation of minimum free energy. With respect to this task the notion of order of a secondary structure has been introduced in [11]. The idea was restated rigorously in [22] where the first formal framework for secondary structures has been introduced. The working hypothesis which makes the evaluation of the free energy $E(S)$ of structure S feasible is that if we decompose S into disjoint substructures $S_1, S_2, \ldots, S_t$, then $E(S) = e(S_1) + e(S_2) + \cdots + e(S_t)$ where $e(S_i)$ denotes the energetic contribution of substructure S_i. One possible method for an efficient prediction of the secondary structure then is to first construct an optimal first-order structure. Using the results from the previous pass, successively higher

order structures are computed in an iterative way. The algorithms of Waterman [22] and Mainville [10] can be seen to work this way.
Many authors have paid attention to enumeration problems related to the combinatorics of RNA secondary structures. Assuming that base-pairing is possible between arbitrary pairs of nucleotides, the set of all possible structures is modelled as a specific kind of planar graph [22]. Parameters of interest are the number of different structures of a given size, the number of structures of given size and order, the expected number of specific substructures but also the systematical treatment of such problems from a mathematical point of view [8, 13, 15, 16, 18, 20, 22, 23, 24]. A more realistic model, the so-called Bernoulli-model, is obtained by a stochastic approach where we assume a Bernoulli distribution of the bases. A parameter p is used to specify the probability that two random bases can be paired. This model was considered for example in [6, 27]. In its final section, the article [6] presents some asymptotics for parameters like the average number of stacks per base or the expected stacklength, all based on the probability p.
The aim of this paper is to judge the quality of the Bernoulli-model for random secondary structures and to detect its major weaknesses. In this way it should be possible to find better, more realistic models which could be handled mathematically, too. For this purpose we first study the influence of the probability p on the expected shape of a secondary structure in the Bernoulli-model. Therefore, we will derive parameters like the expected number of specific substructures like hairpins and bulges, the expected lengths of loops and so on, all depending on p, so that it will be possible to conclude what a typical secondary structure of a given size looks like. Those parameters were not determined in [6] and we will present a different, less classical method to derive them. Afterwards we will rate the quality of the Bernoulli-model by comparing our results to the corresponding parameters of real secondary structures of long subunit ribosomal RNA molecules taken from the database of Wuyts et al.[25].
Before we start, we restate some definitions and prior results such that it becomes possible to state precisely which problems are to be considered here.

2 Definitions and Prior Results

We will first consider the combinatorial model for the RNA secondary structure in order to introduce all the terms needed.

Definition 1 ([20]) *For $\Sigma := \{(, |,)\}$ and $w \in \Sigma^\star$ let $|w|_x$ for $x \in \Sigma$ denote the number of occurrences of symbol x in w. Then a word $w \in \Sigma^n$ is a secondary structure of size n if w satisfies the three following conditions:*

(1) For every factorization $w = u \cdot v$, $|u|_(\geq |u|_)$.

(2) $|w|_(= |w|_)$.

(3) w has no factor $(\,)$.

Within this model a pair of corresponding brackets in a word w represents two bases of the molecule which are paired. The symbol $|$ is used to represent an unpaired nucleotide. The words of $\Sigma^\star$ which satisfy the conditions (1) and (2) are known as *Motzkin words*; words over the alphabet $\{(,)\}$ satisfying the conditions (1) and (2) are usually called *semi-Dyck words*. Condition (3) takes into account that two (with respect to the primary structure) adjacent bases cannot be paired. Thus, condition (3) implies a minimal length for the hairpin-loops (which formally will be defined later) of one (while a value of three would provide a realistic lower bound). However, in that way our definition is equivalent to the graph-theoretic definition given in [22] which makes it possible to compare our results to the rich set of results for the combinatorial model. Using our formal framework, Definition 2.2 of [22] reads as follows:

Definition 2 *Let w be a secondary structure of size n and let w_i denote the i-th symbol of w, $1 \leq i \leq n$.*

(i) The subword $v = w_{i+1} \dots w_{j-1}$ is a (hairpin)-loop*, if $v \in \{|\}^+$ and $w_i w_j = (\,)$ is a corresponding pair of brackets of w.*

(ii) The subword $v = w_{i+1} \dots w_{j-1}$ is a bulge*, if $v \in \{|\}^+$ and $w_i w_j \in \{(,)\}^2$ but $w_i w_j$ does not represent a pair of corresponding brackets of w.*

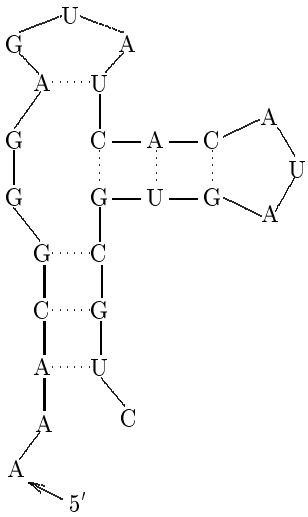
(iii) A tail *is a prefix $v = w_1 \dots w_i$ resp. a suffix $v = w_j \dots w_n$ such that $v \in \{|\}^+$ and w_{i+1} resp. w_{j-1} is in $\{(,)\}$.*

(iv) A hairpin *is a subword $v = w_{i+1} \cdots w_{j-1}$ such that v contains exactly one loop, $w_{i+1} w_{j-1}$ is a corresponding pair of brackets of w, but $w_i w_j$ is none.*

(v) A ladder *consists of two maximal subwords u, v such that $u = w_i \dots w_{i+c}$ and $v = w_j \dots w_{j+c}$ and w_{i+k}, w_{j+c-k} is a pair of corresponding brackets, $0 \leq k \leq c$. The length of a ladder is given by $c + 1$.*

Note that multiple bulges together might form more complex substructures like for instance interior loops or multi-loops. However, in this paper we won't distinguish the contexts where bulges may occur.
Next we will define the order of a secondary structure already mentioned in the first section. For $w = w_1 \cdots w_n$ a semi-Dyck word of length n, a subword $v = w_j \cdots w_{j+2i-1} = (^i\,)^i$ is called *pyramid* of w. The pyramid v is called *maximal* if $w_{j-1} \neq ($ or $w_{j+2i} \neq)$. We define $\Pi(w)$ to be the semi-Dyck word which results from w by deleting all maximal pyramids in w.

Definition 3 *Let w be a secondary structure and let $\alpha(w)$ denote the semi-Dyck word which results from w by the deletion of all symbols $|$ in w. w is said to be of* order *k if k is the smallest integer such that $\Pi^{(k)}(\alpha(w)) = \varepsilon$ holds. Here ε is the empty word and $\Pi^{(k)}$ denotes the k-th iterated of Π.*



$w = ||(((||(||)(((||))))))|$ $\alpha(w) = (((\underline{()} \underline{(((\,)\,)\,)}\,))) \to (((\,))) \to \varepsilon$

Figure 4: An example of a RNA molecule of size 25 and order 2 (left). Its first base (usually called the $5'$ terminus) in the chain is marked by an arrow. The hydrogen bounds are represented as dotted lines. The corresponding representation as a Motzkin word (middle) and the two runs of deleting maximal pyramids necessary to erase $\alpha(w)$ (right). The two maximal pyramids which are deleted in the first run are underlined within $\alpha(w)$.

In Figure 4 we find an example molecule together with its abstract representation as a Motzkin word and the runs performed in order to determine its order. The secondary structure of Figure 4 possesses one bulge (6th and 7th base, subsequence GG) and two hairpin-loops (9th to 11th base, subsequence GUA and 16th to 18th base, subsequence AUA). The two tails are given by the sequences AA and C. Furthermore, it has two hairpins (8th to 12th base, subsequence AGUAU and 13th to 21th base, subsequence CACAUAGUG). The structure has 3 ladders, two of length 3 (3rd to 5th base, 22nd to 24th base and 13th to 15th base, 19th to 21st base) and one of length 1 (8th and 12th base). Let us try to provide an insight into the notion of order. If we traverse a secondary structure along its ladders starting at its tails we may reach a point were the path has to split and we have the choice to continue with at least two ladders. (In the example of Figure 4 this is the case when we have reached the GC-pair at positions 5 and 22). Let us call such a point *bifurcation*. Then the order gives information about the maximal nesting-depth of bifurcations within the structure considered. Therefore it seems to be reasonable that the order should in some way be related to the spatial structure of a RNA molecule: A molecule with a relative small number of bases but a high order cannot stay planar, since there is not enough room for all its nested substructures; those need to make room for each other by leaving the plane. Viennot [20] was the first to notice that there is a close connection

between the order of secondary structures and the *Horton-Strahler number* of binary trees. For a binary tree T we recursively define its Horton-Strahler number $hs(T)$ in the following way:

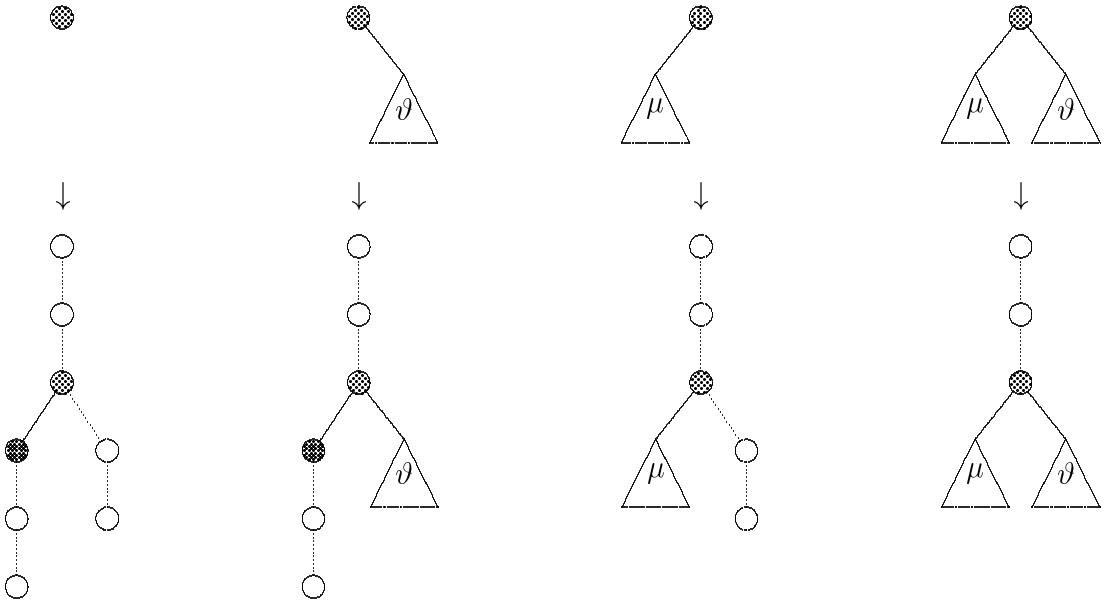
$$hs(T) := \begin{cases} 0 & : \text{ if } T \text{ is either a leaf or empty} \\ hs(T.l)+1 & : \text{ if } hs(T.l) = hs(T.r) \\ \max(hs(T.l), hs(T.r)) & : \text{ otherwise} \end{cases} .$$

Here, $T.l$ (resp. $T.r$) denotes the left (resp. right) subtree of T. Thus, as for the order, the branching-points (internal nodes) of the tree are responsible for a growth of the parameter in the case where both its subtrees possess the same Horton-Strahler number. This is exactly the same for the order of a secondary structure. When we *delete* a secondary structure step by step according to Definition 3, a bifurcation is responsible for a $(k+1)$st iteration whenever at least two of its substructures need k iterations to be deleted and none of them needs more than k. In all other cases, the bifurcation will be *deleted* by the same iteration as its substructure, which needs the largest number of iterations to be deleted. Originally the Horton-Strahler number was used by Horton and Strahler [7, 19] to study the morphological structure of river networks. It is also of interest to numerous other subjects like botany or anatomy in which branching patterns appear. Furthermore, there are several links of the Horton-Strahler number of a binary tree to computer science; for an overview we refer to [21] and the references given there. Viennot [20] noticed that if only the paired bases (symbols (and)) contribute to the size of a secondary structure, then the enumerator generating function of secondary structures of order k coincides with the enumerator generating function of binary trees with a Horton-Strahler number k. We will use this observation in order to derive the generating functions that are needed to conclude our results related to the order. As already done in [13] we will derive our generating functions from corresponding generating functions for binary tree structures presented in [12] just by substitutions for the variables. Since these substitutions must be adjusted to the different parameters considered in the following section, it is essential to have an idea of how these substitutions work in order to understand the methodology. Let $\mathbf{b}(t)$ (resp. $\mathbf{u}(t)$; $\mathbf{l}(t)$) denote the number of nodes of an extended binary tree t with two successors which are no leaves (resp. with one successor which is no leaf; with two successors which are leaves). The ordinary generating functions $\mathbf{T}(x,u,v) := \sum_{t\in\mathcal{T}} x^{\mathbf{b}(t)} u^{\mathbf{u}(t)} v^{\mathbf{l}(t)}$ and $\mathbf{R}_k(x,u,v) := \sum_{t\in\mathcal{T}_k} x^{\mathbf{b}(t)} u^{\mathbf{u}(t)} v^{\mathbf{l}(t)}$ for $\mathcal{T}$ the set of all extended binary trees and $\mathcal{T}_k$ the set of those $t \in \mathcal{T}$, which have a Horton-Strahler number of k, possess the following representations [12]:

$$\mathbf{T}(x,u,v) = \frac{1-2\,u-\sqrt{1-4\,u+4\,u^2-4\,xv}}{2x},$$

$$\mathbf{R}_k(x,u,v) = -\frac{v}{\sqrt{xv}}\frac{1}{U_{2^k-1}\left(\frac{2\,u-1}{2\sqrt{xv}}\right)} = \frac{v\,(1-\omega)\,\omega^{2^{k-1}}}{\sqrt{xv}\sqrt{\omega}\left(1-\omega^{2^k}\right)}. \qquad (1)$$

Here, $U_n(z)$ denotes the n-th Chebyshev polynomial of the second kind (see e.g. [1]) and $\omega := \left(1-\sqrt{1-4\,\frac{xv}{(1-2\,u)^2}}\right)\left(1+\sqrt{1-4\,\frac{xv}{(1-2\,u)^2}}\right)^{-1}$. Every secondary structure w can be reconstructed from the semi-Dyck word $\alpha(w)$ by inserting unary symbols | at the appropriate positions. Thus, since we can identify binary trees and semi-Dyck words (see e.g. [26]), it becomes possible to consider secondary structures based on these generating functions by inserting linear lists. For $|^\star$ a string of zero or more symbols | and $|^+$ a string of at least one symbol |, the cases to be distinguished for this procedure are:
$() \longrightarrow |^\star(|^+)|^\star \qquad ()\vartheta \longrightarrow |^\star(|^+)\vartheta \qquad (\mu) \longrightarrow |^\star(\mu)|^\star \qquad (\mu)\vartheta \longrightarrow |^\star(\mu)\vartheta.$
These cases correspond one-to-one to the following transformations for trees:



Here, a light shaded node descends from the pair of brackets occurring in the corresponding transformation of a semi-Dyck word. A dark shaded node must be inserted since the appropriate list complies with $|^+$, whereas the insertion of the non-shaded nodes is not mandatory as indicated by the dotted edges. The leftmost case corresponds to a node marked by v within $\mathbf{T}$ and $\mathbf{R}_k$. If we let z mark a paired base (an opening or closing bracket) and a mark an unpaired base (a symbol |) then the light shaded node corresponds to z^2 since it represents a pair of bound bases. Furthermore, the three lists which must be attached to this node are given by $\frac{a}{(1-a)^3}$. Thus, the substitution

$$v := z^2\frac{a}{(1-a)^3} \tag{2}$$

takes care of this case. The three other cases can be handled in a similar way leading to

$$u := \frac{1}{2}z^2\left(\frac{a}{(1-a)^2} + \frac{1}{(1-a)^2}\right) \text{ and } x := z^2\frac{1}{1-a}. \tag{3}$$

Remark: The correspondence given above does not work in a way which translates each binary tree with Horton-Strahler number k into a secondary structure of order k and vice versa. However, each binary tree with Horton-Strahler number k is translated into a secondary structure w which has the same number of paired and unpaired bases and the same number of subwords $(\,)$ within $\alpha(w)$ as the suitable structure of order k. Therefore the correspondence can be used for all enumeration purposes considered in this paper but also e.g. in order to enumerate the number of hairpins or bulges in structures of order k as done in [13].

Now, let us use the considerations related to the combinatorial model in order to investigate the Bernoulli-model sketched in the first section. We suppose that the different bases X appear independently with probabilities $p(X)$, $X \in \{A, C, G, U\}$, in a random primary structure. Assuming that only the Watson-Crick base pairs A-U and C-G are possible, $p := 2(p(A)p(U)+p(C)p(G))$ is the probability that two random bases can form a hydrogen bound. Obviously, not every secondary structure w according to Definition 1 is compatible with a given primary structure s since the i-th and j-th base of s might be non-complementary, whereas w_iw_j is a pair of corresponding brackets of w. What we are interested in, is the expected number of different secondary structures and related parameters supposing that only structures which are compatible with a random sequence of bases are counted. We first observe, that the assumption of the Bernoulli-model does not affect the unpaired bases. Since each base is possible in an unpaired position, their probabilities sum up to 1. The situation for the paired bases is the contrary. If we fix one base it determines which base is possible as its counterpart. Thus a random primary structure may have i paired positions only with probability p^i. As a consequence, the probability that a random primary structure of length n is compatible with $|^n$ is one, the probability that it is compatible with $(|)|^{n-3}$ is p and so on. We can translate this behavior into our generating functions by setting z to $z\sqrt{p}$ with the effect that their coefficients now describe the desired expected values instead of absolute numbers. Note that this substitution only provides the generating functions needed. We cannot reuse the asymptotics presented in [13] for the combinatorial model just by substitutions, the asymptotics for the coefficients must be determined from scratch. Note further that we are not restricted to the case of four different bases with only the base pairs A-U and C-G. The considerations presented here are valid for each probability p independent from which number of symbols or which kind of pairings it results. In the style of [9], the probability p will be called *stickiness* in the sequel.

3 Investigation of the Bernoulli-Model

In this section we will derive numerous results related to the Bernoulli-model for secondary structures. We will use the generating functions of the previous section together with the $\mathcal{O}$-transfer method [4] to derive asymptotic estimates for the expected number of secondary structures of size n and many other parameters all depending on the stickiness p. In order to make this article more self-contained we first give a brief description of how the $\mathcal{O}$-transfer method works.
Assume we have a generating function $f(z) = \sum_{n\geq 0} f_n z^n$ with $f_n \geq 0$ for all n and we wish to approximate f_n for large n. In the sequel we will use the notation $[z^n]f(z)$ to denote the coefficient at z^n in the expansion of $f(z)$ around 0. The basic principle of the method is the existence of a correspondence between the asymptotic expansion of $f(z)$ near its dominant singularities and the asymptotic expansion of the coefficients f_n. Here, a singularity is called dominant if it is located on the circle for convergence of $f(z)$ or equivalent if it is a singularity of smallest modulus. It is convenient to consider functions $f(z)$ that are singular at $z = 1$, a restriction that entails no loss of generality. If $f(z)$ is singular at $z = \rho^{-1}$ and $g(z) := f(z/\rho)$, then by the scaling rule of Taylor expansions $[z^n]f(z) = \rho^n[z^n]f(z/\rho) = \rho^n[z^n]g(z)$, where $g(z)$ is singular at $z = 1$. The method applies to so-called algebraic-logarithmic functions, i.e. functions whose singular expansions involve logarithms and fractional powers. Two types of results are used. First, a catalogue of coefficients of standard functions which occur in such singular expansions so that the coefficients of the main terms can be extracted. Second, suitable theorems which allow to extract the asymptotic order of error terms involved. Both will just be presented without proof. We refer the reader to [4, 5] for details.
The following table of commonly encountered functions together with the asymptotic forms of their coefficients contains all estimates which are used within this paper. A similar table with much more entries can be found in [5].

Function	Coefficient at z^n
$(1-z)^{1/2}$	$-\frac{1}{\sqrt{\pi n^3}}\left(\frac{1}{2}+\frac{3}{16n}+\frac{25}{256n^2}+\mathcal{O}(n^{-3})\right)$
$-(1-z)^{1/2}\log(1-z)$	$-\frac{1}{\sqrt{\pi n^3}}\left(\frac{1}{2}\log(n)+\frac{\gamma+2\log(2)-2}{2}+\mathcal{O}(\log(n)/n)\right)$
$(1-z)^{-1/2}$	$\frac{1}{\sqrt{\pi n}}\left(1-\frac{1}{8n}+\frac{1}{128n^2}+\frac{5}{1024n^3}+\mathcal{O}(n^{-4})\right)$
$(1-z)^{\alpha}$, $\alpha \notin \mathbb{N}_0$	$n^{-\alpha-1}/\Gamma(-\alpha)+\mathcal{O}(n^{-\alpha-2})$.

The basic requirement for the method is that the asymptotic expansion of the function should be valid in an area of the complex plain which extends beyond the disk of convergence of the original series. This requirement is described by the notions of Δ-domain and Δ-analyticity which will be introduced in the following definition.

Definition 4 ([5]) *Given two numbers ϕ, R with $R > 1$ and $0 < \phi < \frac{\pi}{2}$, the open domain $\Delta(\phi, R)$ is defined as*

$$\Delta(\phi, R) := \{z \mid |z| < R, z \neq 1, |\arg(z-1)| > \phi\}.$$

A domain is a Δ-domain if it is a $\Delta(\phi, R)$ for some R and ϕ. A function is Δ-analytic if it is analytic in some Δ-domain.

If a function f is Δ-analytic and its asymptotic expansion (including the error term) is valid for the entire Δ-domain then we are allowed to transfer f's expansion term by term into an asymptotic for f's coefficients; the error term of the expansion translates into an error term for the asymptotic. More technically we have

Theorem 1 ([5]) *Assume that $f(z)$ is Δ-analytic and that it satisfies in the intersection of a neighbourhood of 1 and of its Δ-domain the condition*

$$f(z) = \circ\left((1-z)^{-\alpha}(\log\frac{1}{1-z})^{\beta}\right).$$

Then

$$[z^n]f(z) = \circ(n^{\alpha-1}(\log n)^{\beta}).$$

Here, $\circ$ is one of the operators in $\{\mathcal{O}, o\}$.

Analyticity in a Δ-domain is not a stringent requirement since the basic functions $\frac{1}{1-z}$, $\exp(z)$, $-\log(1-z)$ and $\sqrt{1-z}$ are all Δ-analytic and apart from a few degenerated exceptions the composition of these remains Δ-analytic.

Now everything is prepared to derive our results. We start our investigations with the computation of the expected number of secondary structures assuming that the stickiness p is a parameter. As described in the preceding sections we can use the generating function $\mathbf{T}(x, u, v)$ together with the substitutions

$$v := z^2\frac{a}{(1-a)^3},\ u := \frac{1}{2}z^2\left(\frac{a}{(1-a)^2} + \frac{1}{(1-a)^2}\right) \text{ and } x := z^2\frac{1}{1-a}$$

to solve this task. In the sequel we will write $T(z, a)$ to represent $\mathbf{T}(x, u, v)$ with these substitutions inserted. In order to take care of the stickiness we set $z = z\sqrt{p}$ and $a = z$ within $T(z, a)$. The resulting generating function possesses the expected number of secondary structures of size n as its coefficient at z^n. It possesses the representation

$$T(z\sqrt{p}, z) = \frac{1-2z+z^2-z^2p-z^3p-(1-z)\sqrt{1-2z^2p-2z^3p-2z+z^2+z^4p^2}}{2z^2p(1-z)}.$$

The dominant singularity of that function is determined by a zero of the square root and is located at

$$z = z_d(p) := \frac{1 - \sqrt{1 + 4\sqrt{p} + 2\sqrt{p}}}{2p} = \left(\frac{1 + \sqrt{1 + 4\sqrt{p}}}{2} \right)^{-2}.$$

The expansion of the function around $z_d(p)$ is given by (terms relevant for the asymptotic of the coefficients only):

$$-\frac{\sqrt{2}p^{1/4}}{\sqrt{\frac{1+4\sqrt{p}+2p}{\sqrt{1+4\sqrt{p}}} - 1 - 2\sqrt{p}}} \left(1 - \frac{z}{z_d(p)}\right)^{1/2} + \mathcal{O}\left(\left(1 - \frac{z}{z_d(p)}\right)^{3/2}\right).$$

The application of the methodology described at the beginning of this section yields:

Lemma 1 *Under the assumption of the Bernoulli-model with a stickiness p, the expected number of secondary structures of size n is asymptotically given by*

$$\frac{z_d(p)^{-n}}{\sqrt{\pi n^3}} \frac{p^{1/4}}{\sqrt{2}\sqrt{\frac{1+4\sqrt{p}+2p}{\sqrt{1+4\sqrt{p}}} - 1 - 2\sqrt{p}}} + \mathcal{O}\left(z_d(p)^{-n} n^{-5/2}\right),\ n \to \infty.$$

□

Note that this is just another representation of the asymptotic formula already presented in [27]. The next parameter that we will consider is the expected number of unpaired bases in a random secondary structure. Since each unpaired base is marked by variable a within $T(z, a)$, this number can be determined by differentiation with respect to a. The appropriate generating function is given by

$$\left[\frac{\partial}{\partial a} T(z\sqrt{p}, az)\right]_{a=1} = \frac{(z-1)^2(1 + z(pz-1)) - (1 + z(2pz + z - 2))\Upsilon}{2p(z-1)^2 z \Upsilon},$$

$$\Upsilon := \sqrt{1 + z(-2 + z(1 + p(-2 + z(-2 + pz))))}.$$

Again, the only dominant singularity is located at $z = z_d(p)$. The corresponding expansion is given by (terms relevant for the asymptotic only)

$$\frac{p^{1/4}}{\sqrt{2}\sqrt{\sqrt{1+4\sqrt{p}}(1+4\sqrt{p}+2p) - 1 - 6\sqrt{p} - 8p}} \left(1 - \frac{z}{z_d(p)}\right)^{-1/2} + \mathcal{O}\left(\left(1 - \frac{z}{z_d(p)}\right)^{1/2}\right).$$

Thus we find

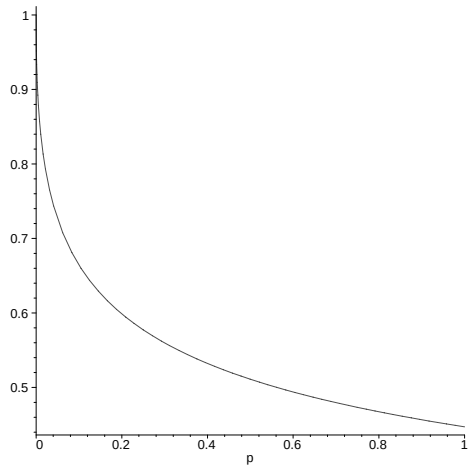


Figure 5: The portion of unpaired bases in a random structure depending on the stickiness p.

Lemma 2 *Under the assumption of the Bernoulli-model with a stickiness p, the expected number of unpaired bases in all secondary structures of size n is asymptotically given by*

$$\frac{z_d(p)^{-n}}{\sqrt{\pi n}} \frac{p^{1/4}}{\sqrt{2}\sqrt{\sqrt{1+4\sqrt{p}}(1+4\sqrt{p}+2p)-1-6\sqrt{p}-8p}} + \mathcal{O}\left(z_d(p)^{-n} n^{-3/2}\right),\ n \to \infty.$$

□

We will compute the *averaged* expected number of unpaired bases in a random secondary structure of size n by dividing the two asymptotic formulæ one by the other. Note that this yields the fraction of the expected values but not the expected value of the fraction.

Theorem 2 *Under the assumption of the Bernoulli-model with a stickiness p, the averaged number of unpaired bases in a secondary structure of size n is asymptotically given by*

$$\frac{n}{\sqrt{1+4\sqrt{p}}} + \mathcal{O}(1),\ n \to \infty.$$

□

Note that this is not the same result as the one presented in [6] since there the authors have assumed a minimal hairpin-loop length of three. Figure 5 shows a plot of the slope $\frac{1}{\sqrt{1+4\sqrt{p}}}$ which describes the portion of unpaired bases in the structure. In Table 1 you find some exact values of the averaged number of unpaired bases compared to our asymptotical equivalent as given in the previous theorem. We observe that the error made by our asymptotic is rather small,

p	1			$\frac{1}{2}$			$\frac{3}{8}$			$\frac{1}{4}$		
n	$\sim$	$=$	/	$\sim$	$=$	/	$\sim$	$=$	/	$\sim$	$=$	/
10	4.47	5.20	1.16	5.11	5.87	1.15	5.38	6.15	1.14	5.77	6.51	1.13
50	22.36	23.11	1.03	25.55	26.38	1.03	26.92	27.78	1.03	28.87	29.77	1.03
100	44.72	45.47	1.02	51.11	51.93	1.02	53.84	54.70	1.02	57.74	58.64	1.02

Table 1: Some exact values for the averaged number of unpaired bases compared to their asymptotic equivalent as given in Theorem 2. The columns labeled with $\sim$ contain the asymptotical-, the columns labeled with $=$ the exact values; the columns labeled with / contain the quotient of the exact divided by the asymptotical value, i.e. the relative error of the asymptotic. All entries are rounded to the second decimal place.

even for small values of n. This is of special interest for our case since functional RNA molecules tend to be short and thus a comparison of our results to real data like e.g. tRNA sequences as contained in the database of Sprinzl et al. [17] would become questionable if large values for n were required. The precision of all the other asymptotics which will be presented throughout this paper is always similar to the one discussed here. Next, we will see what influence on the number of hairpins in a secondary structure is implied by the stickiness. Since every hairpin possesses exactly one loop, it is sufficient to count the number of hairpin-loops. Such a loop is generated exactly at those positions of a secondary structure $w \in \{(,),|\}^\star$ where we had to insert $|^+$ between a pair of brackets $(\,)$ in order to comply with condition (3) of Definition 1. Furthermore, the insertion of $|^+$ is translated into generating functions by means of the geometric series $\frac{a}{1-a}$. Thus we can mark each hairpin by variable h by changing each $\frac{a}{1-a}$ into $\frac{ha}{1-a}$ within the substitutions for x, u and v. Afterwards we set $z = z\sqrt{p}$ and $a = z$ to get

$$\frac{1-z}{2z^2p}\left[1-\frac{z^2p\,(1+hz)}{(z-1)^2}\right.$$
$$\left.-\sqrt{\frac{1+z\left(-4+z\left(6+(-4+z)\,z+p^2z^2\,(-1+hz)^2-2\,p\,(z-1)^2\,(1+hz)\right)\right)}{(z-1)^4}}\right].$$

Taking the first partial derivative with respect to h and setting $h = 1$ afterwards provides the desired generating function

$$\frac{z}{2(z-1)}\left(1-\frac{1+z(pz-1)}{\sqrt{1+z(-2+z(1+p(-2+z(-2+pz))))}}\right).$$

In this case the dominant singularity again stems from a zero of the square root and is located at $z = z_d(p)$ also. The expansion at $z_d(p)$ possesses the following

term relevant for the asymptotic

$$\frac{\left(\frac{p}{1+4\sqrt{p}}\right)^{1/4}}{2\sqrt{p}+\sqrt{1+4\sqrt{p}}-1}\left(1-\frac{z}{z_d(p)}\right)^{-1/2}+\mathcal{O}\left(\left(1-\frac{z}{z_d(p)}\right)^{1/2}\right).$$

Thus, we have for the asymptotic of the coefficient

Lemma 3 *Under the assumption of the Bernoulli-model with a stickiness p, the expected number of hairpins in all secondary structures of size n is asymptotically given by*

$$\frac{z_d(p)^{-n}}{\sqrt{\pi n}}\frac{\left(\frac{p}{1+4\sqrt{p}}\right)^{1/4}}{2\sqrt{p}+\sqrt{1+4\sqrt{p}}-1}+\mathcal{O}\left(z_d(p)^{-n}n^{-3/2}\right),\ n\to\infty.$$

□

Dividing this quantity by the expected number of secondary structures of size n gives some insight into the behavior of a single structure. We have

Theorem 3 *Under the assumption of the Bernoulli-model with a stickiness p, the averaged number of hairpins in a secondary structure of size n is asymptotically given by*

$$\frac{\left(1-\frac{1+\sqrt{p}}{\sqrt{1+4\sqrt{p}}}\right)}{2-\sqrt{p}}n+\mathcal{O}(1),\ n\to\infty.$$

□

In Figure 6 the slope of this formula is plotted against p. In Figure 7 we find a plot of the number of bases which reside in a random secondary structure per single hairpin on the average. Due to the drastic increase of the number of unpaired bases for shrinking p we are faced with the question where all the unpaired bases reside. There are several possibilities. The length of the loops, bulges and tails may increase or the number of bulges may become larger. All possibilities will be considered in the following such that, at the end, the *real* behavior is known. At first, we will consider the number of bulges. The bulges are generated by the insertion of $|^\star$ into the semi-Dyck words and equivalently by the geometric series $\frac{1}{1-a}$ for the generating functions. However, only the insertion of at least one $|$ generates a bulge. Thus we have to decompose $|^\star$ into $\{\varepsilon\}\cup|^+$, i.e. $\frac{1}{1-a}$ into $1+\frac{a}{1-a}$. Then each bulge is marked by variable b by using $1+\frac{ba}{1-a}$ instead of $\frac{1}{1-a}$ within the substitutions for x, u and v. Furthermore, the tails are generated by $|^\star$ also. Thus, these substitutions would overestimate the number of bulges; at two places (two tails) the series $\frac{1}{1-a}$ is replaced even if it should remain unchanged. Since $(1+\frac{ba}{1-a})^2(1-a+ba)^{-2}=(\frac{1}{1-a})^2$ we can get rid of this problem by multiplying the

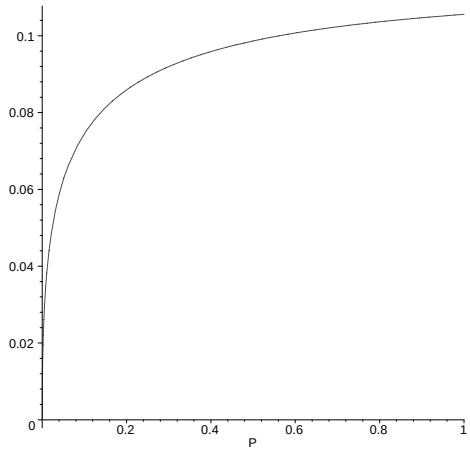


Figure 6: The slope of the number of hairpins in a random structure depending on the stickiness p.

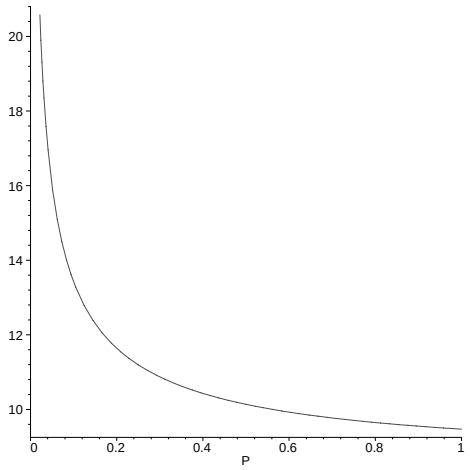


Figure 7: The averaged number of bases per hairpin in a random structure depending on the stickiness p.

resulting generating function by $(1-a+ba)^{-2}$. As done for the enumeration of hairpins we then have to set $z = z\sqrt{p}$ and $a = z$, take the first partial derivative with respect to b and set b to 1 afterwards. This procedure yields the generating function in question which possesses the representation

$$\frac{3-6z+3z^2-pz^2-2pz^3+\frac{-3+z(9+z(-9+3z+p(4+z-(5+p)z^2+2pz^3)))}{\sqrt{1-2z+z^2-2pz^2-2pz^3+p^2z^4}}}{2p(z-1)z}.$$

Its expansion around the dominant singularity $z_d(p)$ is given by (terms relevant for the asymptotic only)

$$\frac{\sqrt{1+4\sqrt{p}}+2}{2(1+\sqrt{1+4\sqrt{p}}+\sqrt{p})}\left(p+4p^{3/2}\right)^{-1/4}\left(1-\frac{z}{z_d(p)}\right)^{-1/2}+\mathcal{O}\left(\left(1-\frac{1}{z_d(p)}\right)^{1/2}\right).$$

The resulting asymptotic for the coefficient is given by

Lemma 4 *Under the assumption of the Bernoulli-model with a stickiness p, the expected number of bulges in all secondary structures of size n is asymptotically given by*

$$\frac{z_d(p)^{-n}}{\sqrt{\pi n}}\frac{\sqrt{1+4\sqrt{p}}+2}{2(1+\sqrt{1+4\sqrt{p}}+\sqrt{p})}\left(p+4p^{3/2}\right)^{-1/4}+\mathcal{O}\left(z_d(p)^{-n}n^{-3/2}\right),\ n\to\infty.$$

□

Averaging in the same way as before we find

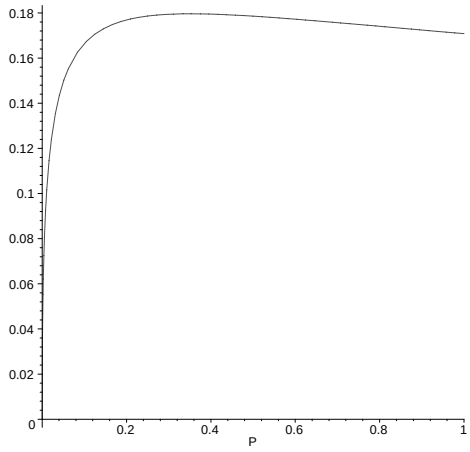


Figure 8: The slope of the number of bulges in a random structure depending on the stickiness p.

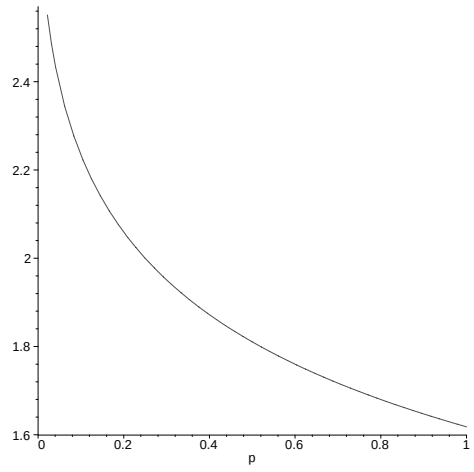


Figure 9: The quotient of the number of bulges and the number of hairpins depending on the stickiness p.

Theorem 4 *Under the assumption of the Bernoulli-model with a stickiness p, the averaged number of bulges in a secondary structure of size n is asymptotically given by*

$$\frac{4\sqrt{p}+\sqrt{1+4\sqrt{p}}-1}{2+\sqrt{1+4\sqrt{p}}(2+5\sqrt{p})+9\sqrt{p}+4p}n+\mathcal{O}(1),\ n\to\infty.$$

□

If we take a look at the plot of the slope of this quantity presented in Figure 8 we observe that at the beginning ($p = 1$) the number of bulges increases. We find numerically, that its maximum is given for $p = 0.351557\ldots$. For $p = 0.1266911\ldots$ it reaches again the level of $p = 1$. Figure 9 presents the quotient of the expected number of bulges and the expected number of hairpins which is given by

$$\frac{1+(2\sqrt{p}-1)\sqrt{1+4\sqrt{p}}}{2p}.$$

As we can see, there are always more bulges than hairpins, with decreasing p the bulges are spreading more and more.
Next, we will consider the length of the loops and bulges. This can be done in a similar way like the one used to enumerate the number of hairpins and bulges. Instead of marking an entire loop or bulge by a variable, we have to mark the bases of them. For the hairpins this means to replace $\frac{a}{1-a}$ by $\frac{ha}{1-ha}$ within our substitutions for x, u and v. Again, we set $z = z\sqrt{p}$, $a = z$, take the first partial

derivative with respect to h and set $h = 1$ afterwards. We find

$$\frac{z(1 - z + pz^2 - \sqrt{1 - 2z + z^2 - 2pz^2 - 2pz^3 + p^2z^4})}{2(1-z)^2\sqrt{1 - 2z + z^2 - 2pz^2 - 2pz^3 + p^2z^4}}.$$

As expected, the dominant singularity is located at $z_d(p)$ where the function possesses the expansion (terms relevant for the asymptotic only)

$$\frac{(\sqrt{1+4\sqrt{p}}-1)^{-1}(1+\sqrt{1+4\sqrt{p}})^3}{2(3+\sqrt{1+4\sqrt{p}})^2(p+4p^{3/2})^{1/4}}\left(1-\frac{z}{z_d(p)}\right)^{-1/2}+\mathcal{O}\left(\left(1-\frac{z}{z_d(p)}\right)^{1/2}\right).$$

Obviously, this expansion can be translated into an asymptotic for the coefficients and we can average by dividing the resulting expression by the expected number of structures.
We find a generating function where each unpaired base of a bulge is marked by variable b by replacing the geometric series $\frac{1}{1-a}$ corresponding to the insertion of $|^\star$ by $\frac{1}{1-ba}$. Again, we must take care of the overestimation which now can be corrected by the factor $(\frac{1-ba}{1-a})^2$. The procedure to find an asymptotic for the coefficients and the averaged number of bulges remains the same and its presentation is therefore left out. At the end, we find that (in the leading term of the asymptotic) the averaged loop-length and the averaged bulge-length behave in exactly the same way. We have

Theorem 5 *Under the assumption of the Bernoulli-model with a stickiness $p > 0$, the averaged asymptotical length of a hairpin-loop in a secondary structure of size n has the same leading term as the averaged asymptotical length of a bulge in a secondary structure of size n and is given by*

$$\frac{1+\sqrt{1+4\sqrt{p}}+2\sqrt{p}-2p}{4\sqrt{p}-2p}+\mathcal{O}(n^{-1}),\ n\to\infty.$$

□

Note that this result implies that the ratio plotted in Figure 9 can also be considered as the ratio of unpaired bases residing in bulges compared to those residing in hairpin-loops. In [13] the loop-length and the length of bulges were determined for the combinatorial model, i.e. for the case $p = 1$. It was shown there, that the difference in the second order term of the asymptotical representations for the two quantities is rather small. A plot of the *expected* loop-length is presented in Figure 10. The realistic lower bound of 3 is reached at $p = \frac{1}{4}(\sqrt{6}-2)^2 = 0.07576\ldots$. Such a small value for the stickiness can only result from a quite asymmetric distribution of the bases. Before we will consider the effect of the stickiness to the order of a secondary structure we want to estimate the number of unpaired bases residing in the tails. The generating function enumerating this quantity results from subtracting the generating functions enumerating the unpaired bases

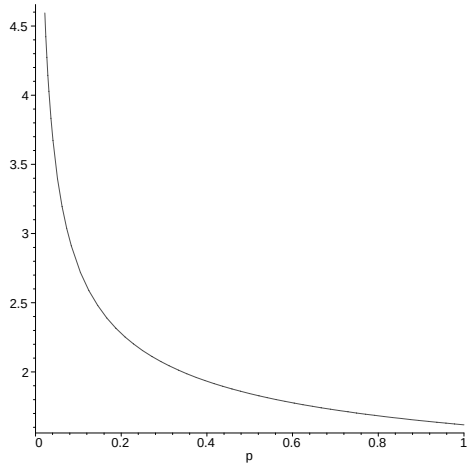


Figure 10: The length of a loop in a random secondary structure depending on the stickiness p.

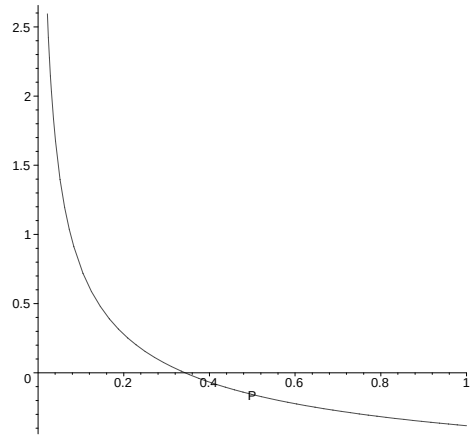


Figure 11: The averaged number of unpaired bases residing in the tails of a secondary structure minus the averaged length of a hairpin-loop depending on the stickiness p.

in loops and bulges from the generating function which enumerates all unpaired bases and is given by

$$\frac{1-2z+z^2-z^2p-z^3p-(1-z)\sqrt{1+z(-2+z(1+p(-2+z(-2+pz))))}}{pz(1-z)^2}.$$

Note that it is impossible to just subtract the asymptotical numbers of unpaired bases in loops and bulges from the asymptotical total number of unpaired bases to get a result, since the number of unpaired bases located in the tails is of smaller order and is therefore hidden within the $\mathcal{O}$-terms of these asymptotics. As a consequence, this subtraction simplifies to zero. The expansion of the generating function around its dominant singularity $z_d(p)$ is given by (terms relevant for the asymptotic only)

$$\frac{-16(1+4\sqrt{p})^{1/4}\sqrt{1+\sqrt{1+4\sqrt{p}}}}{(\sqrt{1+4\sqrt{p}}-1)^{3/2}(2\sqrt{1+4\sqrt{p}}+4\sqrt{p}-2)}\left(1-\frac{z}{z_d(p)}\right)^{1/2}+\mathcal{O}\left(\left((1-\frac{z}{z_d(p)}\right)^{3/2}\right).$$

Applying the $\mathcal{O}$-transfer method and dividing by the expected number of secondary structures of size n finally yields

Theorem 6 *Under the assumption of the Bernoulli-model with a stickiness p, the averaged number of unpaired bases located in the tails of a secondary structure of size n is asymptotically given by*

$$\frac{8}{4\sqrt{p}+2\sqrt{1+4\sqrt{p}}-2}+\mathcal{O}(n^{-1}),\ n\to\infty.$$

□

If we take a look at the plot of this asymptotical number minus the asymptotical number of bases located in a loop as presented in Figure 11 we observe that for $p = 1$ there are less bases located in both tails than in a random loop. We find that for $p = (2 - \sqrt{(2)})^2 = 0.3431\ldots$ both quantities become the same, for any value smaller than this the number of bases in the tails becomes larger than the number of unpaired bases in a random loop.
The last parameter that we want to consider is the order of a secondary structure. While all the parameters considered so far were related to specific substructures of the molecule, the order gives some information about the overall shape of it. We will determine the expected order (the quotient of the expected sum of orders of all structures divided by the expected number of structures) of a structure of size n and the expected number of secondary structures of size n and order k both depending on the stickiness p. In order to compute the first quantity we use the rightmost representation of $\mathbf{R}_k(x, u, v)$ given in (1) together with the substitutions given in (2) and (3). Furthermore we set $z = z\sqrt{p}$ and afterwards $a = z$ which yields the following representation for the generating function $R_k^{(p)}(z)$, counting the expected number of secondary structures of order k

$$R_k^{(p)}(z) = \frac{\sqrt{z}}{1-z}\frac{(1-\varpi)\,\varpi^{2^{k-1}}}{\sqrt{\varpi}\left(1-\varpi^{2^k}\right)},$$

$$\varpi := (1-\epsilon)/(1+\epsilon),\ \epsilon := \sqrt{1-4\frac{p^2z^5}{(-1+z(2+z(-1+p+pz)))^2}}.$$

We have to consider $[z^n]\sum_{k\geq 1} kR_k^{(p)}(z)$ which provides the expected sum of the orders of all secondary structures of size n. By trivial expansions we find

$$\sum_{k\geq 1} kR_k^{(p)}(z) = \frac{\sqrt{z}}{1-z}\frac{1-\varpi}{\sqrt{\varpi}}\underbrace{\sum_{\substack{k\geq 1\\ j\geq 0}} k\varpi^{2^{k-1}(1+2j)}}_{=:\sigma(\varpi)}.$$

The method of choice to handle sums like σ is the Mellin summation technique surveyed in [3]. This technique is based on a direct correspondence between the asymptotic expansion of a function (at either 0 or ∞) and the singularities of the transformed function. Furthermore, the so-called *separation property* of the Mellin transform which leads to a closed-form representation of the transform of certain kind of sums (so-called *harmonic sums*) using zeta functions is essential. In our specific example we compute the Mellin transform of $\sigma(e^{-t})$ which is given by

$$\mathcal{M}(s) := \frac{2^s}{2^s-1}\zeta(s)\Gamma(s),\ \Re(s) > 1,$$

for ζ the Riemann zeta function and $\Gamma(s)$ the complete gamma function. $\mathcal{M}(s)$ possesses poles for $s=1$, $s=0$, $s=-k$, $k\in\mathbb{N}$ and $s=\frac{2\pi ik}{\ln(2)}=:\chi_k$, $k\in\mathbb{Z}\backslash\{0\}$. According to the methodology an expansion of $\sigma(e^{-t})$ for $t\to 0$ is given by the sum of the residues of $\mathcal{M}(s)t^{-s}$ at these poles. We find for those residues:

$$\begin{aligned} s=1 &: 2t^{-1}, \\ s=0 &: \frac{2\ln(t)+2\gamma-2\ln(\pi)-3\ln(2)}{4\ln(2)}, \\ s=-1 &: -\frac{1}{12}t, \\ s=-3 &: \frac{1}{5040}t^3 \text{ and} \\ s=\chi_k &: \frac{\zeta(\chi_k)\Gamma(\chi_k)}{\ln(2)}t^{-\chi_k}. \end{aligned}$$

In general, the residue for $s=-2n$, $n\in\mathbb{N}$, is 0 and that for $s=-2n-1$, $n\in\mathbb{N}_0$, is in $\mathcal{O}(t^{2n+1})$. The expansion of $\sigma(e^{-t})$ at $t=0$ given by the sum of these expressions must be translated into an expansion around the dominant singularity of $\sum_{k\geq 1}kR_k^{(p)}(z)$. The location of this singularity can be determined in the following way: Obviously,

$$[z^n]T(z\sqrt{p},z)\leq[z^n]\sum_{k\geq 1}kR_k^{(p)}(z)\leq n[z^n]T(z\sqrt{p},z)$$

holds. Since $T(z\sqrt{p},z)$ and $z\frac{d}{dz}T(z\sqrt{p},z)$ have the same dominant singularity $z_d(p)$ and since $[z^n]z\frac{d}{dz}T(z\sqrt{p},z)=n[z^n]T(z\sqrt{p},z)$, we can conclude that $z_d(p)$ is the dominant singularity of $\sum_{k\geq 1}kR_k^{(p)}(z)$, too. At $z=z_d(p)$ the square root ϵ evaluates to zero. Thus we expand $t=-\ln((1-\epsilon)/(1+\epsilon))$ around $\epsilon=0$ to find that $t=2\epsilon+\mathcal{O}(\epsilon^3)$. By expanding ϵ around $z_d(p)$ we finally find

$$t=\underbrace{(3+\sqrt{1+4\sqrt{p}})\left(\frac{1}{p}+\frac{4}{\sqrt{p}}\right)^{1/4}}_{=:c_1}\left(1-\frac{z}{z_d(p)}\right)^{1/2}+\mathcal{O}\left(\left(1-\frac{z}{z_d(p)}\right)^{3/2}\right). \quad (4)$$

The factor $\frac{\sqrt{z}}{1-z}\frac{1-\varpi}{\sqrt{\varpi}}$ possesses the following leading term of its expansion at $z_d(p)$

$$\underbrace{4\frac{(1+4\sqrt{p})^{1/4}\sqrt{1+\sqrt{1+4\sqrt{p}}}}{(\sqrt{1+4\sqrt{p}}-1)^{3/2}}}_{=:c_2}\left(1-\frac{z}{z_d(p)}\right)^{1/2}. \quad (5)$$

By setting t to its equivalent (4) within all the residues given above, summing the resulting expressions and multiplying with the expansion (5) we finally find

the expansion of our generating function around its dominant singularity. This expansion is given by

$$2\frac{c_2}{c_1}+\frac{c_2\left(1-\frac{z}{z_d(p)}\right)^{1/2}\ln\left(1-\frac{z}{z_d(p)}\right)}{4\ln(2)}+\frac{2\ln(c_1)+2\gamma-\ln(\pi^2)-\ln(8)}{4\ln(2)}c_2\left(1-\frac{z}{z_d(p)}\right)^{1/2}$$

$$+\frac{c_2}{\ln(2)}\sum_{k\neq 0}\zeta(\chi_k)\Gamma(\chi_k)e^{-2\pi ik\log_2(c_1)}\left(1-\frac{z}{z_d(p)}\right)^{\frac{1-\chi_k}{2}}.$$

This expansion can be translated into an asymptotic for the coefficients by means of the $\mathcal{O}$-transfer method. By dividing the resulting expressions by the expected number of secondary structures of size n we finally find

Theorem 7 *Under the assumption of the Bernoulli-model with a stickiness p, the averaged order of a secondary structure of size n is asymptotically given by*

$$\log_4\left(\frac{32\pi^2}{(3+\sqrt{1+4\sqrt{p}})^2}\sqrt{\frac{p}{1+4\sqrt{p}}}n\right)-\frac{\gamma+2}{2\ln(2)}+\Theta\left(\log_2\left(\frac{n}{c_1^2}\right)\right)+\mathcal{O}(n^{-1}),$$

$n\to\infty$. Here, $c_1=(3+\sqrt{1+4\sqrt{p}})\left(\frac{1}{p}+\frac{4}{\sqrt{p}}\right)^{1/4}$ and $\Theta(x)$ is the periodic function of very small modulus with the following Fourier series:

$$\Theta(x)=\frac{1}{\ln(2)}\sum_{k\neq 0}(\chi_k-1)\Gamma\left(\frac{\chi_k}{2}\right)\zeta(\chi_k)e^{\pi ikx},\ \chi_k:=\frac{2\pi ik}{\ln(2)}.$$

□

Only the constant term of this asymptotic and the period of Θ depend on p. Thus the stickiness affects only marginally the averaged order of a random secondary structure. Figure 12 shows how a change of the stickiness relocates Θ, in Figure 13 you find a plot of the constant term without the contribution of Θ against the stickiness p.

Finally, we will study the expected number of structures of given order k. For that purpose we return to the representation of $\mathbf{R}_k(x,u,v)$ based on the Chebyshev polynomial as given in (1). Applying the substitutions given in (2) and (3), setting z to $z\sqrt{p}$ and setting a to z afterwards yields the following representation of the generating function of the expected number of secondary structures of order k

$$R_k^{(p)}(z)=\frac{\sqrt{z}}{z-1}U_{2^k-1}^{-1}\left(\frac{-1+2z-z^2+pz^2+pz^3}{2z^{5/2}p}\right).$$

The dominant singularity $z_k(p)$ of that function results from a zero of the Chebyshev polynomial for which it is known that

$$U_n\left(\cos\left(\frac{\pi m}{n+1}\right)\right)=0,\ 1\leq m\leq n,$$

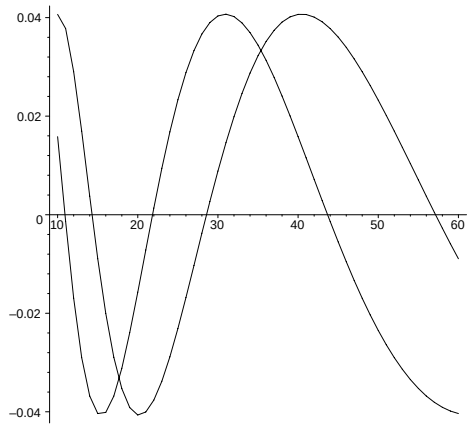


Figure 12: A plot of $\Theta(x)$ for $p = \frac{1}{6}$ (right) and $p = \frac{2}{3}$ (left).

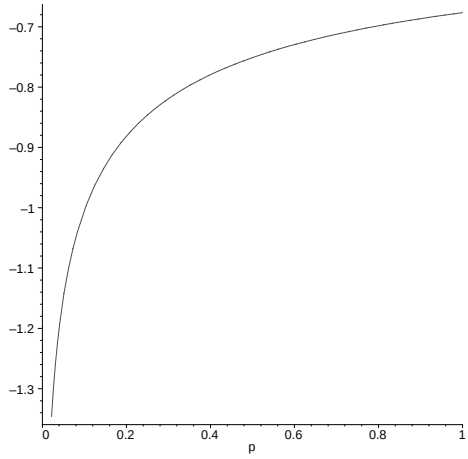


Figure 13: The constant term (without $\Theta(x)$) of the averaged order depending on the stickiness p.

holds. Therefore we have to find the smallest solution of

$$\underbrace{\frac{-1+2z-z^2+pz^2+pz^3}{2z^{5/2}p}}_{=:f(z,p)} = \cos(m2^{-k}\pi),\ 1 \le m \le 2^k - 1.$$

This is not a trivial task, since it is equivalent to determine the roots of a polynomial of sixth degree. However, by discussing the equations $f(z,p) = \cos(m2^{-k}\pi)$ for $1 \le m \le 2^k - 1$, it will be possible to get a quite precise approximation of this solution. Assuming that $p > 0$, we find for all p that $f(z,p) = -1$ has the smallest solution $z = \frac{1-\sqrt{1+4\sqrt{p}+2\sqrt{p}}}{2p} =: z_\ell(p)$. Furthermore, the smallest solution of $f(z,p) = 1$ is given by $z = 1$ for all p. Now, since $f'(z) \ge 0$ for $z \in [z_\ell(p), 1]$ and all p, we know that f is a monotone increasing function within this interval. Thus, the smallest solutions of $f(z,p) = \cos(m2^{-k}\pi)$ result from the choice $m = 2^k - 1$ since this minimizes the value of the cosine. So the dominant singularity $z_k(p)$ is the smallest real solution of $f(z,p) = \cos((2^k-1)2^{-k}\pi) = -\cos(2^{-k}\pi)$. Implied by properties of the cosine we find that the sequence $f(z_k(p),p)$ is monotone decreasing with respect to k. Furthermore, we can argue that $z_k(p)$ is a monotone decreasing sequence itself by means of the positivity of the first derivative f' for all values in the interval $[z_\infty(p), z_1(p)] \subseteq [z_\ell(p), 1]$. Since the value of $f''(z,p)$ is negative for all p and all z in that interval, we can conclude that $\frac{f(z_k(p))-f(z_\infty(p))}{z_k(p)-z_\infty(p)} \ge f'(z_k(p))$ for all possible k and p. Now setting $f(z_k(p)) = -\cos(2^{-k}\pi)$ and expanding the cosine finally proves $0 \le z_k(p) - z_\infty(p) \le 4^{-k}$ for all p. Note that $z_k(p) \xrightarrow{p\to 0} 1$ for each k.

The next step is to find an expansion of $R_k^{(p)}(z)$ around $z_k(p)$. Based on the

representation [1, 22.3.16]

$$U_n(x) = \frac{\sin((n+1)\arccos(x))}{\sin(\arccos(x))}$$

we find

$$\lim_{z \to z_k(p)} \frac{\left(1 - \frac{z}{z_k(p)}\right)}{(f(z,p) + \cos\left(2^{-k}\pi\right))} = \frac{-4 z_k(p)^{5/2} p}{5 - 6 z_k(p) + z_k(p)^2 - p z_k(p)^2 + p z_k(p)^3} \tag{6}$$

and

$$\lim_{x \to -\cos\left(2^{-k}\pi\right)} \left(x + \cos\left(2^{-k}\pi\right)\right) \frac{\sin(\arccos(x))}{\sin(2^k \arccos(x))} = -2^{-k} \sin^2(2^{-k}\pi).$$

Together with the factor $\sqrt{z}/(z-1)$ we find in that way the expansion of $R_k^{(p)}(z)$ at $z_k(p)$

$$\frac{1}{\left(1 - \frac{z}{z_k(p)}\right)} \frac{-4 z_k(p)^3 p 2^{-k} \sin^2(2^{-k}\pi)}{(1 - z_k(p))^2 (p z_k(p)^2 + z_k(p) - 5)},$$

and by means of the $\mathcal{O}$-transfer the following result

Theorem 8 *Under the assumption of the Bernoulli-model with stickiness $p > 0$, the expected number of secondary structures of size n and order k is asymptotically given by*

$$z_k(p)^{-n} \frac{-4 z_k(p)^3 p 2^{-k} \sin^2(2^{-k}\pi)}{(1 - z_k(p))^2 (p z_k(p)^2 + z_k(p) - 5)},\ n \to \infty.$$

Here, $z_k(p)$ is the smallest real solution of the equation

$$\frac{-1 + 2z - z^2 + pz^2 + pz^3}{2 z^{5/2} p} = -\cos(2^{-k}\pi).$$

The following approximation for $z_k(p)$ holds:

$$0 \leq z_k(p) - \frac{1 - \sqrt{1 + 4\sqrt{p}} + 2\sqrt{p}}{2p} \leq 4^{-k}.$$

□

Some numerical values of $z_k(p)$ can be found in the table of Figure 14. By means of Theorem 8 it becomes possible to connect the result for the expected order to some sort of density function which results from the quotient of the expected number of structures of order k and the expected number of structures. A plot of that density is shown in Figure 15. The different curves correspond to structures of size $n = 100$ and the probabilities $p = \frac{j}{10}$ for $j = 1, 2 \ldots, 10$ from left to right.

	p									
k	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$	$\frac{4}{10}$	$\frac{5}{10}$	$\frac{6}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	1
1	0.70759	0.63611	0.59138	0.55877	0.53317	0.51215	0.49437	0.47898	0.46545	0.45339
2	0.65404	0.57850	0.53270	0.50000	0.47470	0.45417	0.43696	0.42220	0.40930	0.39788
3	0.64155	0.56537	0.51951	0.48690	0.46176	0.44140	0.42438	0.40980	0.39708	0.38584
4	0.63848	0.56216	0.51629	0.48371	0.45861	0.43831	0.42133	0.40680	0.39413	0.38292
5	0.63772	0.56136	0.51549	0.48292	0.45783	0.43754	0.42058	0.40605	0.39339	0.38220
6	0.63753	0.56116	0.51529	0.48272	0.45764	0.43735	0.42039	0.40587	0.39321	0.38202
7	0.63748	0.56111	0.51524	0.48267	0.45759	0.43730	0.42034	0.40582	0.39317	0.38198
8	0.63747	0.56110	0.51523	0.48266	0.45758	0.43729	0.42033	0.40581	0.39315	0.38196
9	0.63746	0.56110	0.51523	0.48266	0.45757	0.43728	0.42032	0.40581	0.39315	0.38196
10	0.63746	0.56110	0.51523	0.48266	0.45757	0.43728	0.42032	0.40581	0.39315	0.38196
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	0.63746	0.56110	0.51523	0.48266	0.45757	0.43728	0.42032	0.40581	0.39315	0.38196

Figure 14: Some numerical values of $z_k(p)$ together with the corresponding approximation for $k \to \infty$ (last row).

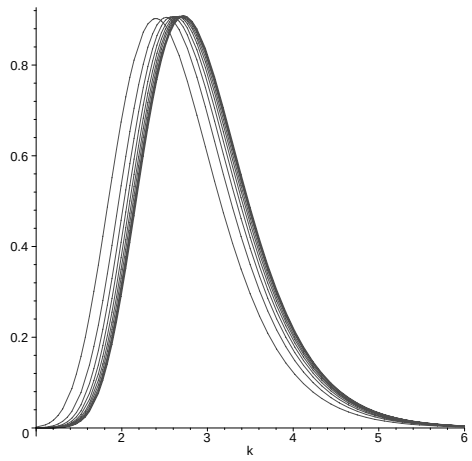


Figure 15: The density of the order for structures of size 100 and $p = \frac{j}{10}$, $j = 1, 2, \ldots, 10$ (from left to right).

We observe that the density is quite peaky such that we have to expect a small variance. Furthermore, in accordance with our results for the averaged order the influence of the stickiness p is rather poor. We want to conclude this section by remarking that for the case $p = 1$ the variance has been determined in [13]. It was shown there, that the variance is asymptotically given by $0.17939\ldots$ plus an oscillating function of small modulus.

	total	average (percentage)
number of unpaired bases	1542431	(52.23% of all bases)
number of paired bases	1410624	(47.77% of all bases)
number of hairpins	66391	52.19 hairpins per structure
number of bulges	232955	183.14 bulges per structure
length of hairpin-loops	489623	7.3748 bases per loop
length of bulges	1052806	4.5194 bases per bulge
order	6044	4.7516

Table 2: Statistical data computed from 1272 entries of the long subunit ribosomal database of Wuyts et el. .

4 Comparison to Real Secondary Structures

In this section we will compare our results for the Bernoulli-model to the data contained in the long subunit ribosomal RNA database of Wuyts et al. . For this purpose we take 1272 sequences contained in the database together with the encoding of their secondary structure and analyze parameters like the average proportion of paired to unpaired bases or the average number of hairpins of a structure. We used a set of simple counting programs in order to compute the total numbers and averages of interest. The corresponding results can be found in Table 2. For the Bernoulli-model we assume that the length n of the structures is equal to the average length of the sequences taken from the database (which is given by 2321.58). Then, we compute the resulting values for the same parameters from our asymptotic formulæ for reasonable choices for the stickiness[1] p. Afterwards we determine those values for p which make our formulæ equal to the corresponding averages of the real world data. All the resulting quantities can be found in Table 3. If we compare the two tables we find that some parameters fit for a reasonable choice of the stickiness p while others are completely out of scope. For instance, the average proportion of paired to unpaired bases in real world molecules equals the one for the Bernoulli-model for $p = 0.44$, for $p = 0.25$, i.e. for the Bernoulli-model with a uniform distribution of the four bases and Watson-Crick base pairs only, the Bernoulli-model is quite close to the real world data (57.74% compared to 52.23% of unpaired bases). Surprisingly, the same holds for the average order of a molecule. Even if this parameter is somehow related to the (global) planar topology of the entire structure, both values coincide for a stickiness $p = 0.2626$, for $p = 0.25$ their difference is given

[1] $p = 1$ corresponds to the combinatorial model, $p = \frac{1}{2}$ corresponds to a binary alphabet of complementary bases. The case $p = \frac{1}{4}$ considers a uniform distribution of the four bases together with Waston-Crick pairings only, $p = \frac{3}{8}$ takes all canonical pairing into account.

	p				
	1	$\frac{1}{2}$	$\frac{3}{8}$	$\frac{1}{4}$	$=$
percentage of unpaired bases	44.72	51.11	53.84	57.74	0.44
percentage of paired bases	55.28	48.89	46.16	42.26	0.44
average number of hairpins per structure	245.10	229.00	220.62	207.36	0.0028
average number of bulges per structure	396.57	414.59	416.94	414.71	0.0050
average length of a hairpin-loop	1.62	1.84	1.96	2.15	0.0066
average length of a bulges	1.62	1.84	1.96	2.15	0.0227
average order of a structure	4.91	4.84	4.80	4.74	0.2626

Table 3: Parameters computed for the Bernoulli-model assuming a structure-size of n=2321.58. The column labeled with = contains those values for p which make the results for the Bernoulli-model equal to those for the real world structures as given in Table 2. The values for the average order were computed without taking the oszillation into account.

by 0.01. Therefore it is justified to belive that the Bernoulli-model is accurate with respect to the average order. Recall that the order only depends marginally on the order which might be the reason for this observation. The other parameters which were considered do not fit at all. The average number of hairpins and the average number of bulges of a structure in the Bernoulli-model equal those of the RNA database for a stickiness much smaller than 1/100. Such a small value for the stickiness is absolutely inadequate. If, for example, we assume a stickiness of 0.0028 as given by an equality of the average number of hairpins, we would get an average proportion of 90.85% of unpaired bases to all bases within the Bernoulli-model. We also find inadequate values for the stickiness in case of the average length of a hairpin-loop and the average length of a bulge, but the order of magnitude differs in both cases by a factor about 3.4. This might be due to the unrealistic assumption of a minimal loop length of 1 for our computations. However, the assumption of a minimal loop-length of for example 3 won't give much change to the observed behavior.
Within our model, a minimal loop-length of $m > 0$ can be considered by the following substitutions:

$$v := z^2 \frac{a^m}{(1-a)^3},\ u := \frac{1}{2} z^2 \left(\frac{a^m}{(1-a)^2} + \frac{1}{(1-a)^2} \right) \text{ and } x := z^2 \frac{1}{1-a}.$$

The assumption of $m = 3$ and a stickiness 0.25 yield an averaged expected loop-length of

$$4.3862\ldots$$

which can be concluded by the same computations as performed for $m = 1$. Thus, even in this *more realistic* setting, the average loop-length in the Bernoulli-model with $p = 0.25$ is only about 60% of the average for the real world data. The average length of a bulge in the Bernoulli-model achieves about 47.6% of the real average bulge length. Since the number of hairpin-loops and bulges is far too large in the Bernoulli-model while the relation of paired to unpaired bases seems to be realistic, we have to expect one major structural difference between structures in the model and real molecules within the length of the ladders. Only for structures with mostly short ladders it is possible to have a huge number of loops and bulges. However, those structures are unstable for a lack of stacking and will therefore not occur in reality. We just want to remark that we also compared our results for the Bernoulli-model to the tRNA database of Sprinzl et al. [17]. The observations made there were quite similar, we thus resigned to present them in detail.

Only the proportion of unpaired bases and the order behave realistic within the Bernoulli-model. Both, in some sense, are global parameters which are determined by information on the entire structure. For such parameters it seems to be sufficient to consider only the underlying combinatorial structure together with some pairing probabilities (as done in the Bernoulli-model) in order to model them in a realistic way. Parameters, which are related to details of the structures and thus are of interest e.g. in relation to the prediction of secondary structures, behave totally unrealistic in our model. We thus have to conclude that significant factors which determine the details of a real secondary structure are not taken into account. The above mentioned disregarded minimal length of stable ladders seems to be one major weakness of the Bernoulli-model in this respect. As already Zuker and Sankoff [27] pointed out, the Bernoulli-model also considers structures which contain pairs of bases which are not joint by a hydrogen bound even though they are in stereochemically favorable positions for base pairing. This is of course an unrealistic behavior of our model, but to consider only satutated[2] structures without any other change of the model would decrease the length of the loops and thus would affect the results into a false direction. Therefore, the search for a realistic model must try to take more details into account than only the existence of complementary bases together with their pairing probabilities.

[2] A model for saturated structures was recently presented at GCB'01 by Evers et al.[2].

5 Conclusions

In this paper we have investigated the Bernoulli-model for RNA secondary structures. Compared to similar considerations of Hofacker et al. [6] we have determined different parameters with different methods. Furthermore, we have compared our results to real world data in order to judge the quality of the model. As already pointed out by Zuker and Sankoff [27], the Bernoulli-model is more realistic than the pure combinatorial point of view which has been considered by numerous authors (see e.g. [6, 13, 15, 20, 22, 24]). However, our comparison of the Bernoulli-model to the data of the large subunit ribosomal RNA database of Wuyts et al. [25] proved that many details of structures in the model like the length and the number of their hairpin-loops or the length of their ladders are far from being realistic. Surprisingly, not only the proportion of unpaired bases but also the order of a structure seems to behave quite realistic in the Bernoulli-model. Of course, all these studies can only be a starting-point for investigations which must try to consider more details of the real structures' folding mechanism. Even if our results do not have a direct influence on applications, the methodology used to derive them is of independent interest and may be of use for future work on RNA structure and related problems.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, 1970.

[2] D. J. Evers and R. Giegerich, *Reducing the Conformation Space in RNA Structure Prediction*, German Conference on Bioinformatics 2001.

[3] P. Flajolet, X. Gourdon and P. Dumas, *Mellin transforms and asymptotics: Harmonic sums*, Theoretical Computer Science **144** (1995), 3-58.

[4] P. Flajolet and A. Odlyzko, *Singularity Analysis of Generating Functions*, SIAM J. Disc. Math. **3** (1990), No. 2, 216-240.

[5] P. Flajolet and R. Sedgewick,*The Average case analysis of algorithms: complex asymptotics and generating functions*, INRIA rapport de recherche **2026**, 1993.

[6] I. L. Hofacker, P. Schuster and P. F. Stadler, *Combinatorics of RNA secondary structures*, Discrete Applied Mathematics **88** (1998), 207-237.

[7] R. E. Horton, *Erosioned development of systems and their drainage basins, hydrophysical approach to quantitative morphology*, Bull. Geol. Soc. of America **56** (1945), 275-370.

[8] J. A. Howell, T. F. Smith and M. S. Waterman, *Computation of Generating Functions for Biological Molecules*, SIAM J. Appl. Math. **39** (1980), 119-133.

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

[10] S. Mainville, *Comparaisons et Auto-comparaisons de Chaînes Finies*, Ph. D. thesis, Université de Montréal, Canada, 1981.

[11] Mitiko Gô, *Statistical Mechanics of Biopolymers and Its Application to the Melting Transition of Polynucleotides*, Journal of the Physical Society Japan **23** (1967), 597-608.

[12] M. E. Nebel, *A Unified Approach to the Analysis of Horton-Strahler Parameters of Binary Tree Structures*, Random Structures & Algorithms, to appear.

[13] M. E. Nebel, *Combinatorial Properties of RNA secondary Structures*, Journal of Computational Biology, to appear.

[14] J. M. Pipas and J. E. McMahon, Method for predicting RNA secondary structures, Proc. Nat. Acad. Sci., U.S.A. **72** (1975), 2017-2021.

[15] M. Régnier, *Generating Functions in Computational Biology: a Survey*, submitted.

[16] W. R. Schmitt and M. S. Waterman *Linear trees and RNA secondary structure*, Discrete Applied Mathematica **51** (1994), 317-323.

[17] M. Sprinzl, K.S. Vassilenko, J. Emmerich and F. Bauer, *Compilation of tRNA sequences and sequences of tRNA genes*, (20 December, 1999) http://www.uni-bayreuth.de/departments/biochemie/trna/.

[18] P. R. Stein and M. S. Waterman *On some new sequences generalizing the Catalan and Motzkin numbers*, Discrete Mathematics **26** (1978), 261-272.

[19] A. N. Strahler, *Hypsometric (area-altitude) analysis of erosonal topology*, Bull. Geol. Soc. of America **63** (1952), 1117-1142.

[20] G. Viennot and M. Vauchaussade de Chaumont, *Enumeration of RNA Secondary Structures by Complexity*, Mathematics in medecine and biology, Lecture Notes in Biomaths. **57** (1985), 360-365.

[21] X. G. Viennot, *Trees Everywhere*, Proc. CAAP'90, LNCS **431** (1990), 18-41.

[22] M. S. Waterman, *Secondary Structure of Single-Stranded Nucleic Acids*, Advances in Mathematics Supplementary Studies **1** (1978), 167-212.

[23] M. S. Waterman and T. F. Smith, *RNA Secondary Structures: A Complete Mathematical Analysis*, Mathematical Biosciences **42** (1978), 257-266.

[24] M. S. Waterman, *Combinatorics of RNA hairpins and cloverleaves*, Studies in Applied Mathematics **1** (1978), 91-96.

[25] Wuyts J., De Rijk P., Van de Peer Y., Winkelmans T., De Wachter R., *The European Large Subunit Ribosomal RNA database*, Nucleic Acids Res. **29** (2001), 175-177.

[26] S. Zaks, *Lexicographic Generation of Ordered Trees*, Theoretical Computer Science **10** (1980), 63-82.

[27] M. Zuker and D. Sankoff, *RNA Secondary Structures and Their Prediction*, Bulletin of Mathematical Biology **46** (1984), 591-621.