Combinatorial Properties of RNA Secondary Structures

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

Abstract

The secondary structure of a RNA molecule is of great importance and possesses influence, e.g. on the interaction of tRNA molecules with proteins or on the stabilization of mRNA molecules. The classification of secondary structures by means of their *order* proved useful with respect to numerous applications. In 1978 Waterman, who gave the first precise formal framework for the topic, suggested to determine the number $a_{n,p}$ of secondary structures of size n and given order p. Since then, no satisfactory result has been found. Based on an observation due to Viennot et al. we will derive generating functions for the secondary structures of order p from generating functions for binary tree structures with Horton-Strahler number p. These generating functions enable us to compute a precise asymptotic equivalent for $a_{n,p}$. Furthermore, we will determine the related number of structures when the number of unpaired bases shows up as an additional parameter. Our approach proves to be general enough to compute the average order of a secondary structure together with all the r-th moments and to enumerate substructures such as hairpins or bulges in dependence on the order of the secondary structures considered.

1 Introduction and Definitions

In this paper we consider combinatorial models for molecules of single-stranded nucleic acids like RNA, mRNA (messenger RNA) or tRNA (transfer RNA). The sequence of bases of such a molecule is known as its *primary structure* and is usually encoded by a word on the alphabet $\{A, U, G, C\}$, where the different letters represent the involved bases Adenine, Uracyl, Guanine and Cytosine. This linear

structure is created by phosphodiester bounds between the bases. Since the bases can additionally be linked together by hydrogen bounds (Adenine with Uracyl and Guanine with Cytosine) the primary structure is folded into a planar graph which is called the *secondary structure* of the molecule. Figure 1 shows an example for such a graph. The secondary structure plays a role in the interaction

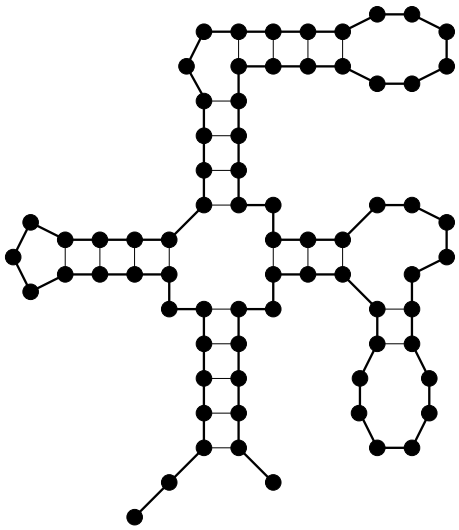


Figure 1: An example for a secondary structure. Each node of the graph represents one of the bases, a thick edge represents a phosphodiester bound, a thin edge represents a hydrogen bound.

of tRNA with proteins [19], in stabilizing mRNA and in packing RNA into virus particles. Many authors have payed attention to the prediction of the secondary structure from the knowledge of the primary structure. The general approach has been to search for configurations of minimum free energy. 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 approach for the efficient prediction of the secondary structure is based on the notion of *order* of a structure introduced in [24] (see also [16], [9]). The prediction-algorithm first constructs an optimal first-order structure. Then, using the results from the previous pass, successively higher order structures are computed in an iterative way. The algorithms of Waterman [24] and Mainville [15] can be seen to work this way.
Here we will focus on enumeration problems which are related to the secondary structure of single-stranded nucleic acids. This sort of studies has a long history starting with the investigations of Waterman [26] who gave the first formal framework for the topic [24]. As shown in [22] the sequences arising in the enumeration

of secondary structures which can occur under various reasonable restrictions may be considered as natural generalizations of the Catalan and Motzkin numbers. For most of the problems considered one finds similar decomposition patterns. This observation is used in [11] to describe algorithms and techniques for computing generating functions for certain RNA configurations. Like in our paper, the authors of [21] use different one-to-one correspondences between secondary structures and trees to derive enumeration results. Based on the traditional approach, i.e. based on the determination of recurrence relations from decomposition properties of the objects, the authors of [10] obtain several enumeration results for restricted configurations of secondary structures. Recently, Régnier showed how symbolic enumeration methods allow to simplify and extend previous results [20]. We will first recall the related definitions as introduced in [24]. Afterwards it will be possible to state precisely which problems are to be considered here. In the sequel we will forget about the different bases that emerge in the secondary structures. We will change to a complete combinatorial point of view where only the topology of the planar secondary structure is considered.

Definition 1 ([24]) *A* secondary structure of size n *is a loop free graph on the set of* n *labeled points* $\{1, 2, \ldots, n\}$ *such that the adjacency matrix* $A = (a_{i,j})$ *has the following three properties:*

(i) $a_{i,i+1} = 1$ *for* $1 \leq i \leq n-1$.

(ii) For each fixed i, $1 \leq i \leq n$, *there is at most one* $a_{i,j} = 1$ *where* $j \neq i \pm 1$.

(iii) If $a_{i,j} = a_{k,l} = 1$, *where* $i < k < j$, *then* $i \leq l \leq j$.

If $a_{i,j} = 1$, i *and* j *are said to be* bounded.

Please note that part (iii) of this definition ensures that the structure remains planar.

Definition 2 ([24]) *Suppose* A *is the adjacency matrix for a secondary structure of size* n.

(i) The point j *is said to be* paired *if there is some point* $i \neq j \pm 1$ *such that* $a_{i,j} = 1$.

(ii) The sequence $i+1, i+2, \ldots, j-1$ *is a* loop, *if* $i+1, i+2, \ldots, j-1$ *are all unpaired and* $a_{i,j} = 1$. *The pair* (i, j) *is said to be the* foundation of the loop.

(iii) The sequence $i+1, i+2, \ldots, j-1$ *is a* bulge *if* $i+1, i+2, \ldots, j-1$ *are all unpaired,* i *and* j *are both paired, and* $a_{i,j} \neq 1$.

(iv) A tail *is a sequence* $1, 2, \ldots, j$ *resp.* $j, j+1, \ldots, n$ *where* $1, 2, \ldots, j$ *resp.* $j, j+1, \ldots, n$ *are unpaired and* $j+1$ *resp.* $j-1$ *is paired.*

(v) *A* ladder *is built by two sequences* $i{+}1, i{+}2, \ldots, i{+}j$ *and* $k{+}1, k{+}2, \ldots, k{+}j$ *such that* $i + j + 1 \leq k$, $a_{i+l,k+j-l+1} = 1$ *for* $1 \leq l \leq j$, $a_{i,k+j+1} = 0$ *and* $a_{i+j+1,k} = 0$. *If* $i + j + 3 = k + 1$, *then this last requirement is dropped.*

(vi) *A* hairpin *is the longest sequence* $i + 1, i + 2, \ldots, j - 1$ *containing exactly one loop such that* $a_{i+1,j-1} = 1$ *and* $a_{i,j} = 0$. *The paired points* $i + 1$ *and* $j - 1$ *will be called the* foundation *of the hairpin.*

In a certain sense, the substructures previously defined can be considered as the building blocks for a secondary structure as can be seen be the next theorem.

Theorem 1 ([24]) *Any secondary structure can be uniquely decomposed into loops, ladders, bulges and tails. Alternatively, every secondary structure can be uniquely decomposed into hairpins and ladders, bulges, and tails which are not members of a hairpin.*

Next we will classify secondary structures by a certain complexity criterion.

Definition 3 ([24]) *Let* A *be the adjacency matrix of a secondary structure. A sequence* $A^{(i)}$ *of adjacency matrices of secondary structures is formed as follows:*

(i) $A^{(0)} := A$.

(ii) *We get* $A^{(i+1)}$ *from* $A^{(i)}$ *by setting* $a_{k,l}^{(i+1)} := a_{l,k}^{(i+1)} := 0$ *whenever* $a_{k,l}^{(i)} = a_{l,k}^{(i)} = 1$, k *and* l *are members of some hairpin, and* $k \neq l \pm 1$.

The secondary structure for A *is said to be of* k-th order *if* $A^{(k)}$ *is the first matrix in the sequence* $\{A^{(i)}\}_{i=0}^{\infty}$ *that has no hairpins.*

In [24] it is proven that every secondary structure possesses a unique order. Furthermore, the number $a_{n,1}$ of secondary structures of size n and order 1 was determined. The enumeration of the number $a_{n,p}$ of secondary structures of size n and order p was left open. Since then, several authors addressed to this problem. Viennot and Vauchaussade de Chaumont [23] used a methodology due to Schützenberger in order to prove the following:

Theorem 2 ([23]) *The generating function* $s_p(t) := \sum_{p \geq 0} a_{n,p} t^n$ *of secondary structures of order* p *is*

$$s_p(t) = \frac{t^{5\,2^{p-1}-2}}{(1-t)Z_1(t)\cdots Z_p(t)},$$

where $Z_1(t), \ldots, Z_p(t)$ *are the polynomials defined by the recurrence*

$$Z_1(t) = 1 - 2t - t^3,\ Z_{p+1}(t) = Z_p^2(t) - 2t^{5\,2^{p-1}}.$$

Unfortunately, no representation for the coefficient $a_{n,p}$ was concluded. Recently Hofacker et al. [10] had a paper where they considered enumeration problems related to RNA secondary structures and especially the number of secondary structures of size n and order p. However, they also did not get satisfactory results. In the present paper we will give a precise asymptotic estimated for the number $a_{n,p}$ and also for the number $a_{m,n,p}$ of secondary structures of order p built from n paired and m unpaired bases. Furthermore, we will derive the expected order of a secondary structure of size n, the related variance and the r-th moments which will also be computed for secondary structures of size n and m unpaired bases. Finally, we will determine the average number of hairpins and bulges in secondary structures of size n and in secondary structures of size n and order p and the corresponding expected length of the loops involved. For our presentation we will assume that the reader is familiar with generating functions and the tools usually denoted as *singularity analysis*. For details on that topics please refer to [7] and [6].

2 A Connection between Secondary Structures and Binary Trees

In this section we will present the connection between secondary structures and binary trees which will be the main tool for all the investigations that will be presented in the subsequent sections. For that purpose we first recall an equivalent definition for secondary structures which was introduced in [23]:

Definition 4 *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 notation a pair of corresponding brackets within a word w represents two bases of the single-stranded nucleic acid which are paired. The symbol | represents an unpaired base. For example, the secondary structure of Figure 1 is represented by the word

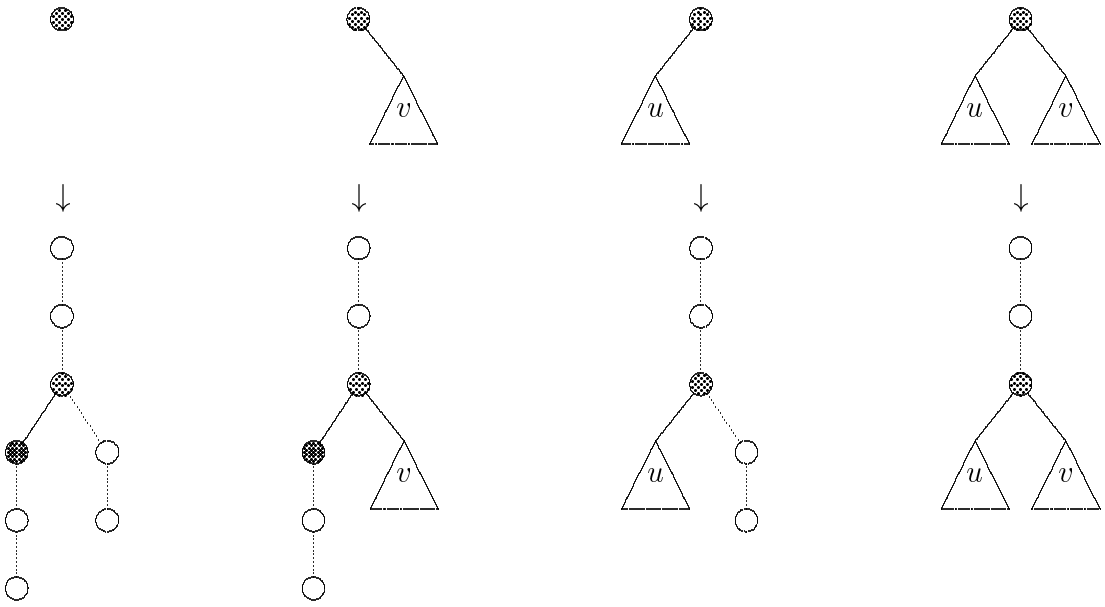
$$||((((((|((((|||))))((((||((((|||||||))))))))|(((|||||((||||||))))))|))))))|.$$

The words of $\Sigma^\star$, which satisfy the conditions (1) and (2) of the previous definition, are called *Motzkin words*. Condition (3) accommodates the fact, that two adjacent bases cannot be linked together by a hydrogen bound. Now let α denote

the homomorphism which deletes all symbols | in a word $w \in \Sigma^\star$. Then for each secondary structure w, $\alpha(w)$ is a semi Dyck-word, i.e. a word over the alphabet $\{(,)\}$ which fulfills the conditions (1) and (2) of Definition 4. The following one-to-one correspondence of semi Dyck-words of length $2n$ and ordered binary trees with n nodes is well-known (see e.g. [27]):

$$(u)v \Longleftrightarrow \begin{array}{c} \circ \\ / \ \backslash \\ u \quad v \end{array}.$$

To put the meaning of this symbolic notation into words, the first bracket in the semi Dyck-word together with its corresponding closing bracket represent the root of the tree, the subword u (resp. v) represents the left (resp. right) subtree (and vice versa). Note, that u and v are semi Dyck-words so that this correspondence is continued recursively. We get a one-to-one correspondence of secondary structures and ordered unary/binary trees by considering the inverse homomorphism α^{-1}. For $\mathcal{D}$ the semi Dyck-language and $\mathcal{S}$ the language of all secondary structures we find that $\beta := \alpha^{-1}(\mathcal{D}) \cap \mathcal{S}$ implies the following cases:
$() \xrightarrow{\beta} |^\star(|^+)|^\star \qquad ()v \xrightarrow{\beta} |^\star(|^+)v \qquad (u) \xrightarrow{\beta} |^\star(u)|^\star \qquad (u)v \xrightarrow{\beta} |^\star(u)v.$
These cases correspond to the following transformations for trees one-to-one:



Here, a light shaded node descends from the pair of brackets occuring 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.
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). In a recent work [17] the author presents a unified analysis of the so-called Horton-Strahler parameters of binary

tree structures. In this paper it is proven, that 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 $\mathbf{R}_p(x,u,v) := \sum_{t\in\mathcal{T}_p} x^{\mathbf{b}(t)}u^{\mathbf{u}(t)}v^{\mathbf{l}(t)}$ for $\mathcal{T}$ the set of all extended binary trees and $\mathcal{T}_p$ the set of those $t \in \mathcal{T}$ which have a Horton-Strahler number of p possess the following representations:

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

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

for $U_n(z)$ the n-th Chebyshev polynomial of the second kind (see e.g. [1]) and $\omega := \left(1-\sqrt{1-4\,\frac{xv}{(1-2\,u)^2}}\right)\left(1+\sqrt{1-4\,\frac{xv}{(1-2\,u)^2}}\right)^{-1}$. Since these generating functions keep track of the different types of internal nodes it becomes possible to translate the above transformations, which make a binary tree (a semi Dyck-word) to become a model for a secondary structure, into substitutions for the variables. For example, consider the leftmost case of the transformations in which a single node has to be expanded in such a way that it becomes a node with three lists attached to it, one of these lists must consist of at least one node. If we let z mark an opening or closing bracket within a secondary structure and if we let a mark a symbol | then this transformation can be considered by setting

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

since a single node in the class of binary trees corresponds to a node with two successors that are leaves in the class of extended binary trees and thus variable v is the one which marks the appropriate node within the generating functions $\mathbf{T}$ and $\mathbf{R}_p$. In the same way, the second and the third transformation correspond to the substitution

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

The rightmost transformation is considered via

$$x := z^2\frac{1}{1-a}.$$

Therefore, the ordinary generating function of all secondary structures where a paired (resp. unpaired) base is marked by z (resp. a) becomes

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

In [23] it was shown, that the number of words of size $2n$ in $\alpha(\mathcal{S}_p)$, for $\mathcal{S}_p$ the set of all secondary structures of order p, equals the number of binary trees with

n internal nodes and a Horton-Strahler number of p. Thus our substitutions applied to $R_p(x,u,v)$ lead to a representation for the generating function of all secondary structures of order p. We find

$$R_p(z,a) := \frac{\sqrt{a}}{a-1} U_{2^p-1}^{-1}\left(\frac{-(a-1)^2+(1+a)z^2}{2z^2\sqrt{a}}\right) \quad (1)$$

$$= \frac{\sqrt{a}\,(1-\omega)\,\omega^{2^{p-1}}}{(1-a)\,\sqrt{\omega}\,(1-\omega^{2^p})}, \quad (2)$$

for $\omega := (1-\varepsilon)\,(1+\varepsilon)^{-1}$, $\varepsilon := \sqrt{1-4\frac{az^4}{((a-1)^2-(1+a)z^2)^2}}$.
Please note, that the generating functions $\mathbf{T}(x,u,v)$ and $\mathbf{R}_p(x,u,v)$ do not count the empty tree. Thus, after substituting the variables, the completely linear secondary structure is not considered.
Remark: The correspondence given above does not work in a way which translates each binary tree with Horton-Strahler number p into a secondary structure of order p and vice versa. However, each binary tree with Horton-Strahler number p 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 p. Therefore the correspondence can be used for all enumeration purposes considered in this paper.

In the subsequent sections we will use these generating functions in order to derive numerous results on combinatorial properties of secondary structures.

3 The Number of Structures of Order p

In [24] Waterman raised the question of determining the number $a_{n,p}$ of secondary structures of size n and order p. Assuming that the prediction of the secondary structure is performed as described in the introduction this is equivalent to determine the size of the configuration space over which the search for the optimal secondary structure is to be performed for the corresponding iteration. Several attempts were made to solve this problem (see e.g. [23],[10]) but so far, no satisfactory solution has been found. In this section we will give precise asymptotic estimates for $a_{n,p}$ but also for the number of secondary structures of order p with n paired and m unpaired bases denoted by $a_{m,n,p}$.
In order to determine an asymptotic for $a_{n,p}$ we start with the representation of $R_p(z,a)$ as given in (1) and set $a := z$, i.e. we do not distinguish between paired and unpaired bases, each base is marked by z. We find

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

For the computation of an asymptotic for the coefficient at z^n we are interested in the dominant singularity of the generating function. This singularity results

from the zeros of the Chebyshev polynomial for which it is known that

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

Thus we have to find the smallest solution of

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

This is not a trivial task, since it is equivalent to determine the roots of a polynomial of sixth degree which can be solved in terms of hypergeometric functions in one variable using Klein's approach [12] to solving the quintic equation. We will first continue our computations by assuming that there is only one dominant singularity which is given by $z(p)$ without knowing its exact value and the related choice for m. Then the next step is to find an expansion of $R_p(z,z)$ around $z(p)$. This can be done by using the representation [1, 22.3.16]

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

We find

$$\lim_{z\to z(p)} \frac{\left(1-\frac{z}{z(p)}\right)}{\left(\frac{z^3+2z-1}{2z^{5/2}} - \cos\left(m2^{-p}\pi\right)\right)} = \frac{-4z(p)^{5/2}}{z(p)^3-6z(p)+5} \tag{3}$$

and

$$\lim_{x\to\cos(m2^{-p}\pi)} \left(x-\cos\left(m2^{-p}\pi\right)\right) \frac{\sin(\arccos(x))}{\sin(2^p\arccos(x))} = -\frac{2^{-p}\sin^2(m2^{-p}\pi)}{\cos(m\pi)}.$$

In that way we get the expansion

$$R_p(z,z) = \frac{1}{\left(1-\frac{z}{z(p)}\right)} \frac{-4z(p)^3}{(1-z(p))(z(p)^3-6z(p)+5)} \frac{2^{-p}\sin^2(m2^{-p}\pi)}{\cos(m\pi)}.$$

By means of the $\mathcal{O}$-transfer method [6] this expansion can be translated into an asymptotic for the coefficient. We find

Lemma 1 *Assuming that $z(p)$ is the only dominant singularity of the generating function $R_p(z,z)$, then there exists $m \in [1:2^p-1]$ such that the number $a_{n,p}$ of secondary structures of size n and order p is asymptotically given by $c(m,p)z(p)^{-n}$, $n\to\infty$, where*

$$c(m,p) := \frac{-4z(p)^3}{(1-z(p))(z(p)^3-6z(p)+5)} \frac{2^{-p}\sin^2(m2^{-p}\pi)}{\cos(m\pi)}.$$

It remains to find a representation of $z(p)$, the correct choice for m and to prove that there is only one singularity on the circle of convergence. For that purpose we discuss the equation $f(z) = \cos(m2^{-p}\pi)$ for $f(z) := \frac{z^3+2z-1}{2z^{5/2}}$ and $1 \leq m \leq 2^p - 1$. We find that $f(z) = -1$ for $z = \frac{3}{2} - \frac{1}{2}\sqrt{5}$ and that the smallest solution of $f(z) = 1$ is given by $z = 1$. Furthermore, since $f'(z) \geq 0$ for $z \in [\frac{3}{2} - \frac{1}{2}\sqrt{5}, 1]$, we know that f is a monotone increasing function within this interval. Thus, the smallest solutions of $f(z) = \cos(m2^{-p}\pi)$ result from the choice $m = 2^p - 1$ since this minimizes the value of the cosine. So $z(p)$ is the smallest real solution of $f(z) = \cos((2^p - 1)2^{-p}\pi) = -\cos(2^{-p}\pi)$. Implied by properties of the cosine we find that the sequence $f(z(p))$ is monotone decreasing with respect to p. Furthermore, we can argue that $z(p)$ is a monotone decreasing sequence itself by means of the positivity of the first derivative f' for all values in the interval $[z(\infty), z(1)] = [\frac{3}{2} - \frac{1}{2}\sqrt{5}, \frac{(108+12\sqrt{177})^{2/3}-24}{6(108+12\sqrt{177})^{1/3}}]$. Since the value of $f''(x)$ is negative for all x in that interval, we can conclude that $\frac{f(z(p))-f(z(\infty))}{z(p)-z(\infty)} \geq f'(z(p))$ for all possible p. Now setting $f(z(p)) = -\cos(2^{-p}\pi)$ and expanding the cosine finally proves

$$0 \leq z(p) - z(\infty) \leq 4^{-p}.$$

It remains to show that there are no additional singularities on the circle of convergence. For $p = 1$ the equation $f(z) = -\cos(2^{-p}\pi)$ can be solved explicitly. We get the solution $z(1)$ as given above and two complex roots of larger modulus. For $p > 1$ we consider the equation $z^3 + 2z - 1 + 2\cos(2^{-p}\pi)z^{5/2} = 0$. We define $g(z) := 2\cos(2^{-p}\pi)z^{5/2} - 3z - 1$ and $h(z) := z^3 + 5z$ and regard the contour $|z| = \frac{1}{2}$. For all $p \in \mathbb{N}$, $p > 1$, we find that $|h(\frac{1}{2}e^{i\phi})| > |g(\frac{1}{2}e^{i\phi})|$, $\phi \in [0, 2\pi]$, such that the theorem of Rouché tells us that $g(z) + h(z)$ and $h(z)$ possess the same number of zeros within the domain $|z| < \frac{1}{2}$. Therefore, we have:

Theorem 3 *The number $a_{n,p}$ of secondary structures of size n and order p is asymptotically given by $c(2^p - 1, p)z(p)^{-n}$, $n \to \infty$, where $z(p)$ is the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}} = -\cos(2^{-p}\pi)$. For all possible p, $0 \leq z(p) - \left(\frac{3}{2} - \frac{1}{2}\sqrt{5}\right) \leq 4^{-p}$ holds.*

In Figure 2 some approximate values for $\frac{1}{z(p)}$ and $c(2^p - 1, p)$ can be found. Figure 3 shows the quotient of the number of secondary structures of order p and the total number of structures for different sizes n. In Figure 4 you can find the quotient of some exact values of $a_{n,p}$ and their approximation as given in Theorem 3. As you can see, the rate of convergence of the asymptotic gets worse the larger p becomes.

Now we will consider the bivariate case, i.e. we will determine an asymptotic for the number $a_{m,n,p}$ of secondary structure of order p with m unpaired and n paired bases. For that purpose we return to (1) and determine the dominant

p	$\frac{1}{z(p)}$	$c(2^p-1,p)$
1	2.20557	0.143725
2	2.51326	0.0195502
3	2.59173	0.00249731
4	2.61145	0.000313867
5	2.61639	0.0000392868
6	2.61762	0.00000491253
7	2.61793	0.000000614118
8	2.61801	0.0000000767664
$\vdots$	$\vdots$	$\vdots$
∞	$\frac{3+\sqrt{5}}{2} \approx 2.61803$	$\to \mathcal{O}(2^{-p}) = 0$

Figure 2: Some numerical values for $\frac{1}{z(p)}$ and $c(2^p-1,p)$.

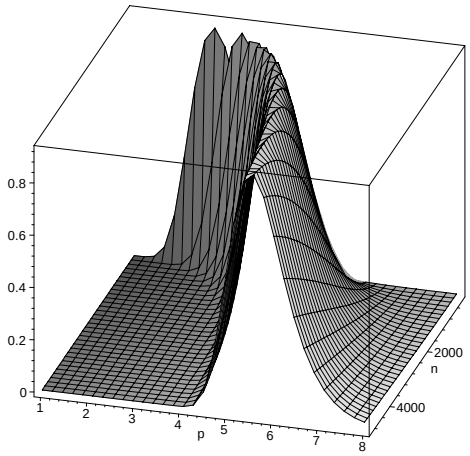


Figure 3: The quota of the secondary structures of order p to all structures.

	p				
n	1	2	3	4	5
10	.99870...	.15770...	0	0	0
50	.99999...	.99463...	.28415...	$< 10^{-5}$	0
100	.99999...	.99999...	.82109...	.00855...	$< 10^{-5}$
250	.99999...	.99999...	.99820...	.45636...	.00006...
1000	.99999...	.99999...	1.00000...	.99797...	.45247...
1500	.99999...	.99999...	1.00000...	.99995...	.76964...

Figure 4: The quotient of some exact values of $a_{n,p}$ and their approximation as given in Theorem 3. An entry of zero indicates that there exists no secondary structure of the given size and order.

singularities now being the solutions of

$$\frac{-(a-1)^2+(1+a)z^2}{2z^2\sqrt{a}} = \cos(m2^{-p}\pi),\ 1 \le m \le 2^p-1.$$

We find two solutions located on the circle of convergence, namely

$$z(a,p) := \frac{1-a}{\sqrt{1+a+2\sqrt{a}\cos(2^{-p}\pi)}},\ \bar{z}(a,p) := -z(a,p).$$

Again we have to find an expansion of $R_p(z,a)$ at $z(a,p)$, the second singularity will only be responsible for a factor $(1+(-1)^n)$ within the asymptotic. We thus

will not consider it in detail. Using the same ideas as in the univariate case we find

$$R_p(z,a) = \frac{1}{\left(1-\frac{z}{z(a,p)}\right)} \frac{a2^{-p}\sin^2(2^{-p}\pi)}{(1-a)(1+a+2\sqrt{a}\cos(2^{-p}\pi))}. \tag{4}$$

Now, we shall apply the following theorem due to Bender [2] in order to derive an asymptotic for the coefficient at $z^n a^m$:

Theorem 4 *Let $f(z,w) := \sum_{n,k\geq 0} a_{n,k} z^n w^k$, $a_{n,k} \geq 0$, and let $-\infty < a < b < \infty$. Define $R(\epsilon) := \{z \mid a \leq \Re(z) \leq b \wedge |\Im(z)| \leq \epsilon\}$. Suppose there exist $\epsilon > 0$, $\delta > 0$, an integer $m \geq 0$, and functions $A(s)$ and $r(s)$ such that*

(i) $A(s)$ is continuous and $A(s) \neq 0$ for $s \in R(\epsilon)$;

(ii) $r(s) \neq 0$ and has bounded third derivative for $s \in R(\epsilon)$;

(iii) for $s \in R(\epsilon)$ and $|z| \leq |r(s)|(1+\delta)$, the function

$$\left(1-\frac{z}{r(s)}\right)^m f(z,e^s) - A(s)\left(1-\frac{z}{r(s)}\right)^{-1}$$

is analytic and bounded;

(iv) $\left(\frac{r'(\alpha)}{r(\alpha)}\right)^2 \neq \left(\frac{r''(\alpha)}{r(\alpha)}\right)$ for $a \leq \alpha \leq b$;

(v) $f(z,e^s)$ is analytic and bounded for $|z| \leq |r(\Re(s))|(1+\delta)$ and $\epsilon \leq |\Im(s)| \leq \pi$.

Then, we have

$$a_{n,k} \sim \frac{n^m e^{-\alpha k} A(\alpha)}{m! r(\alpha)\sigma_\alpha \sqrt{2\pi n}}$$

uniformly for $a \leq \alpha \leq b$, where $\frac{k}{n} = -\frac{r'(\alpha)}{r(\alpha)}$ and $\sigma_\alpha^2 = \left(\frac{k}{n}\right)^2 - \frac{r''(\alpha)}{r(\alpha)}$.

The expansion (4) proves that the assumptions of the theorem are fulfilled. In detail, we find that in our case $m := 0$ and $A(s) := \frac{e^s 2^{-p}\sin^2(2^{-p}\pi)}{(1+e^s+2\sqrt{e^s}\cos(2^{-p}\pi))(1-e^s)}$ together with $r(s) := \frac{1-e^s}{\sqrt{1+e^s+2\sqrt{e^s}\cos(2^{-p}\pi)}}$ is the correct choice. We conclude that

$$-\frac{r'(s)}{r(s)} = -\frac{e^s(3+e^s)+\cos(2^{-p}\pi)\sqrt{e^s}(1+3e^s)}{2(e^s-1)(1+e^s+2\sqrt{e^s}\cos(2^{-p}\pi))}$$

and

$$\sigma_\alpha^2 = \frac{6+6\cosh(\alpha)+4\cos(2^{1-p}\pi)+(15\cosh(\alpha/2)+\cosh(3\alpha/2))\cos(2^{-p}\pi)}{32\,\mathrm{csch}^{-2}(\alpha/2)(\cosh(\alpha/2)+\cos(2^{-p}\pi))^2}$$

hold. Thus the application of the theorem together with the contribution of the second singularity yield:

	p					
ρ	1	2	3	4	5	6
$\frac{1}{10}$	-2.733709	-3.361703	-3.527500	-3.569480	-3.580007	-3.582640
$\frac{1}{5}$	-2.072009	-2.399273	-2.479055	-2.498913	-2.503873	-2.505112
$\frac{1}{3}$	-1.609437	-1.782359	-1.820340	-1.829540	-1.831822	-1.832391
$\frac{1}{2}$	-1.270196	-1.362089	-1.380500	-1.384861	-1.385937	-1.386205
1	-0.7821377	-0.8055456	-0.8096698	-0.8106201	-0.8108530	-0.8109109
2	-0.4409915	-0.4453515	-0.4460703	-0.4462338	-0.4462738	-0.4462837
3	-0.3065120	-0.3079899	-0.3082293	-0.3082837	-0.3082969	-0.3083002
4	-0.2347602	-0.2354266	-0.2355338	-0.2355581	-0.2355641	-0.2355655
5	-0.1901913	-0.1905463	-0.1906032	-0.1906161	-0.1906193	-0.1906200

Figure 5: The appropriate values of α for the asymptotic of $a_{m,n,p}$ with $\rho := \frac{m}{n}$.

Theorem 5 *The number $a_{m,n,p}$ of secondary structures of order p with m unpaired and n paired bases is asymptotically given by*

$$(1+(-1)^n)\frac{4(e^\alpha-1)^{-(n+1)}}{2^p\sqrt{\pi n}\,\mathrm{csch}(\alpha/2)}$$

$$\times\frac{(1+e^\alpha+2\sqrt{e^\alpha}\cos(2^{-p}\pi))^{n/2-1}e^{\alpha-\alpha m}\sin^2(2^{-p}\pi)(\cosh(\alpha/2)+\cos(2^{-p}\pi))}{\sqrt{6+6\cosh(\alpha)+4\cos(2^{-p+1}\pi)+(15\cosh(\alpha/2)+\cosh(3\alpha/2))\cos(2^{-p}\pi)}},$$

uniformly for $\alpha\in]-\infty,0[$, where $\frac{m}{n}=-\frac{e^\alpha(3+e^\alpha)+\cos(2^{-p}\pi)\sqrt{e^\alpha}(1+3e^\alpha)}{2(e^\alpha-1)(1+e^\alpha+2\sqrt{e^\alpha}\cos(2^{-p}\pi))}$.

Please note that the choice $\alpha\in]-\infty,0[$ does not imply any restriction on the application of the asymptotic; all reasonable ratios of m and n lead to a solution of $\frac{r'(\alpha)}{r(\alpha)}$ within this interval. In Figure 5 you find the appropriate values of α for different choices of p and $\rho := \frac{m}{n}$.
Figure 6 shows the quotient of some exact values of $a_{m,n,p}$ and their asymptotical equivalents as given in Theorem 5. For the computation of the approximations we used the values of α as given in Figure 5. We will now turn to another open problem within this context, namely the computation of the average order of a secondary structure and the determination of the related higher moments.

4 The Expected Order of a Secondary Structure and the Related Higher Moments

In this section we will prove exact asymptotic equivalents for the expected order of a secondary structure of size n, for the related variance and the r-th moments

		p			
ρ	n	1	2	3	4
$\frac{1}{10}$	50	.98793	.97009	.21429	0
	100	.99391	.98800	.77609	.00352
	300	.99791	.99600	.99452	.55478
$\frac{1}{3}$	60	.99899	1.0004	.79363	.00535
	90	.99933	1.0002	.95110	.07571
	120	.99948	1.0002	.98875	.22974
1	10	.99833	.73537	0	0
	20	.99909	.97763	.09833	0
	60	.99966	.99963	.87649	.01770
3	10	.99317	.75580	0	0
	20	.99655	.97949	.11230	0
	40	.99830	.99808	.65284	.00041

Figure 6: The quotient of some exact values of $a_{m,n,p}$ and their approximation as given in Theorem 5. An entry of 0 indicates that $a_{m,n,p}$ is zero.

about the origin. Furthermore, we will provide asymptotics for the r-th moments about the origin for the order of a secondary structure with n bases of which m are not paired.

4.1 Expected Order

In order to determine the expected order of a secondary structure of size n we have to consider the sum

$$M(z) := \sum_{p\geq 1} pR_p(z,z).$$

It proves convenient to use the representation of R_p given in (2) in order to evaluate this sum. By obvious series expansion we find

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

The sum $\sigma(\omega)$ can be evaluated by means of the Mellin summation formula [5]. For that purpose we compute the Mellin transform of $\sigma(e^{-t})$ which proves to be given by

$$\Gamma(s)\sum_{\substack{p\geq 1\\ j\geq 0}} p(2^{p-1}(1+2j))^{-s} = \frac{2^s}{2^s-1}\zeta(s)\Gamma(s) =: \mathcal{M}(s)$$

for $\Gamma(s)$ the complete gamma function and $\zeta(s)$ Riemann's zeta function. According to the method, we get an expansion of $\sigma(e^{-t})$ at $t=0$ by summing the residues of $t^{-s}\mathcal{M}(s)$ located left to the fundamental strip of $\mathcal{M}(s)$. The corresponding singularities are the poles at $s=1$, $s=0$ (double pole), $s=-k$ for $k\in\mathbb{N}$ and $s=\chi_k:=\frac{2\pi ik}{\ln(2)}$ for $k\in\mathbb{Z}\backslash\{0\}$. We find for the corresponding 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 $\mathcal{O}(t^{2n+1})$. Besides the sum of these residues we need an expansion of the factor $\varphi:=\frac{\sqrt{z}}{1-z}\frac{1-\omega}{\sqrt{\omega}}$ around the dominant singularity of $M(z)$. This singularity is located at $z_d=\frac{3}{2}-\frac{1}{2}\sqrt{5}$ which can be concluded from the obvious bounds

$$[z^n]T(z,z)\le[z^n]M(z)\le[z^n]\log_2(n)T(z,z)$$

and the fact that $z=\frac{3}{2}-\frac{1}{2}\sqrt{5}$ is the dominant singularity of $T(z,z)$. Furthermore, z_d is the only singularity on the circle of convergence of $M(z)$. The expansion of φ at z_d is given by

$$4\frac{\sqrt{9\sqrt{5}-20}}{5\sqrt{5}-11}\sqrt{1-\frac{z}{z_d}}+\mathcal{O}\left(\left(1-\frac{z}{z_d}\right)^{3/2}\right).$$

Since $t=-\ln\left(\frac{1-\varepsilon}{1+\varepsilon}\right)$ and $\varepsilon=0$ for $z=z_d$ we expand $-\ln\left(\frac{1-\varepsilon}{1+\varepsilon}\right)$ around $\varepsilon=0$ to find $t=2\varepsilon+\mathcal{O}(\varepsilon^3)$ and thus

$$t\sim 4\underbrace{\frac{\sqrt{9\sqrt{5}-20}}{5\sqrt{5}-11}}_{=:\rho}\sqrt{1-\frac{z}{z_d}}. \tag{5}$$

Thus we can conclude that $\varphi=t+\mathcal{O}(t^3)$ holds. If we now multiply the sum of residues by this representation of φ we get

$$2+\left(\frac{2\ln(t)+2\gamma-2\ln(\pi)-3\ln(2)}{4\ln(2)}\right)t+\sum_{k\neq 0}\frac{\zeta(\chi_k)\Gamma(\chi_k)}{\ln(2)}t^{1-\chi_k}+\mathcal{O}(t^2).$$

This expansion can be translated into an expansion around the dominant singularity z_d by substituting t corresponding to (5). In that way we find:

$$\begin{aligned} M(z) &= -\frac{\rho\sqrt{1-\frac{z}{z_d}}\ln\left(\frac{1}{1-\frac{z}{z_d}}\right)}{\ln(2)} + \frac{\rho(\ln(2)+2\ln(\rho)+2\gamma-2\ln(\pi))}{\ln(2)}\sqrt{1-\frac{z}{z_d}} \\ &\quad + \frac{4\rho}{\ln(2)}\sum_{k\neq 0}\Gamma(\chi_k)\zeta(\chi_k)e^{-2\pi ik\log_2(\rho)}\left(1-\frac{z}{z_d}\right)^{\frac{1-\chi_k}{2}} + \mathcal{O}\left(1-\frac{z}{z_d}\right). \end{aligned}$$

Now we can use the $\mathcal{O}$-transfer method in order to get the asymptotic for the coefficient $[z^n]M(z)$. The application of the appropriate formulæ yields

$$\begin{aligned} [z^n]M(z) \sim\ & \frac{\rho}{\ln(2)\sqrt{\pi n^3}}z_d^{-n}\left(\frac{1}{2}\ln(n)+\frac{\gamma+2\ln(2)-2}{2}\right) \\ & -\frac{\rho z_d^{-n}(\ln(2)+2\ln(\rho)+2\gamma-2\ln(\pi))}{2\ln(2)\sqrt{\pi n^3}} \\ & +\frac{4\rho z_d^{-n}}{\ln(2)}\sum_{k\neq 0}\Gamma(\chi_k)\zeta(\chi_k)e^{-2\pi ik\log_2(\rho)}n^{\frac{\chi_k-3}{2}}/\Gamma\left(\frac{\chi_k-1}{2}\right)+\mathcal{O}\left(z_d^{-n}n^{-5/2}\right). \end{aligned}$$

In order to determine the expected value we have to devide this asymptotic by the asymptotical number of secondary structures of size n. This number can be concluded from the following expansion of $T(z,z)$ at $z=z_d$ (terms relevant for the asymptotic only):

$$\begin{aligned} &T(z,z) = \\ &-\frac{\sqrt{30+14\sqrt{5}}}{2}\left(1-\frac{z}{z_d}\right)^{1/2} - \frac{\sqrt{170150+76470\sqrt{5}}}{80}\left(1-\frac{z}{z_d}\right)^{3/2}+\mathcal{O}\left(\left(1-\frac{z}{z_d}\right)^2\right). \end{aligned}$$

We find

$$\begin{aligned} &[z^n]T(z,z) \sim \\ &\frac{z_d^{-n}}{\sqrt{\pi n^3}}\left(\frac{\sqrt{30+14\sqrt{5}}}{2}\left(\frac{1}{2}+\frac{3}{16n}\right)-\frac{3\sqrt{170150+76470\sqrt{5}}}{320n}\right)+\mathcal{O}\left(n^{-7/2}z_d^{-n}\right). \end{aligned}$$

The resulting quotient can by simplified so that we finally find:

Theorem 6 *The expected order of a secondary structure of size n is asymptotically given by*

$$\frac{1}{2}\log_2\left(\frac{2\pi^2}{\rho^2}n\right)-\frac{\gamma+2}{2\ln(2)}+\Delta\left(\log_2\left(\frac{n}{\rho^2}\right)\right)+\mathcal{O}(n^{-1}),$$

$n\to\infty$. Here, $\Delta(x)$ is a periodic function of very small modulus ($|\Delta(x)|\leq 0.040597\ldots$); is has the following Fourier series:

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

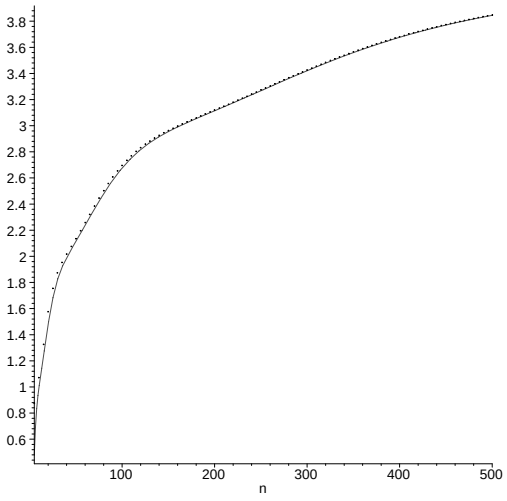


Figure 7: A plot of the exact expected order (dotted line) and its asymptotic equivalent (solid line).

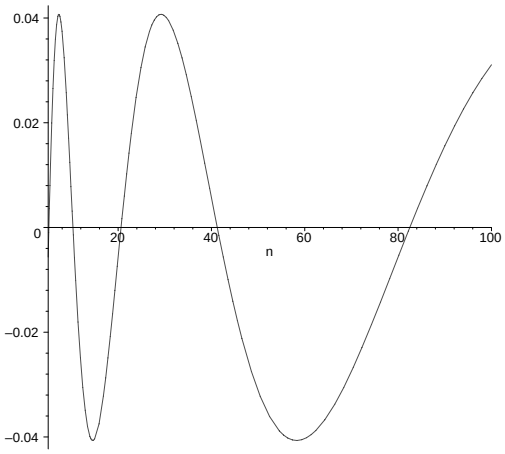


Figure 8: A plot of $\Delta\left(\log_2\left(\frac{n}{\rho^2}\right)\right)$.

Please note that we only had to use the leading term of the asymptotic for $[z^n]T(z,z)$ in order to compute this expectation. However, we will need all terms in order to derive further results which will be presented in subsequent sections. In Figure 7 you find a plot of the exact expected value together with its asymptotical equivalent as given in the previous theorem. Figure 8 shows the oscillation implied by the function Δ within the asymptotic behaviour.

We will continue our investigations by computing asymptotics for the variance and the higher moments.

4.2 Variance and Higher Moments

In this section we will provide an asymptotic equivalent for the variance of the expected order of a secondary structure. We will also prove asymptotic equivalents for the corresponding r-th moments. The latter will be done for both, secondary structures built from n bases and secondary structures built from n bases from which m are unpaired.

For the determination of the r-th moment we have to consider

$$M^{(r)}(z,a) := \sum_{p\geq 1} p^r R_p(z,a) = \frac{\sqrt{a}}{1-a}\frac{1-\omega}{\sqrt{\omega}} \underbrace{\sum_{\substack{p\geq 1\\ j\geq 0}} p^r \omega^{2^{p-1}(1+2j)}}_{=:\sigma^{(r)}(\omega)}.$$

Again, we use the Mellin summation technique to find an expansion of $\sigma^{(r)}(\omega)$ around the dominant singularity. The Mellin transform of $\sigma^{(r)}(e^{-t})$ is given by

$$\mathcal{M}^{(r)}(s) := \frac{\Gamma(s)2^s A_r(2^{-s})\zeta(s)}{(1-2^{-s})^r}.$$

Here, $A_n(x)$ denotes the n-th Eulerian polynomial for which [3, p. 245]

$$\sum_{l\geq 0} l^n u^l = \frac{A_n(u)}{(1-u)^{n+1}}$$

holds. It is the pole at $s=0$ of order $r+1$ which is responsible for the most significant contribution to the asymptotic of the coefficient. Therefore we are interested in the residue of $t^{-s}\mathcal{M}^{(r)}(s)$ at $s=0$. Since for $s \to 0$

$$A_r(1) = r!, \text{ see } [13, 5.1.3(4)],$$

$$\zeta(0) = -\frac{1}{2},$$

$$\Gamma(s) = s^{-1} - \gamma + \mathcal{O}(s),$$

$$\frac{1}{(1-2^{-s})^r} = \frac{1}{\ln^r(2)} s^{-r} + \mathcal{O}(s^{-r+1}), \text{ and}$$

$$t^{-s} = \sum_{i\geq 0} \ln^i(t)\frac{(-1)^i}{i!}s^i,$$

we can conclude that the residue of $t^{-s}\mathcal{M}^{(r)}(s)$ at $s=0$ is given by

$$(-1)^{r+1}\frac{\ln^r(t)}{2\ln^r(2)} + \mathcal{O}(\ln^{r-1}(t)). \tag{6}$$

Now we have to consider the factor $\frac{\sqrt{a}}{1-a}\frac{1-\omega}{\sqrt{\omega}}$. First we set $a := z$, i.e. we do not distinguish between paired and unpaired bases. In this case, the only dominant singularity is again given by z_d and we thus have to multiply (6) by t in order to take care of the factor. After resubstituting t within the resulting expansion of $M^{(r)}(e^{-t}, e^{-t})$ around $t=0$ we find the following leading term for the expansion of our generating function at its dominant singularity z_d:

$$-2^{-r+1}\log_2^r\left(\frac{1}{1-\frac{z}{z_d}}\right)\rho\left(1-\frac{z}{z_d}\right)^{1/2}.$$

The application of the $\mathcal{O}$-transfer method yields:

$$[z^n]M^{(r)}(z,z) \sim 2^{-r}\rho\frac{z_d^{-n}\log_2^r(n)}{\sqrt{\pi n^3}} + \mathcal{O}\left(n^{-3/2}\ln^{r-1}(n)z_d^{-n}\right).$$

Thus, by dividing this asymptotic by the asymptotical number of secondary structures of size n we have proven the following theorem:

Theorem 7 *The r-th moment of the order of a random secondary structure of size n is asymptotically given by*

$$2^{-r}\log_2^r(n)+\mathcal{O}(\ln^{r-1}(n)),\ n\to\infty.$$

Please note, that this asymptotic is not as precise as it ought to be in order to compute the variance since the leading term cancels out when we compute the 2nd moment minus the square of the first moment. Thus, in order to find a representation for the variance, we have to compute further terms for the second moment. We therefore return to $\mathcal{M}^{(2)}(z,z)$ and determine the exact residue at $s=0$ as well as the residues at $s=\chi_k$, $k\in\mathbb{Z}\backslash\{0\}$ (we do not consider $s=1$, since the resulting residue is of order t^{-1} and thus becomes constant after it has been multiplied with the expansion of the factor $\frac{\sqrt{z}}{1-z}\frac{1-\omega}{\sqrt{\omega}}=t+\mathcal{O}(t^3)$). For $\gamma(n):=\lim_{m\to\infty}\left(\sum_{k=1}^m\ln(k)^n/k-\ln(m)^{n+1}/(n+1)\right)$, the residue at $s=0$ is given by

$$\underbrace{\frac{-3\pi^2+4\left(6\gamma(1)-7\ln(2)^2+3\gamma\ln(8)+6\gamma\ln(\pi)-3\ln(\pi)\ln(8\pi)\right)}{24\ln(2)^2}}_{=:f_1}$$

$$-\ln(t)\underbrace{\frac{2\gamma-3\ln(2)-2\ln(\pi)}{2\ln^2(2)}}_{=:f_2}-\ln^2(t)\underbrace{\frac{1}{2\ln^2(2)}}_{=:f_3}.$$

For $\Psi:=\frac{d}{dx}\ln\left(\Gamma(x)\right)$, the residue at $s=\chi_k$ possesses the representation

$$-2\frac{\zeta(\chi_k)\Gamma(\chi_k)\ln(t)}{t^{\chi_k}\ln^2(2)}$$

$$+\frac{\Gamma(\chi_k)((\ln(2)+2\Psi(\chi_k))\zeta(\chi_k)+2\zeta'(\chi_k))}{t^{\chi_k}\ln^2(2)}.$$

After multiplying with t and resubstituting we find that the residue at $s=0$ implies the following contribution to the expansion of $M^{(2)}(z,z)$ at $z=z_d$:

$$4\left(1-\frac{z}{z_d}\right)^{1/2}\rho(f_1-\ln(4\rho)f_2-\ln^2(4\rho)f_3)-\left(1-\frac{z}{z_d}\right)^{1/2}\ln\left(1-\frac{z}{z_d}\right)\rho(2f_2+4\ln(4\rho)f_3)$$

$$-\left(1-\frac{z}{z_d}\right)^{1/2}\ln^2\left(1-\frac{z}{z_d}\right)\rho f_3.$$

This translates into the following contribution to the asymptotic for the coefficient $[z^n]M^{(2)}(z,z)$

$$\frac{z_d^{-n}}{2\sqrt{\pi n^3}}\rho f_3\ln^2(n)-\frac{z_d^{-n}}{\sqrt{\pi n^3}}\rho\ln(n)(f_2-f_3(\gamma-2+\ln(4))+2f_3\ln(4\rho))$$

$$-\frac{z_d^{-n}}{\sqrt{\pi n^3}}\rho\Bigg(2(f_1-\ln(4\rho)f_2-\ln^2(4\rho)f_3)+(f_2+2\ln(4\rho)f_3)(\gamma+2\ln(2)-2)$$
$$-f_3(\frac{1}{2}\gamma^2+2\gamma\ln(2)+2\ln^2(2)-\frac{1}{4}\pi^2-2\gamma-4\ln(2))\Bigg),$$

and into the contribution

$$\frac{1}{2}f_3\ln^2(n)-\ln(n)(f_2-f_3(\gamma-2+\ln(4))+2f_3\ln(4\rho))$$
$$-\Bigg(2(f_1-\ln(4\rho)f_2-\ln^2(4\rho)f_3)+(f_2+2\ln(4\rho)f_3)(\gamma+2\ln(2)-2)$$
$$-f_3(\frac{1}{2}\gamma^2+2\gamma\ln(2)+2\ln^2(2)-\frac{1}{4}\pi^2-2\gamma-4\ln(2))\Bigg),$$
$$=\frac{1}{4}\log_2^2(n)+\frac{1}{2}\log_2(n)\log_2\left(\frac{2\pi^2}{\rho^2}\right)-\frac{1}{2}\log_2(n)\frac{\gamma+2}{\ln(2)}+\mathcal{O}(1)$$

for the second moment by division through the asymptotical number of secondary structures of size n. For the singularities at χ_k we only consider the most significant part of the residue, i.e. the part which possesses the factor $\ln(t)$. Its contribution to the second moment is determined in the same way and proves to be given by

$$\frac{\ln(n)}{\ln^2(2)}\sum_{k\neq 0}(\chi_k-1)\Gamma\left(\frac{\chi_k}{2}\right)\zeta(\chi_k)e^{\pi ik\log_2\left(\frac{n}{\rho^2}\right)}.$$

Thus, not only the terms of order $\ln^2(n)$ cancel out when computing the variance, also the terms of order $\ln(n)$ possess the same coefficient (constant and oscillating part) within the second moment and the square of the first moment. Thus we conclude, that the variance is of order $\mathcal{O}(1)$. The computations for that constant term were only performed numerically without regarding the oscillations. In this way we have found:

Theorem 8 *The order of a random secondary structure of size n possesses a variance which is asymptotically given by*

$$0.17939\ldots+\bar{\Delta}(n),\ n\to\infty,$$

for $\bar{\Delta}(n)$ an oscillating function.

We have to expect that $\bar{\Delta}(n)$ is only of small modulus. This presumption is confirmed by the comparison of the exact variance with our asymptotical estimate which can be found in Figure 9. In Figure 10 you find a plot of the upper bound of the probability that the order of a random secondary structure differs at least by k from its mean implied by Chebyshevs inequality. We used the exact values of the variance in order to prepare the plot.

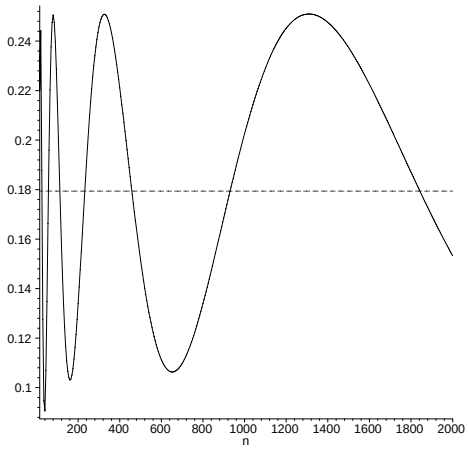


Figure 9: The comparison of the exact variance with the asymptotic estimate (dashed line).

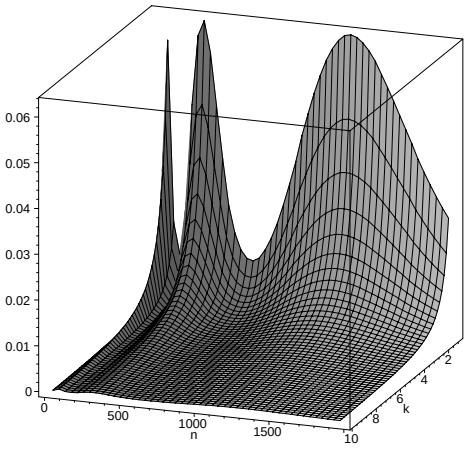


Figure 10: The upper bound for the probability of an order that differs at least by k from the mean.

Now, let us try to get an asymptotic for the r-th moment in the bivariate setting, i.e. for the secondary structures with n bases from which m are unpaired[1]. For this we have to set $a := az$ since in the initial setting our generating functions only mark the paired bases by z. It is obvious, that even after setting $a := az$ the substitution $\omega := e^{-t}$ implies exactly the same Mellin transform $\mathcal{M}^{(r)}(s)$ as in the univariate case. However, we can use $\mathcal{M}^{(r)}(s)$ only in a restricted way since the substitution $\omega := e^{-t}$ introduces a coupling of the actually independent variables a and z. This coupling is responsible for the fact that an asymptotic for the coefficient at $a^m z^n$ which results from this Mellin transform is only valid if m and n grow within a fixed proportion.
Let us start with the determination of the number of secondary structures with n bases from which m are unpaired. The appropriate generating function is $T(z, az)$ which possesses the dominant algebraic singularity $z_d(a) := 1 + \frac{1}{2}a - \frac{1}{2}\sqrt{a(4+a)}$ and the following leading terms

$$\frac{a(1+a) + (a-1)\sqrt{a(4+a)}}{4a-2} - 2\frac{\sqrt{a(4+a)}\sqrt{1-\frac{z}{z_d(a)}}}{\sqrt{-2a(2+a)(4+a) + 2\sqrt{a(4+a)}(2+a(4+a))}}$$

for its expansion at $z_d(a)$. To determine an asymptotic for its coefficients, we can use the Darboux-method for multivariate generating functions due to Drmota which is recalled in the following lemma:

[1] Initially, our generating functions suggest to consider the secondary structures with n paired (variable z) and m unpaired (variable a) bases. However, it proved to be technically convenient to change to this setting where one variable considers all bases (paired and unpaired).

Lemma 2 ([4]) *Let $c(x,z)$ be the generating function for $c_{n,k}$ and suppose that $c(x,z)$ has a positive radius of convergence $r(z)$ for $z = (z_1,\ldots,z_m) \in [a,b] = [a_1,b_1]\times\cdots\times[a_m,b_m]$, $(0 < a_i < b_i)$, $1 \le i \le m$. Furthermore, suppose that there is only one singularity $x = f(z)$ on the circle of convergence $|x| = r(z)$, and that $c(x,z)$ has an expansion of the following form:*

$$c(x,z) = u(x,z) + g(z)A\left(\frac{1}{1-x/f(z)}\right) + o\left(A\left(\frac{1}{1-x/f(z)}\right)\right)$$

for $x \to f(z)$, $z \in R(a,b,\phi) = \{z = (z_1,\ldots,z_m) : |z| \in [a,b], |\arg(z_i)| < \phi\}$ and $x \in T(f(z),\epsilon,\phi) = \{x : |f(z)-x| < \epsilon, |\arg(x-f(z))| < \phi\}$ for some ϵ, $\phi > 0$, where $A(u) = u^r L(u)$, $r \neq -1,-2,\ldots$, $L(u)$ is of slow variation, and $u(x,z)$, $f(z)$ and $g(z) \neq 0$ are analytic for $z \in R(a,b,\phi)$ and $|f(z)-x| < \epsilon$. Set

$$\mu(z) = -\nabla \log(f(e^s))|_{e^s=z},$$

suppose that the matrix

$$\Sigma(z) = \left(-\frac{\partial^2}{\partial s_i \partial s_j}\log(f(e^s))|_{e^s=z}\right)_{i,j=1,\ldots,m}$$

is regular for $z \in R(a,b,\phi)$, and that there exists a $\delta > 0$ such that $c(x,z)$ is analytic and bounded for $|z| \in [a,b]$, $z \notin R(a,b,\phi)$, $|x| \le r(z)(1+\delta)$. Then $det(\Sigma(z)) > 0$ and we have

$$c_{n,k} \sim \frac{g(h(\rho))}{\sqrt{(2\pi n)^m det(\Sigma(h(\rho)))}}\frac{A(n)}{\Gamma(r)n}\frac{1}{f(h(\rho))^n h(\rho)^k}$$

uniformly for $\rho := k/n \in [\mu(a),\mu(b)]$, where $h(t)$ is the inverse function of $\mu(z)$.

We need to be careful since the coefficients $[z^n a^m]T(z,az)$ are different from zero if and only if $n > m$, $n-m$ even. Therefore, the result which will be given by the application of Lemma 2 has to be multiplied by $(1+(-1)^{n-m})$ in order to get the correct asymptotic. Performing the appropriate computations we find that for $\varrho := \frac{m}{n}$

$$[z^n a^{n\varrho}]T(z,az) \sim (1+(-1)^{n-n\varrho})\frac{(n+n\varrho)^{n+\frac{1}{2}}(n^2-n^2\varrho^2)^{n\varrho-\frac{1}{2}}}{4^{n\varrho}(n-n\varrho)^{n+\frac{3}{2}}(n\varrho)^{2n\varrho}\pi}$$

uniformly for $\varrho \in]0,1[$, $n \to \infty$.
Now let us consider $M^{(r)}(z,az)$. As already mentioned, the resulting Mellin transform remains the same, we only have to resubstitute t in another way. Thus, we also can use the approximation for the residue at $t = 0$ as given in (6) to get the most significant term of the expansion around the dominant singularity. This

singularity is also given by $z_d(a)$ which can be seen by the same arguments as in the univariate setting. By means of series expansions we find that

$$t = -\ln(\omega) = \sqrt{24 + 6a + \frac{2}{a}\sqrt{a(4+a)}(2+5a)}\left(1 - \frac{z}{z_d(a)}\right)^{1/2} + \mathcal{O}\left(1 - \frac{z}{z_d(a)}\right).$$

The factor $\frac{\sqrt{az}}{1-az}\frac{1-\omega}{\sqrt{\omega}}$ expands to

$$\frac{\sqrt{2}\sqrt{a(2+a-\sqrt{a(4+a)})}}{2+a(-2-a+\sqrt{a(4+a)})}t.$$

If we now recombine the different parts of the generating function and resubstitute t, then we find for the most significant part of the expansion at $z_d(a)$:

$$-2^{-1/2-r}\sqrt{a(2+a)(4+a)+\sqrt{a(4+a)}(2+a(4+a))}\,\frac{\ln^r\left(\frac{1}{1-\frac{z}{z_d(a)}}\right)}{\ln^r(2)}\sqrt{1-\frac{z}{z_d(a)}}.$$

Again, we can use this expansion together with Lemma 2 to get an asymptotic for the coefficient. This asymptotic also has to by multiplied by $(1+(-1)^{n-m})$ for the same reasons as before. We find for $\varrho := \frac{m}{n} \in]0,1[$ fix:

$$[z^n a^{n\varrho}]M^{(r)}(z,az) \sim (1+(-1)^{n-n\varrho})\frac{\log_2^r(n)(n+n\varrho)^n(n^2-n^2\varrho^2)^{n\varrho}}{2^r 4^{n\varrho}\pi(n-n\varrho)^{n+2}(n\varrho)^{2n\varrho}},$$

for $n \to \infty$. By dividing this asymptotic by the asymptotical number of secondary structures of the appropriate size we get an asymptotic for the r-th moment. We find:

Theorem 9 *The r-th moment of the order of a random secondary structure with n bases of which m are unpaired is asymptotically given by*

$$2^{-r}\log_2^r(n) + \mathcal{O}(\ln^{r-1}(n)),\ \varrho := \frac{m}{n} \in]0,1[\ \textit{fix},\ n \to \infty.$$

As a consequence, the leading term of the r-th moment is identical with the one in the univariate case and therefore independent of the ratio ϱ. However, empirical observations imply the conjecture that the second order term depends on ϱ.

5 The Distribution of the Unpaired Bases

In this section we will investigate the number of unpaired bases in arbitrary secondary structures and in secondary structures of order p. We start with the simple task to determine the average number of unpaired bases and the corresponding

variance for secondary structures of size n. For that purpose we determine the first partial derivative of $T(z, az)$ with respect to a and set $a := 1$ afterwards. We find

$$\left[\frac{\partial}{\partial a}T(z,az)\right]_{a=1} = \frac{(z-1)^2(1+z(z-1))+(z(2-3z)-1)\sqrt{(1+z(z-3))(1+z+z^2)}}{2(z-1)^2 z\sqrt{(1+z(z-3))(1+z+z^2)}}.$$

Again, the dominant singularity stems from a zero of the square roots and is given by $z_d = \frac{3}{2} - \frac{1}{2}\sqrt{5}$. The expansion of $\left[\frac{\partial}{\partial a}T(z,az)\right]_{a=1}$ at z_d possesses the following leading terms:

$$\frac{\sqrt{150+70\sqrt{5}}}{20}\left(1-\frac{z}{z_d}\right)^{-\frac{1}{2}} - \frac{7\sqrt{5}-17}{6\sqrt{5}-14} + \frac{3\sqrt{1710+766\sqrt{5}}}{160}\left(1-\frac{z}{z_d}\right)^{\frac{1}{2}} + \mathcal{O}\left(1-\frac{z}{z_d}\right).$$

The $\mathcal{O}$-transfer method now implies the asymptotic for the coefficient at $[z^n]$ as given below:

$$[z^n]\left[\frac{\partial}{\partial a}T(z,az)\right]_{a=1} = \frac{z_d^{-n}}{\sqrt{\pi n}}\left(\left(\frac{1}{20}-\frac{1}{160n}\right)\sqrt{150+70\sqrt{5}} - \frac{3}{320n}\sqrt{1710+766\sqrt{5}}\right) + \mathcal{O}\left(z_d^{-n}n^{-5/2}\right).$$

By dividing this asymptotic by the asymptotical number of secondary structures of size n as given in Section 4 we find the following theorem:

Theorem 10 *The average number of unpaired bases in a secondary structure of size n is asymptotically given by*

$$\frac{n}{\sqrt{5}} + \frac{3}{10} + \frac{1}{\sqrt{5}} + \mathcal{O}(n^{-1}),\ n \to \infty.$$

Thus about 44% of the bases are unpaired on the average. A similar result concerning the expected number of unpaired bases can be found in [10]. Using the same methods we can also determine an asymptotic for the second factorial moment for the number of unpaired bases in a random secondary structure by considering the second partial derivative $\left[\frac{\partial^2}{\partial a^2}T(z,az)\right]_{a=1}$. We find, that the second factorial moment is asymptotically given by

$$\frac{1}{5}n^2 + \frac{2}{5}n$$

which implies for the variance σ^2:

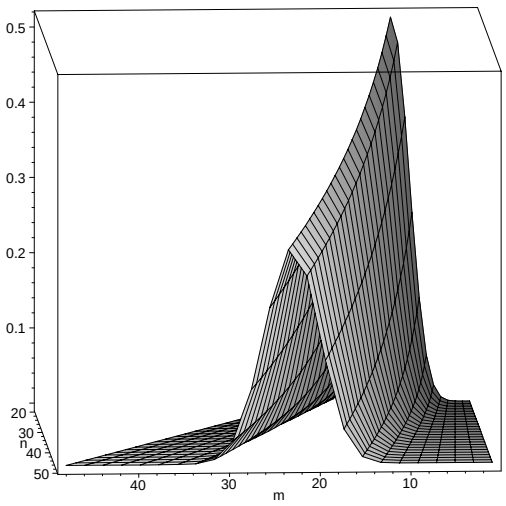


Figure 11: A plot of the distribution of the secondary structures of size n with m unpaired bases.

Theorem 11 *The variance σ^2 of the number of unpaired bases in a random secondary structure of size n is asymptotically given by*

$$\frac{2}{5}\frac{n+1}{\sqrt{5}} + \frac{1}{100},\ n \to \infty.$$

The results we have deduced so far enable us to get additional knowledge. In Section 4 we have computed asymptotics for the number of secondary structures with n bases from which m are unpaired and the number of secondary structures of size n. By dividing these asymptotics one by the other we find:

Theorem 12 *Asymptotically and uniformly for $\frac{m}{n} \in]0,1[$, $n \to \infty$,*

$$\frac{(1+(-1)^{n-m})5\,2^{6-2m-n}(3-\sqrt{5})^n n^{5/2}(n+m)^{n+\frac{1}{2}}(n^2-m^2)^{m-\frac{1}{2}}}{m^{2m}(n-m)^{n+\frac{3}{2}}\sqrt{\pi}(10\sqrt{30+14\sqrt{5}}(8n+3)-3\sqrt{170150+76470\sqrt{5}})} \times 100$$

percent of all secondary structures of size n have m unpaired bases.

If we take a look at Figure 11 which shows a plot[2] of this quantity we get the conjecture that the parameter possesses a normal distribution.

In order to prove this conjecture we use the following theorem:

Theorem 13 ([8]) *Let $F(z,u)$ be a bivariate function that is analytic in a domain*

$$\mathcal{D}_0 = \{(z,u) \mid |z| < \rho, |u| < 1\},$$

[2]Please note that we have replaced the factor $(1+(-1)^{n-m})$ of the asymptotic by 2 and omitted the factor 100 in order to generate this plot.

and has nonnegative coefficients at $(0,0)$. *Assume that there exists* $\epsilon > 0$, $\vartheta < \frac{\pi}{2}$, *and* $r > \rho$ *such that in the domain*

$$\mathcal{D} = \{(z,u) \mid |z| \le r,\ Arg(z-\rho) \in [\vartheta, 2\pi - \vartheta],\ |u-1| < \epsilon\},$$

the function $F(z,u)$ *admits the representation*

$$F(z,u) = A(z,u) + B(z,u)C(z,u)^{-\alpha}(\log C(z,u))^k,$$

where $A(z,u)$, $B(z,u)$, $C(z,u)$ *are analytic for* $(z,u) \in \mathcal{D}$, k *is a nonnegative integer, and* $\alpha \notin \{0,-1,-2,\ldots\}$. *Assume also that the equation*

$$C(\zeta, 1) = 0$$

has only one (simple) root $\zeta = \rho$ *in* $|z| \le r$ *and that* $B(\rho,1) \neq 0$. *Assume finally the "variability condition",*

$$0 < \liminf \frac{\sigma_n^2}{n}.$$

Then, the variable with probability generating function

$$p_n(u) = \frac{[z^n]F(z,u)}{[z^n]F(z,1)}$$

converges in distribution to a Gaussian variable with a speed of convergence that is $\mathcal{O}(n^{-1/2})$.

As postulated in this theorem we expand $T(z,az)$ around its dominant singularity $z(a) := 1 + \frac{1}{2}a - \frac{1}{2}\sqrt{a(a+4)}$ which yields

$$\underbrace{\frac{a(1+a) + (a-1)\sqrt{a(a+4)}}{4a-2}}_{=:A(z,a)} - \underbrace{2\frac{\sqrt{a(a+4)}}{\sqrt{-2a(2+a)(4+a) + 2\sqrt{a(4+a)}(2+a(a+4))}}}_{=:B(z,a)} \left(1 - \frac{z}{z(a)}\right)^{1/2}.$$

Thus we have to set $C(z,a) := 1 - \frac{z}{z(a)}$, $k = 0$ and $\alpha = -\frac{1}{2}$. The equation $C(\zeta,1) = 1 - \frac{\zeta}{\frac{3}{2} - \frac{1}{2}\sqrt{5}} = 0$ possesses exactly one solution, namely $\zeta = \frac{3}{2} - \frac{1}{2}\sqrt{5} =: \rho$. Furthermore, $B(\rho,1) = -\frac{2\sqrt{5}}{\sqrt{14\sqrt{5}-30}} \neq 0$. The variance σ_n^2 has already been computed in Theorem 11 and proves that the variability condition is fulfilled. Thus we can infer:

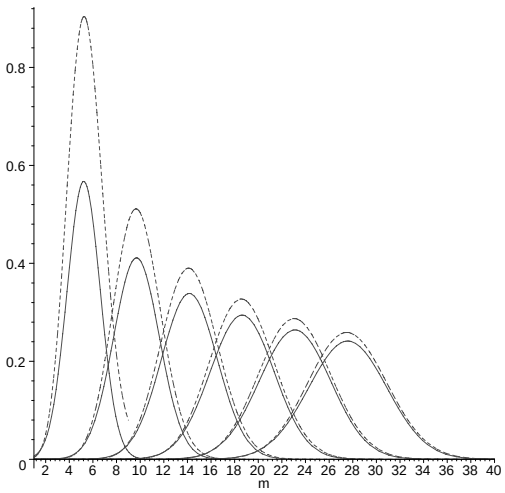


Figure 12: The convergence of the distribution given in Theorem 12 (dashed lines) to the Gaussian distribution with the appropriate mean and variance (solid lines). The plots correspond to the cases $n = 10k$, for $k = 1, 2, \ldots, 6$, in left-to-right order.

Theorem 14 *For random secondary structures of size n the quota of structures with m unpaired bases to all structures converges in distribution to a Gaussian variable with a speed of convergence that is $\mathcal{O}(n^{-1/2})$, $n \to \infty$.*

Figure 12 compares the asymptotic of Theorem 12 to the Gaussian distribution with the variance and the mean as computed before. Please note that the Gaussian distribution function has to be multiplied by two in order to get this convergence. Again, this results from the fact that $[z^n a^m]T(z, az) \neq 0$ if and only if $n > m$, $n - m$ even, and the resulting factor $(1 + (-1)^{n-m})$ which we have substituted by 2 in order to generate the plots.
We conclude this section by considering the question whether or not the order of a secondary structure has an influence on the number of unpaired bases. Therefore we determine the expected number of unpaired bases in a secondary structure of order p. For that purpose we determine the first partial derivative of $R_p(z, az)$ with respect to a and set $a := 1$ afterwards. Based on the representation (1) we find for that derivative:

$$\frac{2^p z^3 (z(z-3)-1) T_{2^p}\left(\frac{-1+2z+z^3}{2z^{5/2}}\right) + (z-1)^{3/2}(1+z(z-1)) U_{2^p-1}\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)}{4z^5 \sin^2\left(2^p \arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)}.$$

Here, $T_n(u)$ denotes the n-th Chebyshev polynomial of the first kind. By applying the identities [1, 22.3.15] and [1, 22.3.16] we can switch to the representation

$$\underbrace{\frac{2^p(z(z-3)-1)\cos\left(2^p \arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)}{4z^2\sin^2\left(2^p \arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)}}_{=:\hat{R}_p^{(1)}(z)}$$

$$+\underbrace{\frac{(z-1)(1+z(z-1))}{4z^{7/2}\sin\left(2^p \arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)\sin\left(\arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)}}_{=:\hat{R}_p^{(2)}(z)}.$$

Both, $\hat{R}_p^{(1)}(z)$ and $\hat{R}_p^{(2)}(z)$, have their singularity of smallest modulus at $z = z(p)$ for $z(p)$ being the smallest real solution of the equation $\frac{-1+2z+z^3}{2z^{5/2}} = \cos\left(\frac{2^p-1}{2^p}\pi\right) = -\cos(2^{-p}\pi)$, $p \in \mathbb{N}$. However, only $\hat{R}_p^{(1)}(z)$ contributes to the leading term of the asymptotic since $z(p)$ is a pole of second order in $\hat{R}_p^{(1)}(z)$ while it is a pole of first order in $\hat{R}_p^{(2)}(z)$. We thus have to expand $\hat{R}_p^{(1)}(z)$ at $z(p)$ which can be done by the following steps. We first consider

$$\frac{(x+\cos(2^{-p}\pi))^2}{\sin^2(2^p \arccos(x))}$$

for $x \to -\cos(2^{-p}\pi)$. Applying l'Hospital's rule we find that this limit is given by

$$\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}.$$

Next we have to consider

$$\frac{\left(1-\frac{z}{z(p)}\right)^2}{\left(\frac{-1+2z+z^3}{2z^{5/2}}+\cos(2^{-p}\pi)\right)^2},\ z \to z(p).$$

This limit can be inferred from (3) and proves to be given by

$$16\frac{z(p)^5}{(z(p)^3-6z(p)+5)^2}.$$

It remains to expand the residual parts of $\hat{R}_p^{(1)}(z)$ which yields

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

Thus, we find the following leading term of the expansion of $\hat{R}_p^{(1)}(z)$ around $z(p)$

$$\frac{1}{\left(1-\frac{z}{z(p)}\right)^2}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(z(p)(z(p)-3)-1)}{4z(p)^2}.$$

The $\mathcal{O}$-transfer method now implies that the asymptotic for the coefficient $[z^n]\left[\frac{\partial}{\partial a}R_p(z,za)\right]_{a=1}$ is given by:

$$n\left(\frac{1}{z(p)}\right)^n \frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(z(p)(z(p)-3)-1)}{4z(p)^2}.$$

Dividing this quantity by the asymptotic given in Theorem 3 provides the next theorem.

Theorem 15 *The expected number of unpaired bases in a random secondary structure of size n and order p is asymptotically given by*

$$n\left(1+\frac{4-4z(p)}{z(p)^2+z(p)-5}\right),\ n\to\infty,$$

where $z(p)$ is the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}}=-\cos(2^{-p}\pi)$ for which $0\le z(p)-\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\le 4^{-p}$ holds.

Since $z(p)$ converges exponentially fast to the reciprocal value of one plus the golden ratio, the factor of n in the previous theorem converges very rapidly. Its

p					
1	2	3	4	5	6
0.4963...	0.4580...	0.4498...	0.4478...	0.4473...	0.4472...

Figure 13: The percentage of unpaired bases to all bases in a large ($n\to\infty$) secondary structure of order p.

limit is numerically given by 0.4472135..., i.e. like in the general case, where secondary structures of arbitrary order are considered, about 44% of the bases are unpaired. Thus, only in the case of a small order it has an influence on the number of unpaired bases. The percentage of unpaired bases for smaller values of p can be found in Figure 13.

6 Additional Results

In this section we will consider hairpins and bulges as defined within the introduction. We will determine the expected number of hairpins and bulges in a secondary structure of size n and in a secondary structure of size n and order p. Furthermore, we will derive the expected length of a hairpin-loop and of a bulge in these structures. Our investigations are of interest since it is known that the contribution of different substructures to the total free energy depends on the

loop type and the number of unpaired bases in the loop of the substructure (see e.g. [28]). It is also in the intention of this section to give the reader an idea of how general our approach can be applied.
Let us start with the enumeration of the hairpins. Since each hairpin possesses exactly one hairpin-loop we will identify an entire hairpin by its loop for the purpose of enumeration. In order to derive the related generating functions we return to section 2 and the derivation of the substitutions presented there. If we consider the representation of a secondary structure as a word over the alphabet $\{(,|,)\}$ and the corresponding cases for the application of β we find that a hairpin-loop is generated exactly at those positions where $|^+$ is inserted in between $(\ )$ within a Dyck-word. Thus we can mark a hairpin-loop by variable h by translating $|^+$ into $h\frac{a}{1-a}$ instead of $\frac{a}{1-a}$ for the generating functions. In this way we get the following substitutions:

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

These substitutions can be applied to $\mathbf{T}$ and $\mathbf{R}_p$ in order to get the generating function in question. In the case of $\mathbf{T}$ we find for $a := z$

$$\frac{1 - 2z - hz^3 - \sqrt{(1-hz^3)(1-z(4+z(zh-4)))}}{2(1-z)z^2},$$

with the first partial derivative with respect to h evaluated at $h = 1$

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

The dominant singularity of this function is located at $z = z_d := \frac{3}{2} - \frac{1}{2}\sqrt{5}$ with the related expansion (terms relevant for the asymptotic only)

$$\frac{1}{20}\sqrt{30\sqrt{5}-50}\left(1-\frac{z}{z_f}\right)^{-1/2} - \frac{1}{160}\sqrt{950+486\sqrt{5}}\left(1-\frac{z}{z_d}\right)^{1/2} + \mathcal{O}\left(\left(1-\frac{z}{z_d}\right)^{3/2}\right).$$

This expansion translates into the following asymptotic for the coefficient at z^n, i.e. the total number of hairpins in all secondary structures of size n:

Lemma 3 *In all secondary structures of size n there are asymptotically*

$$z_d^{-n} \frac{1}{\sqrt{\pi n}} \left(\frac{1}{20}\sqrt{30\sqrt{5}-50}\left(1-\frac{1}{8n}\right) + \frac{1}{320n}\sqrt{950+486\sqrt{5}} \right) + \mathcal{O}\left(z_d^{-n} n^{-5/2}\right),$$

hairpins, $n \to \infty$.

By dividing this total number by the asymptotical number of secondary structures of size n we find:

Theorem 16 *The expected number of hairpins in a random secondary structure of size n is asymptotically given by*

$$\left(1-\frac{2}{5}\sqrt{5}\right)n+\frac{13}{20}-\frac{3}{20}\sqrt{5}+\mathcal{O}(n^{-1}),\ n\to\infty.$$

Now we apply these substitutions to $\mathbf{R}_p$ in order to compute the number of hairpins in a secondary structure of order p. For $a := z$ we find

$$\frac{\sqrt{hz}}{z-1}U_{2^p-1}^{-1}\left(\frac{-1+2z+hz^3}{2\sqrt{hz^5}}\right),$$

with the first partial derivative with respect to h evaluated at $h=1$

$$\underbrace{\frac{2^p(z^2-1+z)\cos\left(2^p\arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)}{4\sin^2\left(2^p\arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)z^2}}_{=:\hat{R}_p(z)}$$

$$-\frac{1+z(z-1)}{2z^{3/2}\sin\left(2^p\arccos\left(\frac{-1+2z+z^3}{2z^{5/2}}\right)\right)\sqrt{4-\frac{(-1+2z+z^3)^2}{z^5}}}.$$

Because of its pole of second order, only the part denoted by $\hat{R}_p(z)$ contributes to the leading term of the asymptotic. Furthermore, besides the non singular factor of $\hat{R}_p(z)$, this generating function is identical with $\hat{R}_p^{(1)}(z)$ of the previous section. Thus, $\hat{R}_p(z)$ possesses the same dominant singularity $z(p)$ as $\hat{R}_p^{(1)}(z)$ and we can reuse most parts of the computations which were necessary to determine the expansion at $z(p)$. In this way we find for $\hat{R}_p(z)$ the expansion

$$\frac{1}{\left(1-\frac{z}{z(p)}\right)^2}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(z(p)^2-1+z(p))}{4z(p)^2}$$

which translates into the following asymptotic for the coefficient at z^n:

Lemma 4 *In all secondary structures of size n and order p there are asymptotically*

$$nz(p)^{-n}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(z(p)^2-1+z(p))}{4z(p)^2}$$

hairpins, $n\to\infty$.

The computation of the expected value yields:

Theorem 17 *The expected number of hairpins in a random secondary structure of size n and order p is asymptotically given by*

$$n\left(1+\frac{4}{z(p)^2+z(p)-5}\right),\ n\to\infty,$$

for $z(p)$ the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}}=-\cos(2^{-p}\pi)$ for which $0\leq z(p)-\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\leq 4^{-p}$ holds.

Please recall, that $z(p)$ converges exponentially fast against $z(\infty):=\frac{3}{2}-\frac{1}{2}\sqrt{5}$. For $z(p)=z(\infty)$ we get exactly the same asymptotic as for secondary structures of size n and an arbitrary order.
Now we will consider the length of the hairpin-loops. For this purpose we have to mark each unpaired base within a hairpin-loop by a special variable (say h). It is obvious, that this leads to the substitutions

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

After setting $a:=z$ the same computations as before lead to the following results:

Theorem 18 *In all secondary structures of size n, there are asymptotically*

$$z_d^{-n}\frac{1}{\sqrt{\pi n}}\left(\frac{5^{3/4}}{10}\left(1-\frac{1}{8n}\right)+\frac{1}{160n}\sqrt{620+981\sqrt{5}}\right),$$

many unpaired bases located in hairpin-loops, $n\to\infty$, $z_d:=\frac{3}{2}-\frac{1}{2}\sqrt{5}$. The expected number of unpaired bases in a random hairpin-loop chosen from all secondary structures of size n is asymptotically given by

$$\phi+\frac{1}{2n}+\mathcal{O}(n^{-2}),$$

$\phi:=\frac{1}{2}+\frac{1}{2}\sqrt{5}$ the golden ratio, $n\to\infty$.

Theorem 19 *In all secondary structures of size n and order p, there are asymptotically*

$$nz(p)^{-n}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(-1+z(p)+z(p)^2)}{4z(p)^2(1-z(p))},$$

many unpaired bases located in hairpin-loops, $n\to\infty$, $z(p)$ the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}}=-\cos(2^{-p}\pi)$ for which $0\leq z(p)-\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\leq 4^{-p}$ holds. The expected number of unpaired bases in a random hairpin-loop chosen from all secondary structures of size n and order p is asymptotically given by

$$(1-z(p))^{-1}\overset{p\to\infty}{\to}\phi,$$

for ϕ the golden ratio, $n\to\infty$.

For small examples, the number of bulges and hairpins differs quite a bit. For example, if we consider the secondary structures of size 6, then there are 17 hairpins but only 2 bulges. Therefore, we are faced with the question whether or not this effect also exists for large structures. Again we can adjust our substitutions in order to count the bulges and the unpaired bases within the bulges. Within our approach, the bulges and the tails of a secondary structure are generated by inserting $|^\star$ at the appropriate places of a Dyck-word. Thus, translating $|^\star$ within the Motzkin word into $1+\frac{ba}{1-a}$ for the generating function marks the tails and each bulge by variable b. The resulting substitutions are given by

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

$$x := z^2\left(1+\frac{ba}{1-a}\right).$$

It remains to take the tails into account, i.e. to prevent that also the tails are marked by variable b which would lead to an overestimated number of bulges, which is an easy task. Each secondary structure can be decomposed into

$$|^\star(\{(,|,)\}^+)|^\star,$$

where the first and the last $|^\star$ correspond one-to-one with the tails. By the above substitutions both $|^\star$ together are translated into $\left(1+\frac{ba}{1-a}\right)^2$ while they should have been translated into $\left(1+\frac{a}{1-a}\right)^2$. Now, since

$$\left(1+\frac{ba}{1-a}\right)^2\frac{1}{(1-a+ba)^2} = \left(1+\frac{a}{1-a}\right)^2,$$

it is sufficient to multiply the resulting generating functions by $\frac{1}{(1-a+ba)^2}$ in order to correct the overestimated number of bulges. Proceeding in this way, we find for $a := z$ the following representations for the generating function of all secondary structures, respectively of all secondary structures of order p, with each bulge marked by b:

$$\frac{2z^3}{(1-z)(1-2z+z^3-2z^3b+z^4b-z^4b^2+\sqrt{\mu})},$$

for $\mu := (-1+z+(-1+b)z^2)(1+2(-1+b)z^2-z^3+(-2+b)(-1+b)z^4)(-1+z(3+(-2+b)z))$, respectively

$$\frac{\sqrt{z}}{(z-1)\sqrt{(1+(b-1)z)^3}}U_{2^p-1}^{-1}\left(\frac{-1+2z+(2b-1)z^3+(b-1)bz^4}{2\sqrt{z^5(1+(b-1)z)^3}}\right).$$

Since the mathematics that has to be done now in order to derive the corresponding results remains the same as for the hairpins, we will present no details.

Theorem 20 *The total number of bulges in all secondary structures of size n is asymptotically given by*

$$z_d^{-n}\frac{1}{\sqrt{\pi n}}\left(\frac{5^{3/4}}{10}\left(1-\frac{1}{8n}\right)-\frac{9}{160n}\sqrt{420+461\sqrt{5}}\right),$$

$n\to\infty$, $zd:=\frac{3}{2}-\frac{1}{2}\sqrt{5}$. For a random secondary structure of size n the expected number of bulges is asymptotically

$$\frac{3\sqrt{5}-5}{10}n-\frac{61}{20}+\frac{21}{20}\sqrt{5}+\mathcal{O}(n^{-1}),\ n\to\infty.$$

Thus there are about 1.618 times as many bulges than hairpins, i.e. the effect that can be observed for small structures has been inverted.

Theorem 21 *The total number of bulges in all secondary structures of size n and order p is asymptotically given by*

$$nz(p)^{-n}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin^2(2^{-p}\pi)}{(2^p)^2}\frac{-2^p(-3+z(p)(3-z(p)))}{4z(p)},$$

$n\to\infty$, $z(p)$ the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}}=-\cos(2^{-p}\pi)$ for which $0\le z(p)-\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\le 4^{-p}$ holds. For a random secondary structure of size n and order p the expected number of bulges is asymptotically

$$n\frac{z(p)(3+z(p)(z(p)-3))}{5-z(p)-z(p)^2},\ n\to\infty.$$

In Figure 14 you find a plot of the slope of the expected number of bulges in a random secondary structure of order p. As you can see the slope converges very fast to its limit.

As for the hairpin-loops we now will determine the number of unpaired bases which belong to the bulges. The appropriate substitutions result from the transformation of $|^\star$ into $\frac{1}{1-ba}$ with the factor $\left(\frac{1-ba}{1-a}\right)^2$ for correcting the overestimated number of unpaired bases. The same methods as before can be used to prove the following theorems:

Theorem 22 *The total number of unpaired bases within bulges of secondary structures of size n is asymptotically given by*

$$z_d^{-n}\frac{1}{\sqrt{\pi n}}\left(\frac{1}{20}\sqrt{50+30\sqrt{5}}\left(1-\frac{1}{8n}\right)-\frac{1}{320n}\sqrt{532890+260246\sqrt{5}}\right),$$

$n\to\infty$, $z_d:=\frac{3}{2}-\frac{1}{2}\sqrt{5}$. For secondary structures of size n the expected number of unpaired bases in a random bulge is asymptotically

$$\phi+\frac{1035+279\sqrt{5}-\sqrt{1301230+532980\sqrt{5}}}{160n}+\mathcal{O}(n^{-2}),$$

$\phi:=\frac{1}{2}+\frac{1}{2}\sqrt{5}$ the golden ratio, $n\to\infty$.

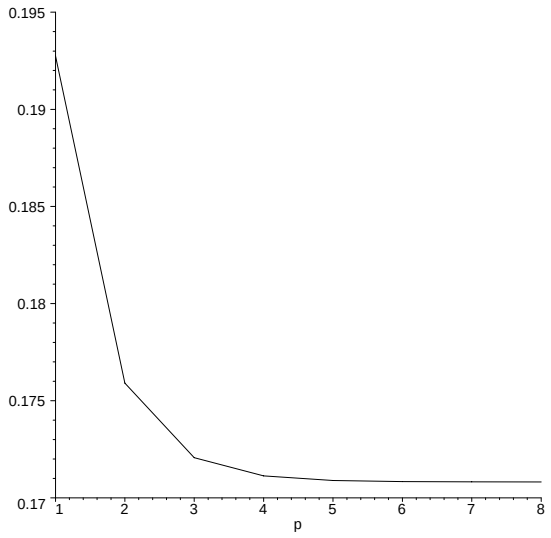


Figure 14: The slope of the expected number of bulges for different values of p.

Thus, the expected length of a hairpin-loop and that of a bulge only differ in the n^{-1} term where the difference is only marginal ($\frac{1}{2n}$ compared to $0.49960\ldots n^{-1}$). For the secondary structures of a given order we find

Theorem 23 *The total number of unpaired bases within bulges of secondary structures of size n and order p is asymptotically given by*

$$nz(p)^{-n}\frac{16z(p)^5}{(z(p)^3-6z(p)+5)^2}\frac{\sin(2^{-p}\pi)}{(2^p)^2}\frac{-2(3+z(p)(z(p)-3))}{4z(p)(z(p)-1)},$$

$n\to\infty$, $z(p)$ the smallest real solution of the equation $\frac{z^3+2z-1}{2z^{5/2}}=-\cos(2^{-p}\pi)$ for which $0\le z(p)-\left(\frac{3}{2}-\frac{1}{2}\sqrt{5}\right)\le 4^{-p}$ holds. For secondary structures of size n and order p the expected number of unpaired bases in a random bulge is asymptotically

$$(1-z(p))^{-1}\overset{p\to\infty}{\to}\phi,$$

for ϕ the golden ratio, $n\to\infty$.

Finally, please note that it is obviously no problem to compute higher moments for the results presented from our generating functions by just considering higher derivatives with respect to h, resp. b.

7 Conclusions

In this paper we have solved the old problem of how to determine the number of secondary structures of size n and order p raised by Waterman in 1978. The

approach that led to the solution of this problem is general enough to allow us to derive many more results in the context of secondary structures, some of them were presented here. Whenever the order p of the structures was used as a parameter, then $z(p)$, the smallest real zero of a polynomial of sixth degree, pops up. This zero has yet not been determined precisely, even though there are methods (based on the work of Klein) which would provide a representation of the zeros of a sextic by means of hypergeometic functions in one variable. It looks like a challenging task to use these methods in order to find an exact representation of $z(p)$.

The combinatorial model considered in this paper disregards numerous details of the folding mechanism for real single stranded nucleic acids. For example we do not take into account that base-pairing is not possible between arbitrary pairs of nucleotides. In reality the positions at which base pairs may occur is dependent on the base composition of the actual sequence. In [28] a stochastic approach to this problem going back to [11] and [25] is discussed. Let us assume that the different bases possess a Bernoulli distribution and that $p(A)$ (resp. $p(C)$; $p(G)$; $p(U)$) denotes the probability for the occurence of base A (resp. C; G; U) in the primary structure. Then $p := 2(p(A)p(U)+p(C)p(G))$ is the probability that any two bases can form a hydrogen bound. This idea can be generalized by considering an arbitrary set of symbols together with an appropriate probability p which then usually is denoted *stickiness* [14]. It is possible to consider random sequences with these Bernoulli distributions by setting z to $z\sqrt{p}$ within our generating functions. In this case the generating functions do not provide explicit enumeration results but their coefficients are related to the expectation of the parameter considered. As an example $[z^n]T(z\sqrt{p}, z)$ is the expected number of secondary structures of size n. Numerous results for this model of a random secondary structure can be found in [18].

References

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

[2] E. A. Bender, *Central and Local Limit Theorems Applied in Asymptotic Enumeration*, J. Comb. Theory (A) **15** (1973), 91-111.

[3] L. Comtet, *Advanced Combinatorics*, D. Reidel, 1974.

[4] M. Drmota, *Asymptotic Distributions and a Multivariate Darboux Method in Enumeration Problems*, Journal of Combinatorial Theory, Series A **67**, 169-184, 1994.

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

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

[7] P. Flajolet and R. Sedgewick, *Analysis of Algorithms*, Addison-Wesley Publishing Company, 1996

[8] P. Flajolet and R. Sedgewick, *The Average Case Analysis of Algorithms: Multivariate Asymptotics and Limit Distributions*, INRIA rapport de recherche **3162** (1997).

[9] de Gennes in *C. Domb and M.S. Green, eds., Phase Transition and Critical Phenomena, vol. 3*, Academic Press, London, 1976.

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

[11] 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.

[12] F. Klein, *Lectures on the Icosahedron and the Solution of Equations of the Fifth Degree*, Dover, 1956.

[13] D. E. Knuth, *The Art of Computer Programming, Vol. 3, Sorting and Searching*, Addison Wesley, 1998.

[14] 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.

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

[16] 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.

[17] M. E. Nebel, *A Unified Approach to the Analysis of Horton-Strahler Parameters of Binary Tree Structures*, Frankfurter Informatik-Berichte 1/01, Institut für Informatik, Johann Wolfgang Goethe-Universität, Frankfurt a. M., 2001.

[18] M. E. Nebel, *Investigation of the Bernoulli-Model for RNA Secondary Structures*, Frankfurter Informatik-Berichte 3/01, Institut für Informatik, Johann Wolfgang Goethe-Universität, Frankfurt a. M., 2001.

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

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

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

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

[23] 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.

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

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

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

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

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