ELSEVIER

Discrete Applied Mathematics 88 (1998) 207–237

DISCRETE
APPLIED
MATHEMATICS

# Combinatorics of RNA secondary structures

Ivo L. Hofacker[a,*], Peter Schuster[a,b], Peter F. Stadler[a,b]

[a] *Institut f. Theoretische Chemie, Univ. Wien, Währingerstr. 17, A-1090 Vienna, Austria*
[b] *The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA*

Received 11 September 1996; received in revised form 11 February 1998; accepted 15 April 1998

**Abstract**

Secondary structures of polynucleotides can be viewed as a class of planar vertex-labeled graphs. We compute recursion formulae for enumerating a variety sub-classes of and classes of sub-graphs (structural elements) of secondary structure graphs. First order asymptotics are derived and their dependence on the composition of the underlying nucleic acid sequences is discussed. © 1998 Elsevier Science B.V. All rights reserved.

*AMS classification*: 05A15; 05A16; 05C30; 92C40

*Keywords*: Planar graphs; Generating functions; Asymptotic enumeration; Secondary structure

## 1. Introduction

Presumably, the most important problem and the greatest challenge in present day theoretical biophysics deals with deciphering the code that transforms sequences of biopolymers into spatial molecular structures. A sequence is properly visualized as a string of symbols which together with the environment encodes the molecular architecture of the biopolymer. In case of one particular class of biopolymers, the ribonucleic acid (RNA) molecules, decoding of information stored in the sequence can be properly decomposed into two steps. Transformation of the string into a planar graph, and folding of the string into a three-dimensional structure under conservation of the neighborhood relation determined by the graph. We are concerned here only with the first step, the transformation of the sequence into the graph (Fig. 1), which is much simpler than other known sequence-to-structure relations in biophysics. We are not concerned here with the physical rules that govern this transformation. Instead we are interested in the combinatorics of RNA secondary structures which in essence is an exercise in combining structural elements into valid structures under certain additional constraints.

[*] Corresponding author. Fax: +43 1 4277 52793; e-mail: ivo@tbi.univie.ac.at.

*PII* S0166-218X(98)00073-0

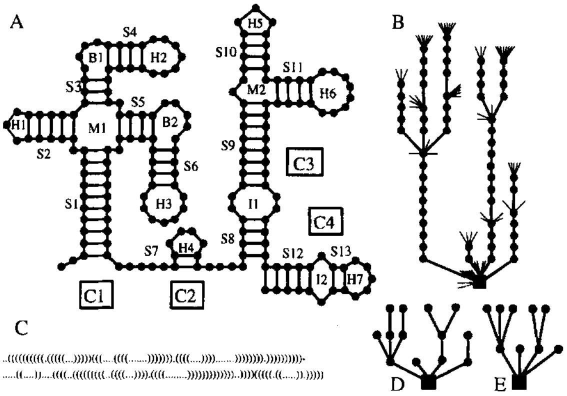

Fig. 1. Representations of secondary structures. The notation **A** is common in biology. Structure elements are indicated as follows: *H* hairpin loops, *I* interior loops, *B* bulges, *M* multiloops; *S* stacks. The structure consists of four components, indicated as C1–C4. **B** is the corresponding tree notation, and **C** is the linear encoding of this tree. For details see Section 2.2. **D** is a coarse grained representation obtained from **B** by contracting each stack to a single vertex and omitting the half-vertices representing the unpaired positions. **E** is the homeomorphically irreducible tree obtained from **D**.

Previous results on combinatorial aspects of secondary structures of RNA molecules are due to Waterman and coworkers [21, 23, 28, 34–37]. Particularly important for the work reported here are a recursion for the number of different secondary structures formed by strings of constant length [34] and the analytical expression for its asymptotic values [28]. Secondary structures are labeled planar graphs and as such they are closely related to the *linked diagrams* of Touchard [13, 14, 26, 27, 32].

In Section 2 we introduce the basic definitions of secondary structures and recall their various representations. Section 3 presents the recursion formulas for the exact enumeration of various types of constrained secondary structures as well as their structural elements. Constrained secondary structures are of primary importance in biophysics since not every conceivable element of a secondary structure will be found in reality. For example, hairpin loops containing one or two nucleotides are so strongly disfavored by the energetics that they do not occur in RNA secondary structures. In Section 4 first-order asymptotics to these recursions are devised. Although the class of graphs formed by secondary structures is interesting in its own rights, secondary structures in biology make sense only when they are related to sequences. Implications resulting from the condition that secondary structures have to be built on sequences are discussed in Section 5. The results reported here are particularly interesting in relation to the data which were obtained from RNA secondary structure statistics performed by folding large ensembles of sequences into minimum free energy structures [6–9]. The asymptotic values show the influence of the logic of base pairing as expressed in terms of *stickiness*. Stickiness accounts for the possible base pairings supported by the nucleotide alphabet but ignores the energetic effect of different strengths of the base pairs. Numerically computed data based on empirical energetic parameters include both effects, and the comparison allows to separate the influence of the pairing logic from the energetics. A detailed comparison can be found in Ref. [30].

## 2. Secondary structures and structural elements

### 2.1. Definitions

**Definition 2.1** (*Waterman* [34]). A *secondary structure* is a vertex-labeled graph on $n$ vertices with an adjacency matrix $A$ fulfilling

(i) $a_{i,i+1} = 1$ for $1 \leqslant i \leqslant n-1$;
(ii) For each $i$ there is at most a single $k \neq i-1, i+1$ such that $a_{ik} = 1$;
(iii) If $a_{ij} = a_{kl} = 1$ and $i<k<j$ then $i<l<j$.

We will call an edge $(i,k)$, $|i-k| \neq 1$ a bond or base pair. A vertex $i$ connected only to $i-1$ and $i+1$ will be called unpaired. A vertex $i$ is said to be *interior* to the base pair $(k,l)$ if $k<i<l$. If, in addition, there is no base pair $(p,q)$ such that $k<p<i<q$ we will say that $i$ is *immediately interior* to the base pair $(k,l)$.

**Definition 2.2.** A secondary structure consists of the following structure elements:

(i) A *stack* consists of subsequent base pairs $(p-k, q+k)$, $(p-k+1, q+k-1)$, $\ldots, (p,q)$ such that neither $(p-k-1, q+k+1)$ nor $(p+1, q-1)$ is a base pair. $k+1$ is the *length* of the stack, $(p-k, q+k)$ is the terminal base pair of the stack.
(ii) A *loop* consists of all unpaired vertices that are immediately interior to some base pair $(p,q)$.
(iii) An *external vertex* is an unpaired vertex which does not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 1 or $n$ it is a free end, otherwise it is called joint.

**Lemma 2.3.** *Any secondary structure $\mathcal{S}$ can be uniquely decomposed into stacks, loops, and external elements.*

**Proof.** Each vertex which is contained in a base pair belongs to a unique stack. Since an unpaired vertex is either external or immediately interior to a unique base pair the decomposition is unique: Each loop is characterized uniquely by its “closing” base pair. □

**Definition 2.4.** A stack $[(p,q), \ldots, (p+k, q-k)]$ is called *terminal* if $p-1=0$ or $q+1=n+1$ or if the two vertices $p-1$ and $q+1$ are not interior to any base pair. The sub-structure enclosed by the terminal base pair $(p,q)$ of a terminal stack will be called a *component* of $\mathcal{S}$. We will say that a structure on $n$ vertices has a terminal base pair if $(1,n)$ is a base pair.

**Lemma 2.5.** *A secondary structure may be uniquely decomposed into components and external vertices. Each loop is contained in a component.*

The proof is trivial. Note that by definition the open structure has 0 components. The loops of a secondary structure graph form its unique minimal cycle basis [16].

**Definition 2.6.** The degree of a loop is given by 1 plus the number of terminal base pairs of stacks which are interior to the closing bond of the loop. A loop of degree 1 is called hairpin (loop), a loop of a degree larger than 2 is called multiloop. A loop of degree 2 is called bulge if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of degree 2 is termed interior loop.

**Definition 2.7.** Let $\mathcal{S}$ be an arbitrary secondary structure. Denote by $\Omega(\mathcal{S})$ the unique secondary structure that is obtained from $\mathcal{S}$ by means of the following procedure:

(i) For each hairpin, open its stack and add the corresponding bases to the hairpin loop.
(ii) If a bulge or interior loop follows, then add its digits also to the hairpin and continue by opening its stack.
(iii) If a multiloop or a joint follows, then add the now unpaired digits to the multiloop and stop.

Waterman [34] used the above procedure to define the order $\omega(\mathcal{S})$ of a secondary structure as the smallest number of repetitions of $\Omega$ necessary to obtain the open structure. Of course, the open structure has order $\omega = 0$ and any structure without a multiloop has order $\omega = 1$.

### *2.2. Representation of secondary structures*

A secondary structure $\mathcal{S}$ can be translated into a rooted ordered tree (linear tree) $\Upsilon$ by introducing an additional root and representing a base pair $(p,q)$ by a vertex $x$ such that the sons $y_1,\ldots,y_k$ of $x$ correspond to the base pairs $(p_1,q_1)\ldots(p_k,q_k)$ immediately interior to $(p,q)$ [6, 7]. For each unpaired vertex $z$ a half-vertex is added to the vertex representing the closing pair of the loop containing $z$. (For external digits this is the root.) The tree-representation of a secondary structure is shown in Fig. 1B.

A string representation **S** can by obtained by the following rules:

(i) If vertex $i$ is unpaired then $\mathbf{S}_i =$ ".".
(ii) If $(p,q)$ is a base pair and $p<q$ then $\mathbf{S}_p =$ "(" and $\mathbf{S}_q =$ ")".

These rules yield a sequence of matching brackets and dots [33] (cf. Fig. 1C). A related representation is derived in Ref. [11].

Waterman's definition of secondary structures implies that each branch of the corresponding tree representation $\Upsilon$ has at least one terminal half-vertex, or equivalently, each matching pair of brackets contains at least one dot. In biological applications the number of unpaired positions is at least 3, implying at least 3 dots within each pair of matching brackets. From the combinatorial point of view it makes perfect sense to consider the general problem with a minimum number $m \geqslant 0$ of unpaired vertices in each hairpin loop. In fact, for $m = 0$ one recovers three well-known Motzkin families [5, 28].

For some applications it is useful to work with simplified representations [24, 25]. A tree $T$ is obtained by denoting a stack by single vertex. In terms of the representation $\Upsilon$ this means that each vertex of degree 2 not carrying a half-vertex (except for the root) is merged with its son and then the half-vertices are removed (cf. Fig. 1D). The number of vertices in $T$ is then just the number of stacks in $\mathscr{S}$, the number of components of $\mathscr{S}$ coincides with the number of sons of the root in $T$. An alternative "coarse grained" representation of secondary structures is the homeomorphically irreducible tree $\mathscr{H}$ corresponding to $\Upsilon$ which is obtained by removing all vertices of degree 2 (except for the root) and all half-vertices. Again the number of components of $\mathscr{S}$ equals the number of sons of the root. Waterman's degree $\omega$ coincides with the height of $\mathscr{H}$ (cf. Fig. 1E).

### 2.3. *The basic recursion*

A secondary structure on $n+1$ digits may be obtained from a structure on $n$ digits either by adding a free end at the right-hand end or by inserting a base pair $(1, k+2)$. In the second case the substructure enclosed by this pair is an arbitrary structure on $k$ digits, and the remaining part of length $n-k-1$ is also an arbitrary valid secondary structure. Therefore, we obtain the following recursion formula for the number $S_n$ of secondary structures:

$$\begin{aligned} &S_{n+1} = S_n + \sum_{k=m}^{n-1} S_k S_{n-k-1}, \quad n \geqslant m+1, \\ &S_0 = S_1 = \cdots = S_{m+1} = 1. \end{aligned} \tag{1}$$

Eq. (1) has first been derived by Waterman [34]; $m$ denotes the minimum number of unpaired digits in a hairpin loop. Note that our definition of $S_n$ differs from Waterman's for $n < m$: he used $S_n = 0$.

The above recursion can be used to develop an algorithm for generating random secondary structures with a uniform distribution

$$\text{Prob}\{\mathscr{S}\} = 1/S_n \tag{2}$$

in the *shape space* of all secondary structures over a given chain length, see [30].

## 3. Recursions

### 3.1. Structures with certain properties

Let $J_n(b)$ denote the number of structures on $n$ vertices with exactly $b$ components. The derivation of the recursion relations parallels the argument leading to Eq. (1):

$$\begin{aligned}&J_{n+1}(b)=J_n(b)+\sum_{k=m}^{n-1} S_k J_{n-k-1}(b-1), \quad b>0,\, n\geqslant m+1,\\ &J_n(b)=0, \quad b>0,\ n\leqslant m+1, \quad J_n(0)=1, \quad n\geqslant 0\end{aligned} \tag{3}$$

because adding an unpaired digit to a structure on $n$ digits does not change the number of components, while introducing an additional bracket makes the bracketed part of length $k$ a single component and does not affect the remainder of the sequence.

Let $H_n(b)$ denote the number of structures with exactly $b$ base pairs (bonds) on $n$ vertices. The recursion

$$\begin{aligned}&H_{n+1}(b)=H_n(b)+\sum_{k=m}^{n-1}\sum_{\ell=0}^{b-1} H_k(\ell)H_{n-k-1}(b-\ell-1), \quad b>0, \quad n\geqslant m+1,\\ &H_n(b)=0, \quad b>0,\ n\leqslant m+1, \quad H_n(0)=1, \quad n\geqslant 0\end{aligned} \tag{4}$$

is also immediate. One just has to observe that an additional sum over the number of unpaired digits in the newly bracketed part of the structure has to be introduced. This recursion has also been considered in Ref. [12]. Recently, Schmitt and Waterman [23] obtained the closed expression

$$H_n(b)=\frac{1}{b}\binom{n-b}{b+1}\binom{n-b-1}{b-1}$$

for the special case $m=1$. Analogously, we obtain

$$\begin{aligned}&E_{n+1}(b)=E_n(b-1)+\sum_{k=m}^{n-1} S_k E_{n-k-1}(b), \quad b>0,\ n\geqslant m+1,\\ &E_{n+1}(0)=\sum_{k=m}^{n-1} S_k E_{n-k-1}(0),\\ &E_n(n)=1, \quad E_n(b)=0 \quad b\neq n,\ n\leqslant m+1\end{aligned} \tag{5}$$

for the number $E_n(b)$ of structures with $b$ external digits.

It is a bit more tricky to find a recursion for the number $N_n(b)$ of structures with a given number of stacks. We introduce the auxiliary variable $Z_n(b)$ counting the number of secondary structures with exactly $b$ stacks *given* that the $3'$ and $5'$ ends are paired.

We obtain then

$$\begin{aligned} N_{n+1}(b) &= N_n(b) + \sum_{k=m}^{n-1}\sum_{\ell=0}^{b} Z_{k+2}(\ell)N_{n-k-1}(b-\ell), \quad b>0,\ n\geqslant m+1, \\ N_n(0) &= 1, \quad N_n(b)=0, \quad b>0,\ n\leqslant m+1. \end{aligned} \tag{6}$$

For the auxiliary variably we find

$$Z_n(b) = Z_{n-2}(b) + N_{n-2}(b-1) - Z_{n-2}(b-1), \quad Z_0(b)=Z_1(b)=0 \tag{7}$$

by enclosing structures on $n-2$ digits by a base pair.

Let $A_n(b)$ denote the number of structures with exactly $b$ hairpins. Since the number of hairpins is unchanged by enclosing a substructure which already contains a base pair in an additional base pair we get

$$\begin{aligned} A_{n+1}(b) &= A_n(b) + \sum_{k=m}^{n-1}\left[\sum_{l=1}^{b} A_k(\ell)A_{n-k-1}(b-l) + A_{n-k-1}(b-1)\right], \\ &\qquad n\geqslant m+1, \\ A_n(b) &= \delta_{0,b}, \quad n\leqslant m+1, \end{aligned} \tag{8}$$

where $\delta_{0,b}$ is Kronecker's $\delta$, i.e. $\delta_{0,0}=1$ and $\delta_{0,b}=0$, $b\neq 0$.

### 3.2. Structure elements

The total number $U_{n+1}$ of unpaired bases in the set of all structures with $n+1$ bases can be computed as follows: adding an unpaired base to each structure on $n$ digits we obtain their $U_n$ unpaired digits plus the $S_n$ newly added ones. Introducing a base pair $(1,k+2)$ we have $S_k$ times all the unpaired digits in the reminder of the sequence plus all the unpaired digits in the newly bracketed part of length $k$ times the the number of structures that can be formed from the reminder of the structure. Summing over $k$ we find

$$\begin{aligned} U_{n+1} &= (U_n + S_n) + \sum_{k=m}^{n-1}[S_k U_{n-k-1} + S_{n-k-1}U_k], \quad n\geqslant m+1, \\ U_n &= n, \quad n\leqslant m+1. \end{aligned} \tag{9}$$

Denote the total number of base pairs by $P_n$. It is clear that $2P_n + U_n = nS_n$. For sake of completeness we state the recursion for $P_n$:

$$P_{n+1} = P_n + \sum_{k=m}^{n-1}\{S_k P_{n-k-1} + S_{n-k-1}(P_k + S_k)\}, \quad P_n = 0,\ n\leqslant m+1. \tag{10}$$

By an analogous reasoning we find for the total number $I_n$ of components in the set of all secondary structures on $n$ vertices:

$$I_{n+1} = I_n + \sum_{k=m}^{n-1} S_k[I_{n-k-1} + S_{n-k-1}], \quad I_n = 0 \; n \leqslant m+1. \tag{11}$$

The number $N_{n+1}$ of stacks in the set of structures on $n+1$ digits consists of all stacks on $n$ digits plus all stacks in the tail times the number of structures with the newly introduced base pair plus all stacks within the newly formed base pair times the number of structures in the tail. The newly formed base pair introduces an additional stack for all the $S_k - S_{k-2}$ structures in its interior without a terminal base pair. (For the $S_{k-2}$ structures with terminal base pair a stack is elongated.) Therefore

$$\begin{aligned} N_{n+1} &= N_n + \sum_{k=m}^{n-1} \{S_k N_{n-k-1} + S_{n-k-1}(N_k + S_k)\} - \sum_{k=m+2}^{n-1} S_{k-2} S_{n-k-1} \\ &\text{for } n \geqslant m+1 \text{ and } \quad N_n = 0, \quad n \leqslant m+1. \end{aligned} \tag{12}$$

Let $Q_n(b)$ denote the number of loops with $b$ unpaired digits in the set of all secondary structures. For $n+1$ vertices we retain all loops from the set of loops on $n$ digits by adding a vertex to the $3'$ end. In addition, we have to count all loops in the tail-substructure for each possible structure that lies interior to the new base pair. The third contribution consists of all loops interior to the new base pair times all possible structures in the tail. A loop with $b$ unpaired vertices remains unchanged and each structure with exactly $b$ external vertices within the new base pair gives rise to an additional loop with $b$ unpaired digit:

$$\begin{aligned} Q_{n+1}(b) &= Q_n(b) + \sum_{k=m}^{n-1} \{Q_{n-k-1}(b) S_k + S_{n-k-1}[Q_k(b) + E_k(b)]\}, \\ & n \geqslant m+1, \; b > 0, \end{aligned} \tag{13}$$

$$Q_n(b) = 0, \quad n \leqslant m+1.$$

The recursion for loops without unpaired digits is slightly different because structures without external digits located within the new base pair do not lead to a loop if they consist of a single component, i.e., if they end in base pair. (In this case the terminal stack is elongated.) There are $S_{k-2}$ such structures on $k$ vertices:

$$\begin{aligned} Q_{n+1}(0) &= Q_n(0) + \sum_{k=m}^{n-1} \{Q_{n-k-1}(0) S_k + S_{n-k-1}[Q_k(0) + E_k(0)]\} \\ & - \sum_{k=m+2}^{n-1} S_{n-k-1} S_{k-2}, \quad n \geqslant m+1, \end{aligned} \tag{14}$$

$$Q_n(0) = 0, \quad n \leqslant m+1.$$

Let $W_n(b)$ denote the number of stacks with exactly $b$ base pairs in the set of secondary structures. From a stack with $b$ base pairs in a structure of length $n$, one can produce a stack of with $b+1$ pairs in a structure of length $n+2$ by inserting a new base pair immediately exterior of the existing stack. Therefore, we have

$$\begin{aligned} &W_{n+2}(b+1)=W_n(b), \quad b>1, \ n\geqslant m, \\ &W_n(b)=0, \quad n\leqslant m+1. \end{aligned} \tag{15}$$

For $b=1$ we have to construct a recursion in the usual way. There are $S_k - S_{k-2}$ structures that will form a new stack of length 1 when enclosed by a new base pair $(1,k+2)$. Conversely, for $S_{k-2}-S_{k-4}$ structures an enclosing stack of length 1 will be elongated by the new pair. We therefore have

$$\begin{aligned} W_{n+1}(1)=W_n(1)&+\sum_{k=m}^{n-1}[W_{n-k-1}(1)S_k - S_{n-k-1}W_k(1)] \\ &+\sum_{k=m}^{n-1}S_kS_{n-k-1} - 2\sum_{k=m+2}^{n-1}S_{k-2}S_{n-k-1} \\ &+\sum_{k=m+4}^{n-1}S_{k-4}S_{n-k-1}, \end{aligned} \tag{16}$$

$$W_n(1)=0 \quad \text{for } n\leqslant m+1.$$

Let $L_n(d)$ denote the number of loops of degree $d$ in the set of all secondary structures. By $Y_n$ and $B_n$, resp., we will denote the number of interior loops and bulges. Let us start with bulges and interior loops: The number of structures that yield an interior loop at their "end" when they are inclosed by an additional base pair equals the number $J_{n-2}(1)$ of structures having a free end on both sides, because structures with zero components would yield a hairpin while structures with more than one components would give rise to a multi-loop. In order to compute the number $X_n^*$ of structures that form a bulge when enclosed by an additional base pair we observe that a bulge is formed if and only the enclosed structure has only a single component and neither a base pair connecting the ends (for these the terminal stack is elongated) nor free ends on both sides. There are $S_{n-2}$ structures resulting in a stack elongation if $n\geqslant m+2$ (and none otherwise). Consequently, we have

$$X_n^* = J_n(1) - J_{n-2}(1) - S_{n-2} \quad n\geqslant m+2. \tag{17}$$

The recursions for loops of degree 2 are now straightforward:

$$\begin{aligned} B_{n+1} &= B_n + \sum_{k=m}^{n-1}\{S_kB_{n-k-1} + S_{n-k-1}[B_k + X_k^*]\}, \\ Y_{n+1} &= Y_n + \sum_{k=m}^{n-1}\{S_kY_{n-k-1} + S_{n-k-1}[Y_k + J_{k-2}(1)]\}, \end{aligned}$$

$$\begin{aligned}L_{n+1}(2) = L_n(2) + \sum_{k=m}^{n-1} \{S_k L_{n-k-1}(2) + S_{n-k-1}[L_k(2) + J_k(1)]\}\\ - \sum_{k=m+2}^{n-1} S_{n-k-1} S_{k-2},\end{aligned} \tag{18}$$

$$B_n = Y_n = L_n(2) = 0, \quad n \leqslant m+1.$$

Hairpins are generated either by stack-elongation of a structure with a single hairpin or by enclosing the open structure into the additional bracket. Thus,

$$L_{n+1}(1) = L_n(1) + \sum_{k=m}^{n-1} \{S_k L_{n-k-1}(1) + S_{n-k-1}[L_k(1) + 1]\} \quad n \geqslant m+1, \tag{19}$$

$$L_n(1) = 0 \quad n \leqslant m+1.$$

For multi-loops, finally, we obtain the recursion

$$\begin{aligned}L_{n+1}(d) = L_n(d) + \sum_{k=m}^{n-1} \{S_k L_{n-k-1}(d) + S_{n-k-1}[L_k(d) + J_k(d-1)]\}\\ \text{for } d \geqslant 2,\ n \geqslant m+1,\end{aligned} \tag{20}$$

$$L_n(d) = 0 \quad \text{for } n \leqslant m+1.$$

Summing over all loop degrees $d$ we recover the recursion for the total number of stacks, since for each stack there is exactly one loop.

The total number of external digits, $E_n$, can be obtained directly as $\sum_b bE_n(b)$. For sake of completeness, we mention that it fulfills the recursion

$$E_{n+1} = E_n + S_n + \sum_{k=m}^{n-1} S_k E_{n-k-1}, \quad n \geqslant m+1, \tag{21}$$

$$E_n = n \quad n \leqslant m+1.$$

### 3.3. Secondary structures of a given order

Let $D_n(c,\omega)$ be the number of secondary structures with $c$ components and order $\omega$. Furthermore, let $D_n^*(\omega)$ be the number of structures which yield a structure of order $\omega$ when enclosed by an additional base pair. The numbers $D_n(c,\omega)$ satisfy the recursion

$$\begin{aligned}D_{n+1}(c,\omega) = D_n(c,\omega) + \sum_{k=m}^{n-1} \Bigg\{ D_k^*(\omega) \sum_{\ell=0}^{\omega-1} D_{n-k-1}(c-1,\ell)\\ + D_{n-k-1}(c-1,\omega) \sum_{\ell=0}^{\omega-1} D_k^*(\ell) + D_k^*(\omega) D_{n-k-1}(c-1,\omega) \Bigg\},\end{aligned} \tag{22}$$

$$D_n(0,0) = 1, \qquad D_n(0,d) = D_n(c,0) = 0, \quad n \leqslant m+1$$

because a structure with a base pair $(1,k+2)$ has order $d$ and $c$ components iff either the bracketed part has order $\omega$ and the tail has a order at most $\omega$ and $c-1$ components or the bracketed part has a degree smaller than $\omega$ and the tail has $c-1$ components and order $\omega$. It remains to calculate $D_n^*(\omega)$. By inspection we find for $n>m$

$$\begin{aligned} D_n^*(0) &= 0,\\ D_n^*(1) &= 1+D_n(1,1),\\ D_n^*(\omega) &= D_n(1,\omega)+\sum_{\ell=2}^{\infty}D_k(\ell,\omega-1), \quad \omega\geqslant 2, \end{aligned} \tag{23}$$

while for $n\leqslant m$ we have $D_n^*(\omega)=0$. There is no structure of order 0 with a bracket in it; order $\omega=1$ is obtained by either bracketing the open structure or by bracketing a structure with a single component and order 1. If the bracketed part has only a single components its order is preserved by adding a terminal bracket. If it consists of more than one components, the addition of the multiloop increases the order by one.

Summing over the number of components we obtain the number of structures with given order $\tilde{D}_n(\omega)$. Let us further introduce the number $D'_n(1)$ of structures of order at most one. It is easy to derive the following system of recursions from the above ones:

$$\begin{aligned} \tilde{D}_{n+1}(\omega) &= \tilde{D}_n(\omega)+\sum_{k=m}^{n-1}\left\{D_k^*(\omega)\sum_{\ell=0}^{\omega-1}\tilde{D}_{n-k-1}(\ell)+\tilde{D}_{n-k-1}(d)\sum_{\ell=0}^{\omega}D_k^*(\ell)\right\},\\ D_k^*(\omega) &= \tilde{D}_k(\omega-1)+D_k(1,\omega)-D_k(1,\omega-1), \quad n\geqslant m+2,\\ D_{n+1}(1,\omega) &= D_n(1,\omega)+\sum_{k=m}^{n-1}D_k^*(\omega),\\ \tilde{D}_n(0) &= 1, \qquad \tilde{D}_n(\omega)=0 \quad \text{for } \omega\geqslant 1, \quad n\leqslant m+1. \end{aligned} \tag{24}$$

For the number of structures with a degree at most one we find

$$\begin{aligned} D'_{n+1} &= D'_n+\sum_{k=m}^{n-1}D_k^*(1)D'_{n-k-1},\\ D^*_{n+1}(1) &= \sum_{k=m}^{n}D_k^*(1). \end{aligned} \tag{25}$$

### 3.4. Secondary structures with minimum stack length

Let $\Psi_n(l)$ be the number of structures with minimal stack length $l$, and let $\Psi_n^*(l)$ be the number of structures on $n$ digits which have only stacks of length at least $l$ if an additional terminal base pair is attached. Furthermore, let $\Psi_n^{**}(l)$ be the number of structures on $n$ digits with all stacks of length at least $l$ for which $(1,n)$ is not a base pair.

These three numbers fulfill for $l > 1$ the coupled recursions

$$\begin{aligned}
&\Psi_{n+1}(l) = \Psi_n(l) + \sum_{k=m+2l-2}^{n-1} \Psi_k^*(l)\Psi_{n-k-1}(l),\\
&\Psi_n^*(l) = \sum_{p=l-1}^{(n-m)/2} \Psi_{n-2p}^{**}(l),\\
&\Psi_n^{**}(l) = \Psi_n(l) - \Psi_{n-2}^*(l),\\
&\Psi_n(l) = \Psi_{n+1}^{**}(l) = 1 \quad n < m + 2l,\\
&\Psi_n^*(l) = 0, \quad m + 2l - 2.
\end{aligned} \tag{26}$$

The first recursion is obvious. A structure which has only stacks of length at least $l$ after addition of the terminal base pair must have a terminal stack of length $p \geqslant l-1$. The remaining part of the structure must have stacks of length at least $l$ without a terminal base pair. Of course, there is no such structure if $n - 2p < m$. For the numbers $\Psi_n^{**}(l)$ we obtain the explicit recursion:

$$\Psi_{n+1}^{**}(l) = \Psi_n(l) + \sum_{k=m+2l-2}^{n-2} \Psi_k^*(l)\Psi_{n-k-1}(l), \tag{27}$$

$$\Psi_n^{**} = 1 \quad n < m + 2l,$$

because structures without a terminal base pair and stacks of length at least $l$ are obtained by adding a new base pair to structures which including this base pair have stacks of sufficient length (first factor in the sum) provided the structures in the remaining part of the structure have also sufficient stack length. Of course, there may not by a terminal base pair by construction. Comparing the sum in (27) and in the recursion for $\Psi_n(l)$ yields the final result. We have of course $\Psi_n(1) = S_n$ for all $n$ and $\Psi_n(l+1) < \Psi_n(l)$ for all $l$ and sufficiently large $n$.

**Remark.** It is possible, of course, to obtain recursions of the above type for the number of structure elements or the number of structures with particular properties also for $l > 1$. If $\Xi_n$ is the counting series of interest one has to replace $S_k\Xi_{n-k-1}$ by $\Psi_k^*\Xi_{n-k-1}$ and $\Xi_k S_{n-k-1}$ by $\Xi_k^*\Psi_{n-k-1}$, where $\Xi^*$ counts the objects of interest subject to the restriction that the secondary structure has a terminal stack of length at least $l$.

## 4. Asymptotics

The symbol $\sim$ has its usual meaning:
$f(n) \sim g(n)$ means $f(n)/g(n) \to 1$ as $n \to \infty$.
If not explicitly stated, asymptotic formulae assume $n \to \infty$.

### 4.1. Asymptotics from generating functions

Most of the published work on the asymptotic behavior of RNA-related counting series makes use of a proposition by E.A. Bender [1, Theorem 5], which was found to be true only under more restrictive conditions than the published ones. It follows from the counterexamples discussed in [2, 18] that Bender's result cannot be applied directly to the RNA problem. Nevertheless, the published expressions for the RNA counting series are correct, as we shall show below. We start from a simplified version of Darboux' theorem [4], see also [29, p. 205].

**Theorem 4.1.** *Suppose $y_n \geqslant 0$ and $y(x) = \sum_{n=0}^{\infty} y_n x^n$ is of the form*

$$y(x) = \beta(x) + g(x)\left(1 - \frac{x}{\alpha}\right)^{\omega}, \tag{28}$$

*where $\alpha > 0$ is real, $\beta(x)$ and $g(x)$ are analytic near $\alpha$, and $\omega$ is real but not a non-negative integer. If $y(x)$ is analytic for $|x| < \alpha$ and $x = \alpha$ is the only singularity of $y$ on its circle of convergence, then*

$$y_n \sim \frac{g(\alpha)}{\Gamma(-\omega)} n^{-1-\omega} \left(\frac{1}{\alpha}\right)^n. \tag{29}$$

**Corollary 4.2.** *Let $\Phi(x, y)$ be a polynomial in $y$ and analytic in $x$ for $|x| < \alpha + \delta$, $\delta > 0$. Suppose $y$ fulfills the conditions of Theorem 4.1 with*

$$y(x) = \beta(x) + \left(1 - \frac{x}{\alpha}\right)^{1/2} g(x). \tag{30}$$

Let the generating function $z(x) = \sum_{n=0}^{\infty} z_n x^n$ be of the form $z = \Phi(x, y)$. Then

$$\lim_{n \to \infty} \frac{z_n}{y_n} = \Phi_y(\alpha, \beta(\alpha)). \tag{31}$$

**Proof.** In the following, we will use the short hand $\beta$ for $\beta(\alpha)$. Expanding $\Phi(x, y)$ around $y = \beta(\alpha)$ one obtains

$$\begin{aligned} \Phi(x, y(x)) &= \Phi(x, \beta(x)) + \Phi_y(x, \beta(x))(y - \beta(x)) + \mathrm{O}((y - \beta(x))^2) \\ &= \Phi(x, \beta(x)) + \Phi_y(x, \beta(x)) g(x)(1 - x/\alpha)^{1/2} + \mathrm{O}((y - \beta(x))^2) \end{aligned} \tag{32}$$

where the $\mathrm{O}((y - \beta(x))^2)$ term does not introduce additional singularities. Darboux' theorem therefore applies and yields

$$z_n \sim \frac{g(\alpha)\Phi_y(\alpha, \beta)}{\Gamma(-\frac{1}{2})} n^{-3/2} \left(\frac{1}{\alpha}\right)^n. \quad \square \tag{33}$$

**Corollary 4.3.** *Let $\Phi(x, y)$ and $y(x)$ have the same properties as in the previous corollary. Assume the coefficients $y_n$ are nonnegative and positive for sufficiently* *large n. Let $z(x)=\sum_{n=0}^{\infty} z_n x^n$ be a generating function of the form*

$$z(x)=\frac{1}{\alpha\beta-xy}\Phi(x,y), \tag{34}$$

*where $\beta=\beta(\alpha)$. Then*

$$\frac{z_k}{y_k}\sim\frac{2\Phi(\alpha,\beta)}{\alpha g^2(\alpha)}n \tag{35}$$

**Proof.** First note that $\alpha\beta-xy$ can be written in the form $\varphi(x)(1-x/\alpha)-xg(x)(1-x/\alpha)^{1/2}$, where $\varphi(x)$ is analytic near $\alpha$. Therefore,

$$\frac{1}{\alpha\beta-xy}=\frac{\varphi(x)}{\varphi(x)^2(1-x/\alpha)-x^2g(x)^2}+\frac{xg(x)}{\varphi(x)^2(1-x/\alpha)-x^2g(x)^2}\left(1-\frac{x}{\alpha}\right)^{-1/2}. \tag{36}$$

Since the $y_n$ are positive $y(|x|)\leqslant y(\alpha)=\beta$ for $|x|\leqslant\alpha$, with equality only for $x=\alpha$. Hence there are no additional zeros of $\alpha\beta-xy$ and $z$ is analytic for $|x|\leqslant\alpha$ with $x=\alpha$ the only singularity on the circle of convergence. Eq. (36), therefore, fulfills the requirements of Theorem 4.1. Multiplying Eq. (36) with $\Phi(x,y)$ and applying Darboux' theorem yields

$$z_k\sim\frac{-\Phi(\alpha,\beta)}{\alpha g(\alpha)\Gamma(1/2)}n^{-1/2}\alpha^{-n}. \tag{37}$$

Using $\Gamma(\frac{1}{2})=-\frac{1}{2}\Gamma(-\frac{1}{2})$ completes the proof. □

**Corollary 4.4.** *Let $y$ as in the previous corollary and let $u,v$ be of the same form as $z$ above. Suppose there is an analytic function $\Phi(x,y)$ such that $u=\Phi(x,y)v$. Then*

$$\lim_{n\to\infty}\frac{u_n}{v_n}=\Phi(\alpha,\beta). \tag{38}$$

**Proof.** Assuming that both $u$ and $v$ are of the form (34) we find from equation (35) that $u_n/v_n=\Phi^u(\alpha,\beta)/\Phi^v(\alpha,\beta)$. The conditions of Corollary 4.3 ensure that this quotient exists and $\Phi=\Phi^u/\Phi^v$. □

### 4.2. The number of secondary structures

The series $S_n$ has been extensively studied in [34]. Consider the series $\Psi_n$ of secondary structures with a prescribed minimum stack length $l$ and minimum size $m$ for hairpin loops. Denote by

$$\psi(x)=\sum_{n=0}^{\infty}\Psi_n x^n,\qquad \phi(x)=\sum_{n=0}^{\infty}\Psi_n^* x^n,\qquad \theta(x)=\sum_{n=0}^{\infty}\Psi_n^{**}x^n \tag{39}$$

the generating functions. We shall use the notation

$$t_m(x)=\sum_{k=0}^{m-1} x^k, \qquad \tau_m(x)=\sum_{k=1}^{m-1} kx^k = x\frac{\mathrm{d}}{\mathrm{d}t}t_m(x) \tag{40}$$

**Theorem 4.5.** *The generating function* $\psi, \phi$ *and* $\theta$ *fulfill the coupled functional equations*

$$\begin{aligned}
\psi &= 1 + x\psi + x^2\phi\psi,\\
\phi &= \frac{x^{2(l-1)}}{1-x^2}(\theta - t_m(x)),\\
\theta &= \psi - x^2\phi.
\end{aligned} \tag{41}$$

**Proof.** The first and third line are obvious. The second line is obtained from

$$\begin{aligned}
\phi &= \sum_{n=0}^{\infty} x^n \sum_{p=l-1}^{(n-m)/2} \Psi^{**}_{n-2p}\\
&= \sum_{n=0}^{\infty}\sum_{p=0}^{n/2} x^2 p\Psi^{**}_{n-2p}x^{n-2p} - \sum_{p=0}^{l-2} x^{2p}\sum_{n=0}^{\infty}\Psi^{**}_{n-2p}x^{n-2p}\\
&\quad - \sum_{p>\frac{n-m}{2}}^{n/2} x^{2p}\sum_{n=0}^{\infty}\Psi^{**}_{n-2p}x^{n-2p} + \sum_{p>\frac{n-m}{2}}^{l-2} x^{2p}\sum_{n=0}^{\infty}\Psi^{**}_{n-2p}x^{n-2p}\\
&= \frac{1}{1-x^2}\theta - \sum_{p=0}^{l-2} x^{2p}\theta - t_m(x) + \sum_{p=0}^{l-2} x^{2p} t_m(x)\\
&= \frac{x^{2(l-1)}}{1-x^2}(\theta - t_m(x)). \qquad \square
\end{aligned} \tag{42}$$

**Corollary 4.6.** *The generating function* $\psi$ *is analytic in a neighbourhood of* 0 *and fulfills*

$$\begin{aligned}
x^{2l}\psi(x) &= [(1-x)(1-x^2+x^{2l}) + x^{2l}t_m(x)]\\
&\quad - \sqrt{[(1-x)(1-x^2+x^{2l}) + x^{2l}t_m(x)]^2 - 4x^{2l}(1-x^2+x^{2l})}.
\end{aligned} \tag{43}$$

**Proof.** From (41) we obtain a quadratic equation for $\psi$, the correct sign of the solution follows from $S_0 = \psi(0) = 1$. Taylor expansion shows that $\psi$ has an analytic continuation at the origin. □

The same generating function has recently been derived by Régnier and Tahi [22, 31].

**Corollary 4.7.** *For $l=1$ we recover the generating function $s(x)=\sum_{n=0}^{\infty} S_k x^k$ for the number of secondary structures. It fulfills the functional equation*

$$F(x,y)=1+\left(2x-\sum_{k=0}^{m+1} x^k\right)y+x^2y^2=0. \tag{44}$$

**Theorem 4.8.**

$$\Psi_n \sim \frac{-g(\alpha)}{2\sqrt{\pi}} n^{-3/2}\left(\frac{1}{\alpha}\right)^n, \tag{45}$$

*where $\alpha$ is the smallest positive solution of*

$$p(x)=[(1-x)(1-x^2+x^{2l})+x^{2l}t_m(x)]^2-4x^{2l}(1-x^2+x^{2l})=0 \tag{46}$$

*that satisfies*

$$g(\alpha)=\frac{-1}{x^{2l}}\sqrt{-\frac{1}{\alpha}\frac{\mathrm{d}p(x)}{\mathrm{d}x}\bigg|_{\alpha}} \neq 0. \tag{47}$$

**Proof.** From Eq. (43) it is clear that the singularities of $\psi(x)$ are branch points which occur when Eq. (46) is fulfilled. With $\alpha$ as given in Eq. (46) $\psi$ can be written in the form required by Theorem 4.1:

$$\psi(x)=\frac{p_1(x)}{x^{2l}}-\frac{\sqrt{p_2(x)}}{x^{2l}}\left(1-\frac{x}{\alpha}\right)^{1/2}, \tag{48}$$

where $p_1(x)$ and $p_2(x)$ are polynomials and $p_2(\alpha)$ can be obtained by differentiation of $p(x)$ to yield Eq. (47). It remains to be shown that $\psi$ can have no other singularity for $x \leqslant \alpha$. Recall that there is no singularity at 0 despite the form of Eq. (48), see Corollary 4.6.

Suppose $u \neq \alpha$, $|u| \leqslant \alpha$ is another singularity, i.e., another solution of (46), and let $v=\psi(u)$. Consider the function

$$\varphi(\psi,x)=(\psi^2-1)x^{2l}+x^2, \tag{49}$$

and let

$$\beta=\psi(\alpha)=\frac{1}{\alpha^l}\sqrt{1-\alpha^2+\alpha^{2l}}. \tag{50}$$

By comparison with Eqs. (43) and (46) we have $\varphi(v,u)=\varphi(\beta,\alpha)=1$. The coefficients of the power series for $(\psi(x)^2-1)$ are strictly positive, except the first one which is 0. Therefore, $|\psi(x)^2-1| \leqslant \psi(|x|)^2-1 \leqslant \beta^2-1$ with equality only for $x=\alpha$. Furthermore,

Table 1
Coefficients for the asymptotics of $\Psi_n$

| $l$ | $m=0$ | 1 | 2 | 3 | 5 | $\infty$ |
|---|---|---|---|---|---|---|
| | $\alpha$ | | | | | |
| 1 | 0.3333 | 0.3820 | 0.4142 | 0.4369 | 0.4658 | 0.5000 |
| 2 | 0.4836 | 0.5081 | 0.5266 | 0.5409 | 0.5610 | 0.5958 |
| 3 | 0.5672 | 0.5828 | 0.5952 | 0.6053 | 0.6204 | 0.6537 |
| 4 | 0.6227 | 0.6336 | 0.6428 | 0.6504 | 0.6623 | 0.6938 |
| 5 | 0.6629 | 0.6712 | 0.6783 | 0.6843 | 0.6941 | 0.7237 |
| 10 | 0.7704 | 0.7737 | 0.7766 | 0.7793 | 0.7840 | 0.8066 |
| 20 | 0.8713 | 0.8518 | 0.8530 | 0.8540 | 0.8559 | 0.8713 |
| 100 | 0.9520 | 0.9521 | 0.9522 | 0.9523 | 0.9525 | 0.9571 |
| $\infty$ | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | $-g(\alpha)/(2\sqrt{\pi})$ | | | | | |
| 1 | 1.4658 | 1.1044 | 0.8766 | 0.7131 | 0.4880 | 0.0000 |
| 2 | 2.7155 | 2.1614 | 1.7742 | 1.4848 | 1.0769 | 0.0000 |
| 3 | 3.9640 | 3.2711 | 2.7558 | 2.3561 | 1.7741 | 0.0000 |
| 4 | 5.2305 | 4.4238 | 3.7990 | 3.3003 | 2.5537 | 0.0000 |
| 5 | 6.5194 | 5.6142 | 4.8923 | 4.3033 | 3.4009 | 0.0000 |
| 10 | 13.309 | 12.026 | 10.921 | 9.962 | 8.3820 | 0.0000 |
| 20 | 28.365 | 26.557 | 24.913 | 23.414 | 20.787 | 0.0000 |
| 100 | 189.31 | 185.30 | 181.41 | 177.63 | 170.40 | 0.0000 |

we have

$$|\varphi(v,u)| \leqslant |v^2-1|\,|u^{2l}| + |u^2| \leqslant (\beta^2-1)\alpha^{2l} + \alpha^2 = 1, \tag{51}$$

which together with $\varphi(v,u)=1$ can only be fulfilled for $u=\alpha$. □

**Corollary 4.12.** *For $l=1$ the above equations simplify to $\beta = 1/\alpha$ and $\alpha$ is the smallest positive solution of*

$$\sum_{k=0}^{m+1} \alpha^k - 4\alpha = 0. \tag{52}$$

*We therefore recover the results from Ref.* [28]. *Numerical values are given in Table* 1.

*Throughout the remainder of this paper we will assume $l=1$ if $l$ is not mentioned explicitly, while $\alpha$ and $\beta$ will denote the solutions of Eqs.* (46) *and* (50), *respectively.*

### *4.3. Average number of structure elements*

Denote by $\Xi_n$ the number of structural elements. From the biological point of view it is very interesting to determine the average number of structural elements in a single structure, i.e. the asymptotic behavior of $\Xi_n/S_n$. It is clear that the counting series for the total number of structure elements, including the total number of base pairs and unpaired digits is bounded from above by $nS_n$.

**Lemma 4.9.** *Let $\alpha$ be the smallest positive solution of Eq.* (52). *Then*

$$\begin{aligned} t_m(\alpha) &= \frac{3\alpha-1}{\alpha^2}, \qquad \tau_m(\alpha) = \frac{3\alpha-1}{\alpha(1-\alpha)} - m\frac{(1-2\alpha)^2}{\alpha^2(1-\alpha)}, \\ g^2(\alpha) &= \frac{(1-2\alpha)(2+m-2m\alpha)}{(1-\alpha)\alpha^3}. \end{aligned} \tag{53}$$

**Theorem 4.10.** *The number of components, $I_n$, fulfills*

$$\lim_{n\to\infty} \frac{I_n}{S_n} = 2\beta(1-\alpha) - 1 = 2/\alpha - 3. \tag{54}$$

**Proof.** Let $i(x) = \sum_{k=0}^{\infty} I_k x^k$ be the generating function for the number of components. The recursion can be brought to the form

$$I_{n+1} = I_n + \sum_{k=0}^{n-1} S_k I_{n-k-1} + \sum_{k=0}^{n-1} S_k S_{n-k-1} - \sum_{k=0}^{m-1} [I_{n-k-1} + S_{n-k-1}]. \tag{55}$$

Multiplying by $x^{n+1}$ and summing over $n$ yields

$$i(x) = xi(x) + x^2 s(x) i(x) + x^2 s^2(x) - x^2 t_m(x)[s(x) + i(x)]. \tag{56}$$

Using twice the functional equation for $s(x)$ we find

$$\begin{aligned} i(x) &= \frac{x^2 s^2(x) - s(x) x^2 t_m(x)}{1 - x - x^2 s(x) + x^2 t_m(x)} = s(x) x^2 s(x)[s(x) - t_m(x)]. \\ &= s^2(x)(1-x) - s(x). \end{aligned} \tag{57}$$

Application of Corollary 4.2 immediately yields the desired result. □

The first equality in Eq. (54) holds for arbitrary minimal stack length $l$, too.

**Theorem 4.11.** *The number of external digits, $E_n$, fulfills*

$$\lim_{n\to\infty} \frac{E_n}{S_n} = 2\alpha\beta = 2. \tag{58}$$

**Proof.** The functional equation for the generating function reads $e(x) = x \cdot s^2(x)$. Corollary 4.2 completes the proof. □

**Theorem 4.12.** *The number of unpaired digits, $U_n$, fulfills*

$$\frac{U_n}{S_n} \sim \frac{2\alpha + m(1-2\alpha)}{2 + m(1-2\alpha)} n. \tag{59}$$

**Proof.** Let $u(x) = \sum_{n=0}^{\infty} U_n x^n$ be the generating function of the number of unpaired digits. From recursion (9) we find immediately the functional equation

$$u = xu + xs + 2x^2 us - x^2 u t_m(x) - x^2 s \tau_m(x). \tag{60}$$

Using the functional equation for $s$, some computations yield

$$u(x)=\frac{1}{1-x^2s^2}\,s^2x(1-x\tau_m). \tag{61}$$

Application of Corollary 4.3 completes the proof. □

Let $p(x)$ be the generating function for the number of base pairs. Since $U_n + 2P_n = nS_n$ we have $u(x)+2p(x)=xs'(x)$.

**Theorem 4.13.** *The number of stacks or loops, $N_n$, fulfills*

$$\frac{N_n}{S_n}\sim\frac{(1-\alpha)^2(1+\alpha)}{2+m-2m\alpha}\,n. \tag{62}$$

**Proof.** Let $v(x)=\sum_{n=0}^{\infty}N_nx^n$ be the generating function of the number of stacks. Observe that

$$\sum_{k=m+p}^{n-1}S_{k-p}S_{n-k-1}=\sum_{k=m}^{n-p-1}S_kS_{n-p-k-1} \tag{63}$$

and, therefore, gives rise to a term $x^{p+2}[(s^2-st_m(x)]=x^p[(1-x)s-1]$ in the functional equation for the generating function. Thus, recursion (12) translates to

$$v=xv+2x^2sv-x^2vt_m(x)+(1-x^2)[s(1-x)-1] \tag{64}$$

or, after some simple rearrangements,

$$v=\frac{1}{1-x^2s^2}s(1-x^2)[s(1-x)-1]. \tag{65}$$

The proof is completed by Corollary 4.3. □

### 4.4. The number of structures with certain properties

**Theorem 4.14.** *The number of secondary structures with $b$ base pairs is*

$$H_n(b)\sim\frac{1}{(b+1)!b!}n^{2b}. \tag{66}$$

**Proof.** From recursion (4) we obtain the functional equation

$$\begin{aligned} h_b &= xh_b+x^2\sum_{k=0}^{b-1}h_{b-k-1}h_k-x^2t_m(x)h_{b-1}, \quad b>0\\ &= xh_b+x^2\sum_{k=1}^{b}h_kh_{b-k-1}+x^{m+2}h_{b-1} \end{aligned} \tag{67}$$

and $h_0(x)=1/(1-x)$. With the ansatz

$$h_b(x)=\eta_b(x)\,\frac{1}{1-x}\left(\frac{x}{1-x}\right)^{2b}, \tag{68}$$

we find that the functions $\eta_b(x)$ must be polynomials fulfilling

$$\eta_b(x)=\sum_{k=1}^{b-1}\eta_k(x)\eta_{b-k-1}(x)+x^m\eta_{b-1}(x), \quad \eta_0(x)=1. \tag{69}$$

Theorem 4.1 assures now that

$$H_n(b)\sim\frac{\eta_b(1)}{\Gamma(2b+1)}\,n^{2b}. \tag{70}$$

Since $\eta_0(1)=1$, Eq. (69) reduces to the well known recursion for the Catalan numbers

$$\eta_b(1)=C_b=\frac{1}{b+1}\binom{2b}{b}. \quad \Box \tag{71}$$

**Theorem 4.15.** *The number of structures with exactly b stacks is*

$$N_n(b)\sim\frac{C_b}{2^b(3b)!}\,n^{3b}. \tag{72}$$

**Proof.** Let $\nu_b(x)=\sum_{n=0}^{\infty}N_n(b)x^n$ be the generating function for the number of structures with exactly $b$ stacks and denote by $\zeta_b(x)$ the generating function for the auxiliary variable $Z_n(b)$. It is straightforward to derive the functional equations

$$\begin{aligned}\zeta_b&=\frac{x^2}{(1-x)(1+x)}\,[\nu_{n-1}-\eta_{b-1}],\\ \nu_b&=\frac{x^2}{(1-x)}\sum_{l=1}^{b}\zeta_l-\nu_{b-l}.\end{aligned} \tag{73}$$

One easily verifies that these generating functions are of the form

$$\begin{aligned}\nu_b(x)&=\mu_b(x)\frac{1}{(x+1)^b}\frac{1}{(x-1)^{3b+1}},\\ \zeta_b(x)&=\xi_b(x)\frac{1}{(x+1)^b}\frac{1}{(x-1)^{3b+1}},\end{aligned} \tag{74}$$

where $\mu_b(x)$ and $\xi_b(x)$ are polynomials. We cannot use the simplified version of Darboux' Theorem 4.1 in this case since there are two singularities on the circle of convergence. Expanding by partial fractions we have the identity

$$\frac{1}{(x+1)^b}\frac{1}{(x-1)^{3b+1}}=\frac{A(x)}{(x+1)^b}+\frac{B(x)}{(x-1)^{3b+1}}, \tag{75}$$

where $A(x)$ and $B(x)$ are polynomials of degree $3b$ and $b-1$, respectively, satisfying $B(x)(x+1)^b+A(x)(x-1)^{3b+1}=1$ and, hence, $B(1)=2^{-b}$ and $A(-1)=(-2)^{-3b-1}$.

A more general version of Darboux' theorem, for instance [20, Theorem 11.7], now shows that

$$N_n(b) \sim \frac{1}{2^b}\,\frac{\mu_b(1)}{\Gamma(3b+1)}\,n^{3b} + \frac{1}{(-2)^{3b+1}}\,\frac{\mu_b(-1)}{\Gamma(b)}\,n^{b-1}(-1)^n. \tag{76}$$

Clearly, the second term is $o(n^{3b})$ and hence does not contribute to the asymptotic behavior. The coefficients $\mu_b(1)$ and $\xi_b(1)$ satisfy the recursions

$$\xi_b(1) = \mu_{b-1}(1), \qquad \mu_n(1) = \sum_{l=1}^{b} \xi_l(1)\mu_{b-l}(1) = \sum_{l=0}^{b-1} \mu_l(1)\mu_{b-l-1}(1). \tag{77}$$

Again, the coefficients $\mu_b(1)$ are the Catalan numbers. □

**Theorem 4.16.** *The number of structures with b hairpins fulfills*

$$A_n(b) \sim \frac{4}{2^{(3+m)b}b!(b-1)!}n^{2(b-1)}2^n. \tag{78}$$

**Proof.** Let $a_b(x) = \sum A_n(b)x^n$ denote the generating function. From recursion (8) we obtain after some simple rearrangements

$$a_b = xa_b + x^2\sum_{i=1}^{b} a_i a_{b-i} + x^2 t_m a_{b-1}, \quad b>0 \tag{79}$$

and $a_0(x) = 1/(1-x)$. Collecting all terms containing $a_b(x)$ yields

$$(1-2x)a_b = x^{m+2}a_{b-1} + x^2\sum_{i=1}^{b-1} a_i a_{b-i}. \tag{80}$$

With the ansatz

$$a_b(x) = \left(\frac{x^{m+2}}{1-x}\right)^b \frac{1}{(1-2x)^{2b-1}}\eta_b(x), \tag{81}$$

we find the following recursion for the polynomials $\eta_b(x)$:

$$\eta_b(x) = (1-2x)(1-x)\eta_{b-1} + x^2\sum_{i=1}^{b-1} \eta_i(x)\eta_{b-1}, \quad \eta_1(x) = 1. \tag{82}$$

Theorem 4.1 now implies that the relevant singularity occurs at $x = \frac{1}{2}$ leaving us with the recursion

$$\eta_b(\tfrac{1}{2}) = \frac{1}{4}\sum_{i=1}^{b-1} \eta_i(\tfrac{1}{2})\eta_{b-i}(\tfrac{1}{2}). \tag{83}$$

It is easy to verify that recursion (82) is solved by

$$\eta_b(\tfrac{1}{2}) = \frac{1}{2^{2(b-1)}}C_{b-1}. \tag{84}$$

From Theorem 4.1 we find now that

$$A_n(b) \sim \frac{C_{b-1}}{2^{2(b-1)}2^{b(m+1)}\Gamma(2b+1)} n^{2(b-1)}2^n. \tag{85}$$

A simple calculation completes the proof. □

**Theorem 4.17.** *The number of structures with b components, $J_n(b)$, fulfills*

$$\lim_{n\to\infty} J_n(b)/S_n = \frac{\alpha^2}{(1-\alpha)^3}\, b\left(\frac{1-2\alpha}{1-\alpha}\right)^{b-1}. \tag{86}$$

**Proof.** Let $j_b(x) = \sum_{n=0}^{\infty} J_n(b)x^n$ be the generating function for the number of secondary structures with exactly $b$ components. It is straightforward to derive

$$j_b(x) = \left[\frac{x^2}{1-x}(s - t_m(x))\right]^b j_0(x), \quad b \geqslant 1 \tag{87}$$

and from $J_n(0) = 1$ we obtain $j_0(x) = 1/(1-x)$. From Corollary 4.2 we find that

$$\lim_{n\to\infty} J_n(b)/S_n = \frac{1}{1-\alpha}\left(\frac{\alpha^2}{1-\alpha}\right)^b b(\beta - t_m(\alpha))^{b-1}. \quad □ \tag{88}$$

**Theorem 4.18.** *The number of structures with b external digits, $E_n(b)$, fulfills*

$$\lim_{n\to\infty} E_n(b)/S_n = \tfrac{1}{4}(b+1)\left(\tfrac{1}{2}\right)^b. \tag{89}$$

**Proof.** Let $e_b(x)$ be the generating function of the number of secondary structures with exactly $b$ external digits. Recursion (5) yields the functional equation

$$e_b - \delta_{0b} = xe_{b-1} + x^2se_b - x^2e_bt_m(x). \tag{90}$$

Substituting the functional equation for $s$ and some algebra finally yields $e_0 = s/(1+xs)$ and $e_b = [xs/(1+xs)]e_{b-1}$. Therefore,

$$e_b = \left(\frac{xs}{1+xs}\right)^b \frac{s}{1+xs}. \tag{91}$$

Corollary 4.2 and observing $\alpha\beta = 1$ yields the desired expression. □

**Theorem 4.19.** *For any finite order $\omega$ there is a positive constant $\varepsilon_\omega$ such that*

$$\lim_{n\to\infty} \frac{\tilde{D}_n(\omega-1)e^{n\varepsilon_\omega}}{\tilde{D}_n(\omega)} = 0. \tag{92}$$

**Proof.** We will need the generating functions

$$\Delta_\omega = \sum_{n=0}^{\infty} \tilde{D}_n(\omega)x^n, \qquad \Delta_\omega^* = \sum_{n=0}^{\infty} D_n^*(\omega)x^n, \qquad \Delta_\omega' = \sum_{n=0}^{\infty} D_n(1,\omega)x^n. \tag{93}$$

Recursion (24) yields the following system of coupled functional equations for the above generating functions:

$$\begin{aligned} \Delta_\omega &= x\Delta_\omega + x^2\Delta_\omega^*\sum_{i=0}^{\omega-1}\Delta_i + x^2\Delta_\omega\sum_{i=0}^{\omega}\Delta_i^*,\\ \Delta_\omega^* &= \Delta_{\omega-1} + \Delta'_\omega - \Delta'_{\omega-1}, \quad \omega\geqslant 2, \\ \Delta'_\omega &= x\Delta'_\omega + x^2\Delta_\omega^*\frac{1}{1-x}. \end{aligned} \tag{94}$$

For $\omega=0$ we have $\Delta_0=1/(1-x)$ and for $\omega=1$, we find explicitly

$$\begin{aligned} \Delta_1^*(x) &= \frac{1-x}{1-2x}x^m,\\ \Delta_1(x) &= \frac{x^{m+2}}{1-x}\,\frac{1}{1-2x-x^{m+2}}. \end{aligned} \tag{95}$$

Eliminating $\Delta'_\omega$ we find for $\omega\geqslant 2$

$$\begin{aligned} \Delta_\omega^* &= \frac{(1-x)^2}{1-2x}\Delta_{\omega-1} - \frac{x^2}{1-2x}\Delta_{\omega-1}^*,\\ \Delta_\omega &= \frac{x^2\Delta_\omega^*\sum_{i=0}^{\omega-1}\Delta_i}{1-x-x^2\sum_{i=0}^{\omega}\Delta_i^*}. \end{aligned} \tag{96}$$

Unfortunately these expressions become to involved to be of much practical use. Denote $f_\omega(x)=1-x-x^2\sum_{i=0}^{\omega}\Delta_i^*$ and let $\lambda$ be the unique solution of $1-2x-x^{m+2}$ in the interval $[0,\frac{1}{2}]$. Obviously, $f_\omega(x)$ is strictly decreasing and has at least one zero in $(0,\alpha^*)$, where $\alpha^*$ denotes the position of the singularity with the smallest $x$ value among the function $\Delta_i(x)$, $i<\omega$. Therefore, $\Delta_\omega(x)$ has a singularity $\alpha_\omega<\alpha^*$. By induction, therefore, $\alpha_\omega<\alpha_{\omega-1}$ for all $\omega$, since explicitly we have $\alpha_1=\lambda$ and the first singularity in $\Delta_\omega^*$ occurs at $x=\alpha_{\omega-1}$. By Theorem 4.1 we have $\Delta_n(\omega)\sim c_1 n^{c_2}\alpha_\omega^n$. The inequality $1/\alpha_\omega>1/\alpha_{\omega-1}$ completes the proof. □

Numerical estimates for the constants $\alpha_\omega$ have been obtained by explicitly calculating $\Delta_\omega(x)$ with the help of `Mathematica` and by solving numerically for the smallest zero of the denominator in (96,2). The results are compiled in Table 2. The case $m=1$, $\omega=1$ has already been treated by Waterman [34, 36], the generating function for $m=1$ has been derived in [33].

### 4.5. *The distribution of structure elements*

**Theorem 4.20.** *The number of loops with b unpaired digits, $Q_n(b)$, fulfills*

$$\lim_{n\to\infty}\frac{Q_n(b)}{N_n} = \frac{\alpha^2}{(1-\alpha^2)(1-2\alpha)}\left[\frac{1}{2\alpha 2^b} - \Theta(m-b)\alpha^b - (1-2\alpha)\delta_{b0}\right]. \tag{97}$$

Table 2
Secondary structures with order $\omega$. The base of the exponential part of the asymptotic is given

| | $\alpha_\omega$ | | |
|---|---|---|---|
| $\omega$ | $m=0$ | $m=1$ | $m=3$ |
| 0 | 1 | 1 | 1 |
| 1 | 0.41421256 | 0.4533977 | 0.4863890 |
| 2 | 0.37597060 | 0.4221456 | 0.4680050 |
| 3 | 0.35978154 | 0.4076474 | 0.4577424 |

**Proof.** Let $q_b(x)=\sum_{k=0}^{\infty}Q_n(b)x^n$ denote the generating function for the number of loops with $b$ unpaired digits. From recursion (13) we find immediately

$$\begin{aligned} q_b &= xq_b+2x^2sq_b+x^2se_b-x^2q_bt_m-\Theta(m-b)x^b,\\ q_0 &= xq_0+2x^2sq_0+x^2se_0-x^2q_bt_m-\Theta(m)-x^2[s(1-x)-1], \end{aligned} \tag{98}$$

where $\Theta(n)$ denotes the Heavyside function, $\Theta(n)=1$ for $n>0$ and $\Theta(n)=0$ for $n\leqslant 0$. A simple calculation confirms

$$\begin{aligned} q_b &= \frac{1}{1-x^2s^2}x^2s^2[e_b-\Theta(m-b)x^b], \quad b>0,\\ q_0 &= \frac{1}{1-x^2s^2}x^2s[se_b-s(1-x)+1-\Theta(m-0)]. \end{aligned} \tag{99}$$

The substitution of $e_b$ from Eq. (91) and Corollary 4.3 prove the assertion. □

**Theorem 4.21.** *The asymptotic distribution of stack lengths is geometric*:

$$\lim_{n\to\infty}\frac{W_n(b)}{N_n}=\frac{1-\alpha^2}{\alpha^2}\alpha^{2b}. \tag{100}$$

**Proof.** Let $w_b(x)=\sum_{k=0}^{\infty}W_n(b)x^n$ denote the generating function for the number of stacks of length $b$. From recursion (14) we find

$$w_{b+1}(x)=x^2w_b(x), \quad b>1. \tag{101}$$

Using $v(x)=\sum_b w_b(x)$ determines $w_1(x)$ and yields

$$w_b=x^{2b-2}(1-x^2)\,v(x). \tag{102}$$

Corollary 4.4 completes the proof. □

### 4.6. *Loop types*

**Theorem 4.22.** *The distribution of loop degrees fulfills*

$$\lim_{n\to\infty}\frac{L_n(d)}{N_n}=\frac{\alpha^2}{(1-\alpha^2)(1-2\alpha)}\times\left[\frac{1}{1-2\alpha}\left(\frac{1-2\alpha}{1-\alpha}\right)^d-\begin{cases}\frac{3\alpha-1}{\alpha^2}, & d=1\\ (1-2\alpha), & d=2\\ 0, & d>2\end{cases}\right]. \tag{103}$$

**Proof.** Let $\ell_d(x)=\sum_{n=0}^{\infty}L_n(d)x^n$ be the generating function for the number of loops with degree $d$. For hairpins one finds from recursion (19)

$$\ell_1=x\ell_1+2x^2\ell_1 s-x^2\ell_1 t_m(x)+\frac{x^{m+2}}{1-x}s. \tag{104}$$

Similar functional equations can be obtained for loops of higher degree from recursions (19) and (20). They can be brought to the form

$$\begin{aligned}\ell_1&=\frac{1}{1-x^2s^2}\frac{x^{m+2}}{1-x}s^2,\\ \ell_2&=\frac{1}{1-x^2s^2}[x^2s^2[j_1(x)-(1-x)]+x^2s],\\ \ell_d&=\frac{1}{1-x^2s^2}x^2s^2j_{d-1}(x),\quad d>2.\end{aligned} \tag{105}$$

Using the explicit expressions for $j_d$ and Corollary 4.3, some tedious algebra finally yields Eq. (103). □

The average loop degree $\bar{d}$ can be most easily calculated from the balance equation

$$\sum_{\text{loops }\lambda}\deg(\lambda)=2\#[\text{stacks}]-\#[\text{components}], \tag{106}$$

which holds for all secondary structures. From Eqs. (54) and (62), we find immediately that the average loop degree fulfills

$$\lim_{n\to\infty}\bar{d}_n=2. \tag{107}$$

**Theorem 4.23.** *The ratio of bulges and true interior loops fulfills*

$$\lim_{n\to\infty}\frac{B_n}{Y_n}=\frac{2}{\alpha}(1-\alpha). \tag{108}$$

**Proof.** Denote by $b(x)$ and $y(x)$ the generating function for the number of bulges and interior loops, respectively. By construction they fulfil $b(x)+y(x)=\ell_2(x)$. It is thus sufficient to calculate $y(x)$ from recursion (18). We find

$$y(x) = \frac{1}{1-x^2s^2} s^2x^4 j_1(x) \tag{109}$$

and, thus,

$$b(x) = \ell_2(x) - y(x) = \frac{1}{1-x^2s^2} x^2 s[s(1-x^2)j_1 - (1-x)s + 1]. \tag{110}$$

Corollary 4.4 and a simple calculation complete the proof. □

## 5. Secondary structures on a sequence

So far we have neglected the fact that secondary structures are built on sequences. Not all secondary structures can be formed by a given biological sequence, since not all combinations of nucleotides form base pairs. The results of the previous sections will be generalized to this situation in the remaining part of the paper.

**Definition 5.1.** Let $\mathscr{A}$ be some finite alphabet of size $\kappa$, let $\Pi$ be a symmetric Boolean $\kappa \times \kappa$-matrix and let $\Sigma = [\sigma_1 \dots \sigma_n]$ be a string of length $n$ over $\mathscr{A}$. A secondary structure is *compatible* with the sequence $\Sigma$ if $\Pi_{\sigma_p,\sigma_q} = 1$ for all base pairs $(p,q)$.

Following [12, 37] the number of secondary structures $\mathscr{S}$ compatible with some string can be enumerated as follows: Denote by $S_{p,q}$ the number of structures compatible with the substring $[\sigma_p \dots \sigma_q]$. Then

$$S_{l,n+1} = S_{l,n} + \sum_{k=l}^{n-m} S_{l,k-1} S_{k+1,n} \Pi_{\sigma_k,\sigma_{n+1}}. \tag{111}$$

Consider a random sequence with a Bernoulli distribution of the characters. In this case the expected number $\bar{S}_n$ of compatible structures is then [38]

$$\bar{S}_{n+1} = \bar{S}_n + p \sum_{k=1}^{n-m} \bar{S}_{k-1} \bar{S}_{n-k} = \bar{S}_n + p \sum_{k=m}^{n-1} \bar{S}_k \bar{S}_{n-k-1}, \tag{112}$$

where

$$p = \frac{1}{\kappa^2} \sum_{i,j=1}^{\kappa} \Pi_{ij} \tag{113}$$

is called the *stickiness* [15]. Note that Eq. (112) is not true if the characters along the sequence are correlated as it is the case for instance in a Markov model of the sequence. In the following, we will write $X_n$ to mean the expected value of $X$ on sequences of length $n$ with Bernoulli distributed characters.

A secondary structure compatible with a given sequence with maximal number of base pairs can be determined by a dynamic programming algorithm [19]. This observation was the starting point for the construction of reliable energy-directed folding algorithms (see, e.g., [35, 38, 17, 10]) and a recursive computation of the density of states [3].

All recursions in Section 3 are sums of linear terms of the form $A_n$ and quadratic terms of the type

$$\sum_{k=m}^{n-1} B_k C_{n-k-1} = \sum_{k=1}^{n-m} C_{k-1} B_{n-k}. \tag{114}$$

The corresponding recursions for structures compatible with a string can then be found by the rule

$$\begin{aligned} &A_n \rightarrow A_{l,n}, \\ &\sum_{k=1}^{n-m} C_{k-1} B_{n-k} \rightarrow \sum_{k=l}^{n-m} C_{l,k-1} B_{k+1,n} \Pi_{\sigma_k,\sigma_{n+1}}. \end{aligned} \tag{115}$$

For expected numbers assuming Bernoulli distributed sequences these rules simplify to

$$\begin{aligned} &A_n \rightarrow A_n, \\ &\sum_{k=m}^{n-1} B_k C_{n-k-1} \rightarrow p \sum_{k=m}^{n-1} B_k C_{n-k-1}. \end{aligned} \tag{116}$$

As an example we compute the expected fraction of unpaired digits in a secondary structure compatible with a random sequence with stickiness $p$. Applying these rules to Eq. (9) leads to the recursion

$$\begin{aligned} &U_{n+1} = (U_n + S_n) + p \sum_{k=m}^{n-1} [S_k U_{n-k-1} + S_{n-k-1} U_k], \quad n \geqslant m+1, \\ &U_n = n, \quad n \leqslant m+1. \end{aligned} \tag{117}$$

From Eqs. (112) and (117) we obtain the functional equations

$$1 = s[1 - x - px^2 s + px^2 t_m] \tag{118}$$

for the generating function $s$ of the number of secondary structures, and

$$u = xu + xs + p[2x^2 us - x^2 ut_m - x^2 s\tau_m] \tag{119}$$

for the generating function $u$ of the number of unpaired digits.

The asymptotics for $\bar{S}_n(p)$ can be calculated in analogy to Theorem 4.8. The functional equation for $s(x)$ yields

$$\begin{aligned} &\alpha\beta = 1/\sqrt{p}, \\ &\frac{1}{\sqrt{p}} - \left(2 + \frac{1}{\sqrt{p}}\right)\alpha + \sqrt{p}\alpha^2 t_m(\alpha) = 0. \end{aligned} \tag{120}$$

Furthermore, we have the following generalization of Lemma 4.9 for arbitrary $p \leqslant 1$:

**Lemma 5.3.**

$$\begin{aligned} t_m(\alpha) &= \frac{(1+2\sqrt{p})\alpha - 1}{p\alpha^2}, \\ \tau_m(\alpha) &= \frac{\alpha - 1 + 2\alpha\sqrt{p}}{\alpha p(1-\alpha)} - m\frac{(\alpha - 1 + \alpha\sqrt{p})^2}{\alpha^2 p(1-\alpha)}, \\ g^2(\alpha) &= \frac{(1-\alpha-\sqrt{p}\alpha)(2+m(1-\alpha-\sqrt{p}\alpha))}{\sqrt{p}^3(1-\alpha)\alpha^3}. \end{aligned} \tag{121}$$

Combining Eqs. (119) and (118) $u$ simplifies to

$$u = \frac{s^2 x(1 - p\tau_m x)}{1 - ps^2x^2} \tag{122}$$

and Corollary 4.3 implies that

$$\lim_{n\to\infty} \frac{U_n}{nS_n} = \frac{1}{\alpha g^2(\alpha)p}\left[\frac{1}{\sqrt{p}\alpha} - \sqrt{p}\tau_m(\alpha)\right] = \frac{2\alpha + m(1-\alpha-\sqrt{p}\alpha)}{2+m(1-\alpha-\sqrt{p}\alpha)}. \tag{123}$$

Note that $U_n/S_n$ refers to the fraction of the expected values of $U_n$ and $S_n$, not to the expected value of the fraction!

The asymptotics of the most important series are given below without proofs which do not differ significantly from the proof of the $p=1$ case. Some numerical values are given in Table 3. The stickiness value $p=0.5$ corresponds to a binary alphabet of complementary bases, while $p=0.25$ corresponds to a four letter alphabet with two pairs of complementary bases as in the (such as the biophysical **AUCG** with Watson–Crick pairing rules). Biological RNA structures frequently contain **G–U** pairs. Therefore they are best modeled by a value of $p=\frac{3}{8}$.

Number of Loops and stacks:

$$v = \frac{s(1-s(1-x))(px^2-1)}{1-ps^2x^2},$$

$$\lim_{n\to\infty} \frac{N_n}{S_n} = \frac{(1-\alpha)(1-\alpha^2 p)}{2+m(1-\alpha-\alpha\sqrt{p})}. \tag{124}$$

Number of components:

$$i = s^2(1-x) - s,$$

$$\lim_{n\to\infty} \frac{I_n}{S_n} = 2\beta(1-\alpha) - 1. \tag{125}$$

Table 3
Asymptotics of some structure elements as a function of stickiness

| $p$ | 1 | 0.5 | 0.375 | 0.25 |
|---|---|---|---|---|
| | | GC | GCAU | GCXK |
| $\alpha$ | 0.4369 | 0.5092 | 0.5391 | 0.5809 |
| $U_n/nS_n$ | 0.5265 | 0.5897 | 0.6147 | 0.6487 |
| $P_n/nS_n$ | 0.2368 | 0.2051 | 0.1926 | 0.1756 |
| $N_n/nS_n$ | 0.1915 | 0.1786 | 0.1717 | 0.1608 |
| $I_n/S_n$ | 1.5776 | 1.7266 | 1.7918 | 1.8855 |
| $L_n(1)/N_n$ | 0.2769 | 0.3062 | 0.3183 | 0.3352 |
| $L_n(2)/N_n$ | 0.5082 | 0.4692 | 0.4537 | 0.4325 |
| $B_n/Y_n$ | 2.5776 | 1.9280 | 1.7096 | 1.4428 |
| Stacklength | 1.2363 | 1.1487 | 1.1220 | 1.0924 |
| Loopsize | 2.7493 | 3.3018 | 3.5801 | 4.0342 |
| $E_n/S_n$ | 2 | 2.828 | 3.266 | 4 |

Loops with degree 2, i.e., interior loops and bulges:

$$l_2 = \frac{psx^2[(1-x)^2 - s(1-x)^3 + psx^2(s-t_m)]}{(1-x)^2(1-ps^2x^2)},$$

$$\lim_{n\to\infty} \frac{L_n(2)}{N_n} = \frac{(2-\alpha)\alpha^3 p}{(1-\alpha)^2(1-\alpha^2 p)},$$

$$\lim_{n\to\infty} \frac{B_n}{Y_n} = 2/\alpha - 2. \tag{126}$$

Hairpins:

$$l_1 = \frac{ps^2x^2(1-(1-x)t_m)}{(1-x)(1-ps^2x^2)},$$

$$\lim_{n\to\infty} \frac{L_n(1)}{N_n} = \frac{1-\alpha-\alpha\sqrt{p}}{1-\alpha-\alpha^2 p+\alpha^3 p}. \tag{127}$$

A detailed comparison of the structure statistics derived here with numerical data obtained by energy directed folding of RNA molecules is discussed in [30]. As could be expected, structures obtained by energy minimization tend to contain longer stacks and as a consequence more base pairs. The distribution of loop sizes and loop degrees, on the other hand, seems to be dominated by the combinatorics.

## Acknowledgements

Stimulating discussions with Drs. Walter Fontana and Danielle Konings are gratefully acknowledged. This work was supported financially by the Austrian *Fonds zur*

*Förderung der Wissenschaftlichen Forschung*, Projects No. S 5305-PHY and P 8526-MOB.

## References

[1] E.A. Bender, Asymptotic methods in enumeration, SIAM Rev. 16 (1974) 485–515.

[2] E.R. Canfield, Remarks on an asymptotic method in combinatoric, J. Combin. Theory A 37 (1984) 348–352.

[3] J. Cupal, I.L. Hofacker, P.F. Stadler, Dynamic programming algorithm for the density of states of RNA secondary structures, in: R. Hofstädt, T. Lengauer, M. Löffler, D. Schomburg (Eds.), Computer Science and Biology 96, Proc. German Conf. on Bioinformatics, Universität Leipzig, Leipzig, Germany, 1996, pp. 184–186.

[4] G. Darboux, Mémoir sur l'approximation des fonctions de très grande nombres, et sur une classe étendu de développements en série, J. Math. Pure Appl. 4 (1878) 5–56.

[5] R. Donaghey, L.W. Shapiro, Motzkin numbers, J. Combin. Theor. A 23 (1977) 291–301.

[6] W. Fontana, T. Griesmacher, W. Schnabl, P.F. Stadler, P. Schuster, Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures, Monatsh. Chem. 122 (1991) 795–819.

[7] W. Fontana, D.A.M. Konings, P.F. Stadler, P. Schuster, Statistics of RNA secondary structures, Biopolymers 33 (1993) 1389–1404.

[8] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, P. Schuster, Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks, Monatsh. Chem. 127 (1996) 355–374.

[9] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I.L. Hofacker, P.F. Stadler, P. Schuster, Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering, Monatsh. Chem. 127 (1996) 375–389.

[10] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M. Tacker, P. Schuster, Fast folding and comparison of RNA secondary structures, Monatsh. Chemie 125 (1994) 167–188.

[11] P. Hogeweg, B. Hesper, Energy directed folding of RNA sequences, Nucleic acids research 12 (1984) 67–74.

[12] J.A. Howell, T.F. Smith, M.S. Waterman, Computation of generating functions for biological molecules, SIAM J. Appl. Math. 39 (1980) 119–133.

[13] W.N. Hsieh, Proportions of irreducible diagrams, Studies Appl. Math. 52 (1973) 277–283.

[14] D. Kleitman, Proportions of irreducible diagrams, Studies Appl. Math. 49 (1970) 297–299.

[15] 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.

[16] J. Leydold, P.F. Stadler, Minimal cycle bases of outerplanar graphs, Elec. J. Combin. 5 (1998) R16.

[17] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers 29 (1990) 1105–1119.

[18] A. Meir, J.W. Moon, On an asymptotic method in enumeration, J. Combin. Theory A 51 (1989) 77–89.

[19] R. Nussinov, G. Piecznik, J.R. Griggs, D.J. Kleitman, Algorithms for loop matching, SIAM J. Appl. Math. 35 (1978) 68–82.

[20] A.M. Odlyzko, Asymptotic enumeration methods, in: R.L. Graham, M. Grötschel, L. Lovász (Eds.), Handbook of Combinatorics, vol. II, Elsevier, Amsterdam, 1995, pp. 1021–1229.

[21] R.C. Penner, M.S. Waterman, Spaces of RNA secondary structures, Adv. Math. 101 (1993) 31–49.

[22] M. Régnier, F. Tahi, Enumeration and asymptotics in computational biology, in: Mathematical Analysis for Biological Sequences Workshop, Trondheim, Norway, 1996.

[23] W.R. Schmitt, M.S. Waterman, Linear trees and RNA secondary structure, Discr. Appl. Math. 12 (1994) 412–427.

[24] B.A. Shapiro, An algorithm for comparing multiple RNA secondary structures, CABIOS 4 (1988) 387–397.

[25] B.A. Shapiro, K. Zhang, Comparing multiple RNA secondary structures using tree comparisons, CABIOS 6 (1990) 309–318.

[26] P.R. Stein, On a class of linked diagrams, I. Enumeration, J. Combin. Theory A 24 (1978) 357–366.

[27] P.R. Stein, C.J. Everett, On a class of linked diagrams. II. Asymptotics, Disc. Math. 22 (1978) 309–318.

[28] P.R. Stein, M.S. Waterman, On some new sequences generalizing the Catalan and Motzkin numbers, Disc. Math. 26 (1978) 261–272.

[29] G. Szegö, Orthogonal Polynomials, Amer. Math. Soc. Coll. Publ. vol. XXIII, Amer. Math. Soc., New York, 1959.

[30] M. Tacker, P.F. Stadler, E.G. Bornberg-Bauer, I.L. Hofacker, P. Schuster, Algorithm independent properties of RNA structure prediction, Eur. Biophy. J. 25 (1996) 115–130.

[31] F. Tahi, Méthodes formelles d'analyse des séquences de nucléotides, Ph.D. Thesis, Université de Paris XI, Orsay, 1997.

[32] J. Touchard, Sur une problème de configurations et sur les fractions continues, Canad. J. Math. 4 (1952) 2–25.

[33] X.G. Viennot, M.V. de Chaumont, Enumeration of RNA's secondary structures by complexity, in: V. Capasso, E. Grosso, S.L. Paveri-Fontana (Eds.), Mathematics in Medicine and Biology, Lect. Notes in Biomath., vol. 57, Springer, Berlin, 1985, pp. 360–365.

[34] M.S. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math. Suppl. Studies 1 (1978) 167–212.

[35] M.S. Waterman, Introduction to Computational Biology: Maps, Sequences, and Genomes, Chapman & Hall, London, 1995.

[36] M.S. Waterman, T.F. Smith, Combinatorics of RNA hairpins and cloverleaves, Studies Appl. Math. 60 (1978) 91–96.

[37] M.S. Waterman, T.F. Smith, RNA secondary structure: a complete mathematical analysis, Math. Biosci. 42 (1978) 257–266.

[38] M. Zuker, D. Sankoff, RNA secondary structures and their prediction, Bull. Math. Biol. 46 (4) (1984) 591–621.