Available online at www.sciencedirect.com

SCIENCE DIRECT®

Discrete Mathematics 285 (2004) 67–82

DISCRETE MATHEMATICS

www.elsevier.com/locate/disc

Enumerative aspects of secondary structures

Tomislav Došlić[a], Dragutin Svrtan[b], Darko Veljan[b,*]

[a] *University of Zagreb, Faculty of Agriculture, Svetošimunska c. 25, Zagreb, Croatia*
[b] *Department of Mathematics, University of Zagreb, Bijenička 30, Zagreb, Croatia*

Received 20 March 2001; received in revised form 9 January 2004; accepted 13 April 2004

Abstract

A secondary structure is a planar, labeled graph on the vertex set $\{1,\ldots,n\}$ having two kind of edges: the segments $[i,i+1]$, for $1 \leqslant i \leqslant n-1$ and arcs in the upper half-plane connecting some vertices i,j, $i \leqslant j$, where $j-i > l$, for some fixed integer l. Any two arcs must be totally disjoint. We enumerate secondary structures with respect to their size n, rank l and order k (number of arcs), obtaining recursions and, in some cases, explicit formulae in terms of Motzkin, Catalan, and Narayana numbers. We give the asymptotics for the enumerating sequences and prove their log-convexity, log-concavity and unimodality. It is shown how these structures are connected with hypergeometric functions and orthogonal polynomials.

MSC: 05A15; 05A16; 05A20; 05B50; 05C30; 05C70; 33C45; 92D20

Keywords: Secondary structures; Motzkin path; Motzkin numbers; Dyck path; Narayana numbers; Log-convexity; Orthogonal polynomials

1. Introduction

Let us first explain briefly the biological background of secondary structures and then turn to their mathematics. In biology, many important molecules belong to the class of linear polymers. Linear polymers are long chains built from simpler building blocks, and these blocks are called monomers. For example, polysacharides are long chains of simple sugars while proteins are chains of amino-acids. The DNA molecules are also linear polymers, as well as the very similar molecules of RNA. Together they make the class of nucleic acids; the building blocks of nucleic acids are called nucleotides. The nucleic acids play an important role in coding, transferring and retrieving genetic information, and in directing cell metabolism.

There are four different kinds of nucleotides, which only differ by one part, called *base*. Hence, one usually identifies nucleotides and bases. The bases are denoted by letters A, C, G and U. Each nucleotide is a polar molecule with two differing ends, usually denoted by $5'$ and $3'$. With the exception of one terminal nucleotide, the $5'$ end of one nucleotide fits to the $3'$ end of another nucleotide forming a *p-bond* (p stands for phosphorus); a sequence of such bonds is called a *backbone* of the molecule. The sequence of nucleotides (or bases), when read from the $5'$ end of the chain constitutes the *primary structure* of an RNA molecule. As each nucleotide can bind by a p-bond to another nucleotide, the primary structure of an n-bases RNA molecule is then simply an n-letter word in the alphabet $\{A,C,G,U\}$. The number of all possible primary structures for an n-base RNA molecule is thus equal to 4^n.

Certain pairs of bases, namely C and G, A and U, and G and U, exhibit mutual chemical affinity. They tend to form *h-bonds* (h for hydrogen), which cause folding of the molecular backbone into configurations of minimal energy. The planar folding of an RNA molecule is then called its *secondary structure*; the non-planar folding is the molecule's *tertiary*

* Corresponding author.
E-mail addresses: doslic@math.hr (T. Došlić), dsvrtan@math.hr (D. Svrtan), dveljan@math.hr (D. Veljan).

0012-365X/$ - see front matter
doi:10.1016/j.disc.2004.04.001

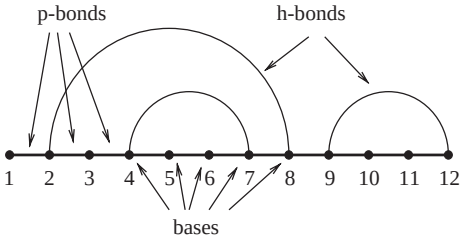

Fig. 1. An example of a secondary structure.

structure. Secondary and tertiary structures are important because they determine the three-dimensional shape of an RNA molecule, which in turn contributes to determining its biological function.

The secondary structures of a given molecule are subject to certain stereo-chemical constraints. We describe them here in some detail to explain the conditions to be imposed on the mathematical objects representing secondary structures. Firstly, no base can participate in more than one h-bond. Secondly, a base cannot be paired by an h-bond to a base that is too close along the backbone, due to the rigidity of the backbone's p-bonds. Thirdly, h-bonds may not cross. It means that, if there is an h-bond pairing the bases i and j, and an h-bond pairing the bases k and m, then either $i < j < k < m$ or $i < k < m < j$. In biological terminology, this constraint prohibits *pseudoknots*. (Of course, the Nature employs the pseudoknots freely, but we consider them as the elements of the molecule's tertiary structure.)

Besides their importance in molecular biology, where secondary structures serve as the phenotypes in evolution experiments in vitro, and where the folding of RNA molecules into structures is the simplest known case of a genotype–phenotype mapping, secondary structures raise many mathematically interesting questions concerning their enumeration and prediction. Some of these questions are treated in [4–6,9,10,13,14,16,18,20,21] We consider here mostly those that can be formulated and answered in the framework of enumerative combinatorics.

The natural setting for the combinatorial modeling of secondary structures is graph theory. We represent the bases by vertices and the bonds by edges of certain graphs. The stereo-chemical constraints translate quite naturally into the graph-theoretical language. We refer the reader to [22] for all graph-theoretical terms not defined here.

A *secondary structure* of *size* $n \geqslant 1$ and *rank* $l \geqslant 0$ is a labeled non-oriented graph S on the vertex set $V(S) = [n] = \{1, 2, \ldots, n\}$ whose edge set $E(S)$ consists of two disjoint subsets, $P(S)$ and $H(S)$, satisfying the following conditions:

(a) $\{i, i+1\} \in P(S)$, for all $1 \leqslant i \leqslant n-1$;
(b) $\{i, j\} \in H(S)$ and $\{i, k\} \in H(S) \Rightarrow j = k$;
(c) $\{i, j\} \in H(S) \Rightarrow |i - j| > l$;
(d) $\{i, j\} \in H(S), \{k, m\} \in H(S)$ and $i < k < j \Rightarrow i < m < j$.

Obviously, the set $P(S)$ contains the edges corresponding to the p-bonds of a molecule's backbone. The set $H(S)$, which may be empty, contains the edges representing the h-bonds, and we will often use the term h-bond when referring to an edge from $H(S)$. Any two vertices connected by an edge from $H(S)$ are called paired; if a vertex of S is not incident to any edge from $H(S)$, it is called unpaired. The cardinality of $H(S)$ is the *order* of S, and the parameter l is the structure's *rank*.

An example of a secondary structure of size 12, rank 2, and order 3 is shown on Fig. 1. Note that every h-bond "leaps" over at least two bases.

We denote the set of all secondary structures of size n and rank l by $\mathscr{S}^{(l)}(n)$. The set of all such structures of order k is denoted by $\mathscr{S}_k^{(l)}(n)$. The cardinalities of these sets will be denoted by $S^{(l)}(n)$ and $S_k^{(l)}(n)$, respectively. By definition, we put $S^{(l)}(0) = 1$, for all l.

There are many ways of representing secondary structures graphically. Three of them, the loop diagram, the chord diagram, and the arc diagram, are shown in Fig. 2(a), (b), and (c), respectively. Some other representations are discussed in [14]. One among them, the so-called "mountain representation" puts the secondary structures into the well-researched context of lattice paths. Here we remind the reader on some basic families of lattice paths and their corresponding number sequences.

A *Dyck path* of length $2n$ is a lattice path in the coordinate plane (x, y) from $(0, 0)$ to $(2n, 0)$ with steps $(1, 1)$ (Up) and $(1, -1)$ (Down), never falling below the x-axis. We denote the set of all Dyck paths of length $2n$ by $\mathscr{D}(n)$. A *peak* of a Dyck path is a consecutive Up–Down step pair. The set of all Dyck paths of length $2n$ with exactly k peaks $(1 \leqslant k \leqslant n)$ is denoted by $\mathscr{D}_k(n)$. It is well known that Dyck paths are enumerated by Catalan numbers, i.e. that $|\mathscr{D}(n)| = C_n$, where $C_n = (1/(n+1))\binom{2n}{n}$, see Exercise 6.19 of [17].

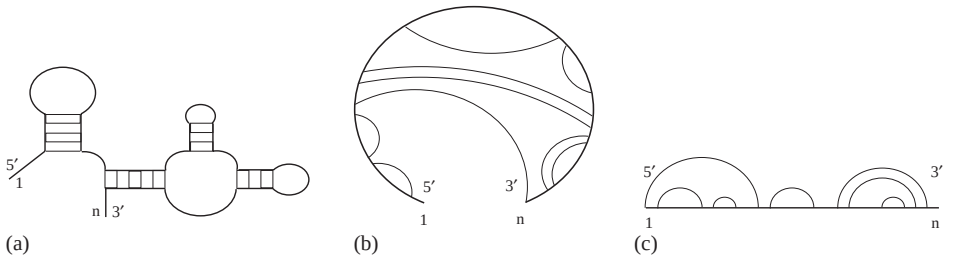

Fig. 2. Three representations of secondary structures.

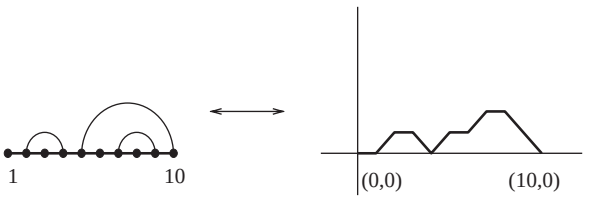

Fig. 3. A secondary structure and the corresponding Motzkin path.

A *Motzkin path* of length n is a lattice path in (x, y) plane from $(0,0)$ to $(n,0)$ with steps $(1,1)$ (Up), $(1,-1)$ (Down) and $(1,0)$ (Level), never falling below the x-axis. We denote the set of all Motzkin paths of length n (i.e. with exactly n steps) by $\mathscr{M}(n)$. The number $M_n = |\mathscr{M}(n)|$ is nth *Motzkin number*. By definition, $M_0 = 1$. A *plateau* of length l is a sequence of l consecutive Level steps, immediately preceded by an Up step, and immediately followed by a Down step. Let us denote by $\mathscr{M}^{(l)}(n)$ the set of all paths from $\mathscr{M}(n)$ whose every plateau is at least l steps long. For $l = 0$ we get simply $\mathscr{M}^{(0)}(n) = \mathscr{M}(n)$.

The following folklore result has been used for enumeration purposes in [20,9].

Proposition 1.1. *There is a bijection between* $\mathscr{M}^{(l)}(n)$ *and* $\mathscr{S}^{(l)}(n)$, *for all* $n \geqslant 1,\ l \geqslant 0$.

A secondary structure from $\mathscr{S}^{(1)}(10)$ and the corresponding Motzkin path from $\mathscr{M}^{(1)}(10)$ are shown in Fig. 3.

Let us now examine the effect of relaxation of our constraints by allowing the parameter l to be -1. Now we allow an h-bond to terminate at the same base from which it begun, i.e. we allow loops. This situation is biologically highly unrealistic, but the result emphasizes the intrinsic connection between secondary structures and other objects counted by the Catalan–Motzkin family of sequences, and concurs with the appearance of the Catalan numbers in the title of [18].

Proposition 1.2. $S^{(-1)}(n) = C_{n+1}$ *for all* $n \geqslant 0$.

Proof. We will exhibit a bijection between $\mathscr{S}^{(-1)}(n)$ and $\mathscr{D}(n+1)$. Take a Dyck path on $2(n+1)$ steps. Discard the first and the last step and divide the remaining steps in groups of pairs of consecutive steps. Assign to each pair a vertex in a secondary structure according to the following rule. To a pair of two Up steps assign a vertex in which an h-bond starts; to a pair of two Down steps assign a vertex in which an h-bond terminates; to a pair consisting of (Up, Down) pair we assign a vertex with a loop attached to it, and to a pair (Down, Up) we assign an unpaired vertex. By definition of Dyck paths, we get a valid secondary structure of rank -1, and the construction is obviously bijective. □

An example of described correspondence is shown in Fig. 4. It is clear from the above construction that all the peaks in a Dyck path that correspond to loops in a given secondary structure of rank -1 must have even altitudes. Since the secondary structures without loops are counted by Motzkin numbers, we have the following result.

Corollary 1.3. *The nth Motzkin number,* M_n, *counts the number of Dyck paths on* $2n$ *steps without peaks at even altitudes.*

Almost all our results in forthcoming sections can be extended also to the case $l = -1$. However, in all cases they reduce to known facts about Catalan numbers. Hence, through the rest of this paper we assume $l \geqslant 0$.

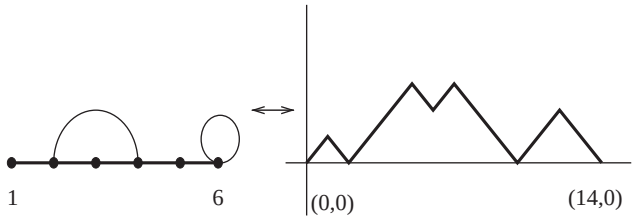

Fig. 4. The correspondence between secondary structures of rank −1 and Dyck paths.

2. Sequences $S^{(l)}(n)$

2.1. Convolutions and generating functions

We start this section by stating some results needed for our investigations [5,18].
For a fixed integer $l \geqslant 0$ we have

$$S^{(l)}(n+1) = S^{(l)}(n) + \sum_{k=l}^{n-1} S^{(l)}(k)S^{(l)}(n-k-1), \qquad n \geqslant l+1,$$

$$S^{(l)}(0) = S^{(l)}(1) = \cdots = S^{(l)}(l+1) = 1. \tag{1}$$

Let $S_l(x) = \sum_{n\geqslant 0} S^{(l)}(n)x^n$ be the *generating function* of the sequence $S^{(l)}(n)$, $n \geqslant 0$. From (1) it follows easily that the function $S_l(x)$, $l \geqslant 0$, satisfies the functional equation

$$x^2[S_l(x)]^2 + \omega_l(x)S_l(x) + 1 = 0, \tag{2}$$

where

$$\omega_l(x) = x - 1 - x^2 - x^3 - \cdots - x^{l+1} = 2x - (1 + x + x^2 + \cdots + x^{l+1}) = x - 1 - x^2\,\frac{1-x^l}{1-x}.$$

The following explicit formula for $S_l(x)$ will be useful for establishing the asymptotic behavior of numbers $S^{(l)}(n)$:

$$S_l(x) = \frac{-\omega_l(x) - \sqrt{\omega_l^2(x) - 4x^2}}{2x^2} = \frac{1 - x + \cdots + x^{l+1} - \sqrt{(1 + x + \cdots + x^{l+1})(1 + x + \cdots + x^{l+1} - 4x)}}{2x^2}. \tag{3}$$

By dividing Eq. (3) for $l+1$ by $xS_{l+1}(x)$ and using $\omega_{l+1}(x) = \omega_l(x) - x^{l+2}$, we get the following simple recursion for the generating function $S_l(x)$, $l \geqslant 0$:

$$xS_{l+1}(x) + \frac{1}{xS_{l+1}(x)} = xS_l(x) + \frac{1}{xS_l(x)} + x^{l+1}.$$

The initial condition for this recurrence is

$$S_0(x) = \frac{1 - x - \sqrt{1 - 2x - 3x^2}}{2x^2} = M(x),$$

the generating function for Motzkin numbers M_n. By expanding the term $\sqrt{1-2x-3x^2}$ in $M(x)$ via binomial series and then taking the coefficient of x^{n+2} on both sides of $1 - x - \sqrt{1-2x-3x^2} = 2x^2M(x)$, after some manipulations we obtain the following expression of M_n.

Proposition 2.1.

$$M_n = \left(\frac{3}{2}\right)^{n+2} \sum_{k\geqslant 1} \frac{1}{3^k}\, C_{k-1} \binom{k}{n+2-k}, \qquad n \geqslant 0.$$

2.2. Asymptotic behavior of $S^{(l)}(n)$

From the explicit form of the generating function (3) we can get the asymptotics for the numbers $S^{(l)}(n)$ from a version of Darboux theorem [11].

Table 1
Exact values of parameters α_l

l	α_l
0	3
1	$\left(3+\sqrt{5}\right)/2$
2	$\sqrt{2}+1$
3	$\frac{3}{4}+\frac{1}{12}\left[\sqrt{57+6w}+\sqrt{114-6w+126\sqrt{3}/\sqrt{19+2w}}\right], \quad w=\left(908+12\sqrt{993}\right)^{1/3}+\left(908-12\sqrt{993}\right)^{1/3}$
4	$\frac{1}{6}\left(172+12\sqrt{177}\right)^{1/3}+\frac{8}{3}\left(172+12\sqrt{177}\right)^{-1/3}+\frac{2}{3}$
5	$\frac{1}{6}\left(188+12\sqrt{93}\right)^{1/3}+\frac{14}{3}\left(188+12\sqrt{93}\right)^{-1/3}+\frac{1}{3}$
6	$\frac{1}{2}+\frac{1}{6}\sqrt{9-3u}+\left(\sqrt{3}/6\right)\sqrt{6+u+18/\sqrt{9-3u}}, \quad u=\left(54+6\sqrt{129}\right)^{1/3}+\left(54-6\sqrt{129}\right)^{1/3}$

Proposition 2.2.

$$S^{(l)}(n)\sim\frac{h_l(r_l)n^{-3/2}}{\Gamma(-1/2)r_l^n},$$

where

$$h_l(x)=-\frac{\sqrt{-\omega_l(x)+2x}\sqrt{w_l(x)}}{2x^2},$$

$$w_l(x)=\frac{-\omega_l(x)-2x}{1-x/r_l}$$

and r_l is the smallest positive root of the discriminant $\Delta_l(x)=\omega_l^2(x)-4x^2=0$.

In other words, $S^{(l)}(n)\sim\left(h_l(r_l)/\Gamma\left(-\frac{1}{2}\right)\right)n^{-3/2}\alpha_l^n$, where $\alpha_l=1/r_l$.

Let us investigate in more detail how $r=r_l$ depends on l.

$$\Delta_l(x)=\omega_l^2(x)-4x^2=(-\omega_l(x)+2x)(-\omega_l(x)-2x)=(1+x+\cdots+x^{l+1})(1+x+\cdots+x^{l+1}-4x).$$

As the first factor is strictly positive for all positive x, r_l must satisfy

$$1+x+\cdots+x^{l+1}=4x.$$

This equation can be solved exactly for $l=0,1,2$, and 3. The solutions for 0, 1, and 2 are $r_0=\frac{1}{3}$, $r_1=\left(3-\sqrt{5}\right)/2$, and $r_2=\sqrt{2}-1$, respectively. The solution r_3 of the equation $x^4+x^3+x^2-3x+1=0$ can also be obtained in explicit form, using Ferrari method.

It is also possible to obtain the exact solutions of the equation $\Delta_l=0$ for the cases $l=4,5$, and 6. Observe that the above equation can be written in the form

$$\frac{1-x^{l+2}}{1-x}=4x$$

or, equivalently, $x^{l+2}=(2x-1)^2$. For even values of l we then get $(x^{l/2+1}-2x+1)(x^{l/2+1}+2x-1)=0$. The first factor is positive on the interval $\left[0,\frac{1}{2}\right]$, and the second factor changes its sign there, so r_l must be a root of the second factor. It is clear that this is the only root of $x^{l/2+1}+2x-1$ in this interval. The case $l=4$ gives an equation $x^3=1-2x$, which we solve explicitly using the Cardano formula. In the case $l=6$, r_6 is obtained by solving the equation $x^4=1-2x$ using Ferrari method. The case $l=5$ does not fit into this pattern, but its discriminant factors in a nice way, thus enabling us to get an exact value of r_5, also.

The exact values of the parameters $\alpha_l=1/r_l$ for $0\leqslant l\leqslant 6$ are listed in Table 1. For the values of $l\geqslant 7$, the equation $x^{l/2+1}=1-2x$ cannot be solved in radicals. However, using the generalized binomial series [3, p. 200], one can easily obtain the following representation of r_l.

Proposition 2.3. *Let r_l be the smallest positive root of the equation $x^{l/2+1} = 1 - 2x$. Then*

$$r_l = \sum_j \frac{1}{2 + 2j + jl} \binom{1 + j + jl/2}{j} \left(-\frac{1}{2^{l/2+1}}\right)^j.$$

Proof. The generalized binomial series $\mathscr{B}_t(z)$ is defined for arbitrary complex values of t and z by

$$\mathscr{B}_t(z) = \sum_{k \geqslant 0} (tk)^{\underline{k-1}} \frac{z^k}{k!}.$$

Here $x^{\underline{m}}$ denotes the falling factorial, $x^{\underline{m}} = x(x-1)\cdots(x-m+1)$ [3, p. 47]. It is shown in [3, p. 363] that $\mathscr{B}_t(z)$ satisfies the following identity:

$$z\mathscr{B}_t(z)^t = \mathscr{B}_t(z) - 1.$$

The solution r_l of the equation $x^{l/2+1} = 1 - 2x$ must satisfy $r_l^{l/2+1} = 1 - 2r_l$. This can be rewritten as

$$\left(\frac{1}{2^{l/2+1}}\right)(2r_l)^{l/2+1} = 1 - 2r_l.$$

By multiplying this equation by -1, and by setting $l/2 + 1 = t$, $-1/2^{l/2+1} = z$, we obtain

$$z(2r_l)^t = 2r_l - 1.$$

Hence,

$$2r_l = \mathscr{B}_t(z) = \mathscr{B}_{l/2+1}\left(-\frac{1}{2^{l/2+1}}\right).$$

The claim of the proposition now follows by plugging the values of t and z into the formula

$$\mathscr{B}_t(z) = \sum_j \binom{jt+1}{j} \frac{z^j}{jt+1},$$

given on p. 363 of [3]. □

2.3. Short recursions

In the beginning of this section we stated convolutional recurrences for the secondary structure numbers. But the generating functions of $S^{(l)}(n)$ are algebraic of degree 2 for each $l \geqslant 0$, and hence D-finite [17, p. 190]. So, there must exist short, non-convolutional recurrences for the numbers $S^{(l)}(n)$. More precisely, there are polynomials $P_0, P_1, \ldots, P_e$ with $P_e \neq 0$ such that

$$P_e(n)S^{(l)}(n+e) + \cdots + P_0(n)S^{(l)}(n) = 0, \quad n \in \mathbb{N}.$$

We start from the expression of (3), that is,

$$S_l(x) = \frac{1}{2x^2}\left[-\omega_l(x) - \sqrt{\omega_l^2(x) - 4x^2}\right].$$

Then we set

$$\omega_l^2(x) - 4x^2 = A(x) = \sum_{k=0}^{2l+2} a_k x^k,$$

where

$$a_k = \begin{cases} 1, & k = 0, \\ k - 3, & 1 \leqslant k \leqslant l + 1, \\ l - 3, & k = l + 2, \\ 2l + 3 - k, & l + 3 \leqslant k \leqslant 2l + 2. \end{cases}$$

Now we set $\sum_{n\geqslant 0} f(n)x^n = \sqrt{A(x)}$, and use formula (6.38) in [17, p. 190]:

$$\sum_{j=0}^{r} a_{r-j}(dn+dj-r+j)f(n+j)=0.$$

Substituting $d=2,\ r=2l+2$ in this formula, we get the following recursion for $f(n)$:

$$\sum_{j=0}^{2l+2} a_{2l+2-j}(2n+3j-2l-2)f(n+j)=0.$$

After some manipulations, we obtain

$$\begin{aligned}(n+2l+2)f(n+2l+2) &= (2n+4l+1)f(n+2l+1)+(n+2l-1)f(n+2l)\\ &\quad -\frac{1}{2}\left[\sum_{j=l+1}^{2l-2}(2l-j-1)(2n+3j-2l-2)f(n+j)+(l-3)(2n+l-2)f(n+l)\right.\\ &\quad \left.+\sum_{j=0}^{l-1}(j+1)(2n+3j-2l-2)f(n+j)\right].\end{aligned}$$

Taking now the coefficient of x^{n+2} in

$$2x^2S_l(x)+\omega_l(x)=-\sqrt{\omega_l^2(x)-4x^2},$$

we get $2S^{(l)}(n)=-f(n+2)$ and hence

$$\begin{aligned}(n+2)S^{(l)}(n) &= (2n+1)S^{(l)}(n-1)+(n-1)S^{(l)}(n-2)\\ &\quad -\frac{1}{2}\left[\sum_{j=l+1}^{2l-2}(2l-j-1)(2n-6l+3j-2)S^{(l)}(n-2l+j-2)+(l-3)(2n-3l-2)S^{(l)}(n-l-2)\right.\\ &\quad \left.+\sum_{j=0}^{l-1}(j+1)(2n-6l+3j-2)S^{(l)}(n-2l+j-2)\right].\end{aligned}$$

Finally, we can express $S^{(l)}(n)$ in terms of $S^{(l)}(n-k),\ k=1,\ldots,2l+2$.

Theorem 2.4.

$$(n+2)S^{(l)}(n)=\sum_{k=1}^{2l+2}A^{(l)}(n,k)S^{(l)}(n-k),$$

where

$$A^{(l)}(n,k)=\begin{cases}-\frac{1}{2}(k-3)(2n+4-3k), & 1\leqslant k\leqslant l+1,\\ -\frac{1}{2}(l-3)(2n-3l-2), & k=l+2,\\ -\frac{1}{2}(2l+3-k)(2n+4-3k), & l+3\leqslant k\leqslant 2l+2\end{cases}$$

with the initial conditions $S^{(l)}(0)=S^{(l)}(1)=\cdots=S^{(l)}(l+1)=1$, $S^{(l)}(l+m)=2^{m-1}$ *for* $2\leqslant m\leqslant l+1$.

In particular, for $l=0,1$, and 2, we have:

Corollary 2.5.

$$M_n=\frac{2n+1}{n+2}M_{n-1}+\frac{3n-3}{n+2}M_{n-2},\quad n\geqslant 2,$$

$$M_0=M_1=1,$$

$$S^{(1)}(n)=\frac{2n+1}{n+2}S^{(1)}(n-1)+\frac{n-1}{n+2}S^{(1)}(n-2)+\frac{2n-5}{n+2}S^{(1)}(n-3)-\frac{n-4}{n+2}S^{(1)}(n-4),$$

$$S^{(1)}(0)=S^{(1)}(1)=S^{(1)}(2)=1,\quad S^{(1)}(3)=2,$$

$$S^{(2)}(n)=\frac{2n+1}{n+2}S^{(2)}(n-1)+\frac{n-1}{n+2}S^{(2)}(n-2)+\frac{n-4}{n+2}S^{(2)}(n-4)-\frac{2n-11}{n+2}S^{(2)}(n-5)-\frac{n-7}{n+2}S^{(2)}(n-6),$$

$$S^{(2)}(0)=S^{(2)}(1)=S^{(2)}(2)=S^{(2)}(3)=1,\quad S^{(2)}(4)=2,\quad S^{(2)}(5)=4. \tag{4}$$

2.4. Log-convexity results

We conclude this section by proving log-convexity of the sequences $S^{(l)}(n)$ for fixed $l\geqslant 0$. Recall that a sequence a_n of positive numbers is called *logarithmically convex*, simply *log-convex*, if $a_n^2\leqslant a_{n-1}a_{n+1}$, for all n. Equivalently, a sequence a_n of positive numbers is log-convex if the sequence $x_n=a_n/a_{n-1}$ is increasing.

The log-convexity of Motzkin numbers was proved in [1] algebraically, and recently in [2] purely combinatorially. Unfortunately, none of these approaches is suitable for generalization to the cases $l>0$. We present here a method based on interlacing the sequence of quotients of successive elements of $S^{(1)}(n)$ with a suitably chosen increasing sequence [19].

Theorem 2.6. *The sequence $S^{(1)}(n)$ is log-convex.*

Proof. We start from the short recursion (4). Dividing this relation by $S^{(1)}(n-1)$, and denoting $S^{(1)}(n)/S^{(1)}(n-1)$ by x_n, we obtain the following recursion for the numbers x_n:

$$x_n=\frac{1}{n+2}\left[2n+1+\frac{n-1}{x_{n-1}}+\frac{2n-5}{x_{n-1}x_{n-2}}-\frac{n-4}{x_{n-1}x_{n-2}x_{n-3}}\right],\quad n\geqslant 4$$

with initial conditions $x_1=x_2=1,\ x_3=2$.

Define now a sequence $a_n=(2n/(2n+3))\phi^2$, where $\phi=\left(1+\sqrt{5}\right)/2$ is the golden ratio. Note that $\phi^2=\alpha_1=\left(3+\sqrt{5}\right)/2$, the constant from the asymptotics of the sequence $S^{(1)}(n)$. The sequence a_n is obviously increasing, and its limit is ϕ^2. We claim that this sequence is interlaced with our sequence x_n, i.e. that $a_n\leqslant x_n\leqslant a_{n+1}$, starting from $n=6$.

We will prove, by induction, that $(n+2)a_n\leqslant(n+2)x_n\leqslant(n+2)a_{n+1}$ for $n\geqslant 6$. First, we check directly the cases $n=6,7,8$ and 9. Now, take $n\geqslant 9$. From the induction hypothesis, we have

$$(n+2)x_n\geqslant 2n+1+\frac{n-1}{a_n}+\frac{n-1}{a_na_{n-1}}+(n-4)\frac{a_{n-3}-1}{a_na_{n-1}a_{n-2}}.$$

We would like the right-hand side to be at least $(n+2)a_n$. But this is equivalent to

$$(2n+1)a_na_{n-1}a_{n-2}+(n-1)a_{n-1}(a_{n-2}+1)+(n-4)(a_{n-3}+1)-(n+2)a_n^2a_{n-1}a_{n-2}\geqslant 0.$$

Plugging in the formulae for a_n's, we get

$$\frac{12\left(5+\sqrt{5}\right)n^4-2\left(241+121\sqrt{5}\right)n^3+2\left(847+382\sqrt{5}\right)n^2-3\left(341+146\sqrt{5}\right)n+126+54\sqrt{5}}{(2n-3)(2n-1)(2n+1)(2n+3)^2}\geqslant 0.$$

The denominator is positive for all integers $n\geqslant 2$. Now take the numerator, denote it by $L(n)$ and shift its argument by 6. The polynomial $L(n+6)$ has only positive coefficients, so it cannot have a positive root. From there follows that $L(n)$ cannot have a root $\gamma\geqslant 6$. So the above inequality is valid for all $n\geqslant 6$, and hence $x_n\geqslant a_n$.

To prove the other inequality, note that the induction hypothesis implies

$$(n+2)x_n\leqslant 2n+1+\frac{n-1}{a_{n-1}}+\frac{n-1}{a_{n-1}a_{n-2}}+(n-4)\frac{a_{n-2}-1}{a_{n-1}a_{n-2}a_{n-3}}.$$

The condition that the right-hand side of this inequality does not exceed $(n+2)a_{n+1}$ is equivalent to

$$(2n+1)a_{n-1}a_{n-2}a_{n-3}+(n-1)a_{n-3}(a_{n-2}+1)+(n-4)(a_{n-2}-1)-(n+2)a_{n+1}a_{n-1}a_{n-2}a_{n-3}\leqslant 0.$$

Plugging in the formulae for a_n's, we get

$$-3\frac{\left(82+42\sqrt{5}\right)n^3-\left(572+248\sqrt{5}\right)n^2+\left(1103+474\sqrt{5}\right)n-\left(529+247\sqrt{5}\right)}{(2n-3)(2n-1)(2n+1)(2n+5)}\leqslant 0.$$

If we put $n+5$ instead of n in the numerator, we get a polynomial with all the coefficients positive, and from this we conclude that the numerator does not change the sign for $n\geqslant 6$. So, we have proved the inequality $x_n\leqslant a_{n+1}$, and completed the induction step. This proves the theorem. □

Corollary 2.7. *The sequence* $x_n = S^{(1)}(n)/S^{(1)}(n-1)$ *is strictly increasing for all* $n \geqslant 5$. *The sequence is bounded from above by* ϕ^2 *and* $x_n \to \phi^2 = \left(3+\sqrt{5}\right)/2$ *as* $n \to \infty$.

The similar approach will give us the log-convexity of the secondary structure numbers for values $l = 2, 3$ and 4, but the details become more and more complicated. Hence, we only state the following result.

Theorem 2.8. *The sequences* $S^{(l)}(n)$ *are log-convex, for* $l = 2, 3$ *and* 4.

Remark 1. An indication that $S^{(l)}(n)$, $n \geqslant 0$ is (at least asymptotically) log-convex for any l is as follows. Let $x_n^{(l)} = S^{(l)}(n)/S^{(l)}(n-1)$. From the remark after Proposition 2.2, it follows that the ratio $S^{(l)}(n)/S^{(l)}(n-1)$ is of the form

$$x_n^{(l)} \sim \alpha_l \left(1 - \frac{1}{n}\right)^{3/2}$$

and this increasingly tends to α_l. In other words, $x_n^{(l)}$ asymptotically behaves as an increasing sequence tending to α_l as $n \to \infty$.

3. The numbers $S_k^{(l)}(n)$

We have so far examined how the number of secondary structures for RNA molecules depends on the size n and the rank l. Let us now consider the role of the parameter defining its order, i.e. the number of its h-bonds.

The *Narayana numbers* $N(n,k)$ are defined for integers $n, k \geqslant 1$ by

$$N(n,k) = \frac{1}{n}\binom{n}{k}\binom{n}{k-1} = \frac{1}{k}\binom{n}{k-1}\binom{n-1}{k-1}$$

with the initial value $N(0,0) := 1$ and the boundary values $N(n,0) = 0$, $N(n,1) = 1$ for $n \geqslant 1$ [8].

A combinatorial proof that the Narayana numbers $N(n,k)$ enumerate Dyck paths on $2n$ steps with exactly k peaks and thus decompose the Catalan numbers, i.e. that $C_n = \sum_{k \geqslant 0} N(n,k)$ is given in the solution of Exercise 6.36 on p. 272 of [17].

3.1. Explicit formulae

The following theorem states our main result in this section. We assume that $\binom{r}{k} = 0$ for any $r < 0$.

Theorem 3.1. *Let* n *and* k *be positive integers. Then*

$$S_k^{(l)}(n) = \sum_{p=1}^{k} N(k,p)\binom{n-lp}{2k} \tag{5}$$

for all $l \geqslant 0$. *For* $k = 0$, *we have* $S_0^{(l)}(n) = 1$, *for all* $n \in \mathbb{N}$ *and for all* $l \geqslant 0$.

Proof. We employ the lattice-path representation of secondary structures. Consider a Dyck path P on $2k$ steps with exactly p peaks. There are $N(k,p)$ such paths. By inserting l horizontal steps between every two steps forming a peak, we get a path P' in $\mathscr{M}^{(l)}(2k+lp)$, the set of all Motzkin paths on $2k+lp$ steps with plateaus of length at least l. There are exactly $N(k,p)$ such paths. Now take additional m horizontal steps and distribute them at will in the path P'. From every path $P' \in \mathscr{M}^{(l)}(2k+lp)$ we can get $\binom{2k+m}{m}$ paths in $\mathscr{M}^{(l)}(2k+lp+m)$. Denoting the total number of steps by n, we get $N(k,p)\binom{n-lp}{2k}$ as the number of paths from $\mathscr{M}^{(l)}(n)$ which have exactly p plateaus. Summing over all p's from 1 to k, we get the number of paths in $\mathscr{M}^{(l)}(n)$ having k ascends and having plateaus of length at least l. By recalling the folklore correspondence between such Motzkin paths and secondary structures of rank l (Proposition 1.1), we get the claim of the theorem. □

For $n = 0$, we put $S_0^{(l)}(0) = 1$.

In the special case $l = 0$, we get $S_k^{(0)}(n) = C_k\binom{n}{2k}$ for all $n, k \geqslant 0$, and hence, as a corollary, the well-known relation

$$M_n = S^{(0)}(n) = \sum_{k \geqslant 0} S_k^{(0)}(n) = \sum_{k \geqslant 0} C_k \binom{n}{2k}.$$

Using this explicit formula, it is easy to prove the following result.

Corollary 3.2. *The sequence* $(S_k^{(0)}(n))_{k=0}^{n/2}$ *is log-concave and hence unimodal in* k. *The maximum value is attained for two consecutive values of* k *if* $n \equiv 2 \bmod 3$, *and is unique otherwise*.

When $l = 1$, we can express the sum in (9) in a closed form. We present here a combinatorial proof.

Proposition 3.3. $S_k^{(1)}(n) = N(n-k, k+1)$ *for all* $n, k \geqslant 0$.

Proof. We construct a bijection $\varphi_{n,k} : \mathscr{S}_k^{(1)}(n) \to \mathscr{D}_{n-2k}(n-k)$. Take a secondary structure S from $\mathscr{S}_k^{(1)}(n)$. There are $n-2k$ unpaired vertices in S. Starting from the vertex 1 and proceeding to the vertex n construct a lattice path with steps Up and Down as follows. If the vertex under consideration is a beginning vertex of an h-bond, add an Up step to the already constructed part of the path. If an h-bond ends in the considered vertex, add a Down step. Finally, if the vertex is unpaired, add a pair of steps (Up, Down) forming a peak. Obviously, the number of peaks will be equal to the number of unpaired vertices, and the total number of steps will be $2(n-2k)+2k = 2(n-k)$. It is also clear that the path will never fall below the x-axis. So, we have constructed a path in $\mathscr{D}_{n-2k}(n-k)$, and the construction is obviously bijective. The claim now follows by using the symmetry of Narayana numbers, $N(n,k) = N(n, n+1-k)$. □

Another combinatorial proof of this formula, based on the correspondence between secondary structures and linear trees, was given in [13].

As another consequence of Theorem 3.1 we have:

Corollary 3.4.

$$N(n,k) = \sum_{j \geqslant 1} N(k-1, j) \binom{n+k-j-1}{2k-2}.$$

3.2. Log-concavity results

We first examine the log-concavity behavior of $S_k^{(1)}(n)$ for a fixed n.

Theorem 3.5. *The sequence* $S_k^{(1)}(n)$ *is log-concave* (*and hence unimodal*) *in* k *for any fixed* $n \geqslant 0$. *Moreover*, *for* $n > 3$, *the maximal value is attained for unique* k.

Proof. The log-concavity of this sequence is easy. Namely, the condition $(S_k^{(1)}(n))^2 \geqslant S_{k-1}^{(1)}(n) S_{k+1}^{(1)}(n)$ is equivalent to

$$(k+2)(n-k-1)(n-2k+1)^2(n-2k+2) \geqslant k(n-k+1)(n-2k-1)^2(n-2k-2),$$

which, in turn, follows from $(k+2)(n-k-1) \geqslant k(n-k+1)$, for $n \geqslant 2k+1$.

The log concavity implies that the maximal value can be attained for at most two consecutive values of k. To show that the maximum is attained for a unique k, we begin by considering the quotient

$$\frac{S_k^{(1)}(n)}{S_{k-1}^{(1)}(n)} = \frac{(n-2k)(n-2k+1)^2(n-2k+2)}{(n-k+1)(n-k)(k+1)k}.$$

Put $a := n-2k+1$. We claim that for all positive integers a, k, with $a \geqslant 3$, the quotient

$$\frac{(a-1)a^2(a+1)}{(a+k)(a+k-1)k(k+1)} \tag{6}$$

is never equal to 1. Assume the contrary, i.e. that it is equal to 1. Then

$$k^4 + 2ak^3 + (a^2+a-1)k^2 + (a^2-a)k + a^2 - a^4 = 0 \tag{7}$$

or equivalently

$$a^4 = a^2(k^2+k+1) + ak(2k^2+k-1) + k^2(k^2-1). \tag{8}$$

Multiplying both sides of (8) by 64 and rearranging it, we get

$$64a^4 = (8ak + 8k^2 + 4a)^2 + 48a^2 - 64ak - 64k^2$$

or equivalently

$$(8a^2-3)^2=(8ak+8k^2+4a)^2+9-64ak-64k^2<(8ak+8k^2+4a)^2. \tag{9}$$

Thus,

$$(8a^2-3)^2\leqslant(8ak+8k^2+4a-1)^2.$$

On the other hand,

$$\begin{aligned}(8a^2-3)^2&=(8ak+8k^2+4a-4)^2+9-64ak-64k^2+64ak+64k^2+32a-16\\&=(8ak+8k^2+4a-4)^2+32a-7>(8ak+8k^2+4a-4)^2.\end{aligned}$$

So, the set of all possible values of the term $8a^2-3$ has only three elements, $8ak+8k^2+4a-3, 8ak+8k^2+4a-2$ and $8ak+8k^2+4a-1$. We can exclude the middle one since it is even and $8a^2-3$ is odd. Let us examine the two remaining possibilities. The equality $8a^2-3=8ak+8k^2+4a-1$ implies that $4a^2-1=4ak+4k^2+2a$, but this is impossible because of parity. The only remaining possibility is $8a^2-3=8ak+8k^2+4a-3$. Hence

$$2a^2=2ak+2k^2+a. \tag{10}$$

The left-hand side of (9) can be written in the form

$$(8a^2-3)^2=(8ak+8k^2+4a-3)^2+9-64ak-64k^2+48ak+48k^2+24a-9.$$

Hence,

$$3a=2ak+2k^2.$$

Plugging this into (10), we get $2a^2=3a+a$, and then $a=2,\ k=1$. So, there are no other solutions of Eq. (7) in the positive integers. □

Corollary 3.6. *The only solutions* (x,y) *in positive integers of the equation*

$$\binom{x}{2}\binom{x+1}{2}=\binom{y}{2}\binom{y-x+1}{2}$$

are $(1,1)$ *and* $(2,3)$.

Proof. The claim follows by expressing (6) via binomial coefficients and equating this expression to one. □

Next we examine, asymptotically, the position of the maximum value of $S_k^{(1)}(n)$. Let k_n be the value of k for which the maximal value of $S_k^{(1)}(n)$ is attained, i.e.

$$S_{k_n}^{(1)}(n)=\max_k\{S_k^{(1)}(n)\}$$

for a fixed n.

Proposition 3.7.

$$\lim_{n\to\infty}\frac{k_n}{n}=\frac{5-\sqrt{5}}{10}.$$

Proof. It is enough to find the value of k which maximizes the binomial coefficient $\binom{n-k}{k}$, because

$$S_k^{(1)}(n)=\frac{1}{k+1}\,\frac{n-2k}{n-k}\binom{n-k}{k}^2$$

and the term $(1/(k+1))((n-2k)/(n-k))$ can affect the location of the maximum by at most a constant term. The value of k maximizing the term $\binom{n-k}{k}$ is easily found to be

$$k_n=\tfrac{1}{10}\left(5n+7-\sqrt{5n^2+10n+9}\right)$$

and the claim follows. □

For higher-rank secondary structures, i.e. for $l > 1$, no closed form for the numbers $S_k^{(l)}(n)$ exists, but we can still establish their log-concavity in k. First, we need the two following lemmas.

Lemma 3.8. *Let $(a_i),(b_i)$ and (c_i) be positive, log-concave sequences, $i = 0, 1, \ldots, n$. Then*

(i) *the sequence $(a_i b_i)_{i=0}^n$ is log-concave;*
(ii) *the sequence $(C_i)_{i=0}^n$, where $C_k = \sum_{i=0}^k c_i$, is log-concave.*

Lemma 3.9. *Let $(a_n)_{n\geqslant 0}$ be a positive log-concave sequence of real numbers. Then any subsequence of (a_n) whose indices form an arithmetic progression is a log-concave sequence.*

Proof. Let $(b_m)_{m\geqslant 0}$ be such a subsequence of a positive log-concave sequence $(a_n)_{n\geqslant 0}$. Then from $b_m = a_n$ follows $b_{m+1} = a_{n+k}$, for some fixed positive integer k and all $m, n \geqslant 0$. But the log-concavity of a_n is equivalent with $a_n/a_{n-1} \geqslant a_{n+1}/a_n$. Iterating this and multiplying all inequalities $a_{n-i+1}/a_{n-i} \geqslant a_{n+i}/a_{n+i-1}$, $i = 1, \ldots, k$, we get $a_n/a_{n-k} \geqslant a_{n+k}/a_n$, and this is precisely the condition of log-concavity of the sequence b_m. □

Corollary 3.10. *The sequence $\binom{n-lj}{2k}$ is log-concave in j for all $l \geqslant 0$.*

Proof. The claim follows from the log-concavity of a column of the Pascal triangle and the previous lemma. □

Now we can prove the log-concavity of $S_k^{(l)}(n)$ in k for all $l \geqslant 0$.

Theorem 3.11. *The sequence $S_k^{(l)}(n)$ is log-concave in k for all $n \geqslant 0$ and $l \geqslant 0$.*

Proof. For the case $l=0$ the claim follows by straightforward computation. The case $l=1$ has been proven in Theorem 3.5, and for $l > 1$ the claim follows by applying Lemma 3.8 to the sequences $a_j = N(k, j)$, $b_j = \binom{n-lj}{2k}$ and $c_j = a_j b_j$. □

3.3. Hypergeometric representations

Although no simple closed form expression for $S_k^{(l)}(n)$ exists for $l \geqslant 2$, we can obtain a representation of $S_k^{(l)}(n)$ in terms of hypergeometric functions.

Theorem 3.12. *For all $n, l \geqslant 0$, $0 \leqslant k \leqslant \lfloor (n-l)/2 \rfloor$, we have*

$$S_k^{(l)}(n) = \binom{n-l}{2k} {}_{l+2}F_{l+1}\left(\begin{matrix} 1-k, -k, \vec{p}_l \\ 2, \vec{q}_l \end{matrix} \middle| \, 1 \right),$$

where

$$(\vec{q}_l)_i = \frac{l-n+i-1}{l}, \quad (\vec{p}_l)_i = (\vec{q}_l)_i + \frac{2k}{l}, \quad i = 1, \ldots, l.$$

Proof. The claim follows by considering the quotients of successive terms in the sum $\sum_{p=1}^k N(k,p)\binom{n-lp}{2k}$ and reading off the parameters of hypergeometric series. □

For $l = 0$, we get

$$S_k^{(0)}(n) = \binom{n}{2k} {}_2F_1\left(\begin{matrix} 1-k, -k \\ 2 \end{matrix} \middle| \, 1 \right)$$

and then

$${}_2F_1\left(\begin{matrix} 1-k, -k \\ 2 \end{matrix} \middle| \, 1 \right) = C_k$$

for $k \geqslant 0$. From there we get a representation of Motzkin numbers in terms of hypergeometric functions.

Proposition 3.13.

$$M_n = {}_2F_1\left(\begin{matrix}\frac{1-n}{2}, -\frac{n}{2}\\ 2\end{matrix}\,\middle|\, 4\right)$$

for $n \geqslant 0$.

Proof. We start from the equality $M_n = \sum_{k\geqslant 0} S_k^{(0)}(n)$. Plugging in the explicit values for $S_k^{(0)}(n)$, considering the quotients of successive terms and reading off the parameters of the corresponding hypergeometric function yields the formula. □

By repeating the same procedure with the explicit formulae for $S_k^{(1)}(n)$, we get the following result.

Proposition 3.14.

$$S^{(1)}(n) = {}_4F_3\left(\begin{matrix}1-\frac{n}{2}, \frac{1}{2}-\frac{n}{2}, \frac{1}{2}-\frac{n}{2}, -\frac{n}{2}\\ 2, -n, -n+1\end{matrix}\,\middle|\, 16\right).$$

The explicit formula from Proposition 3.3 can be easily obtained from this result via the Saalschütz identity.

Note now that the elements of the parameter vectors $\vec{p}_l$ and $\vec{q}_l$ differ by a constant value, $2k/l$. For the values $l = 1$ and 2 this difference will be an integer, $2k$ and k, respectively. We can use this to reduce the number of parameters in our hypergeometric functions.

Proposition 3.15.

$${}_3F_2\left(\begin{matrix}1-k, -k, 1-n+2k\\ 2, 1-n\end{matrix}\,\middle|\, z\right) = \frac{z^n}{(1-n)^{\overline{2k}}}\,\frac{\mathrm{d}^{2k}}{\mathrm{d}z^{2k}}\left[z^{2k-n}\,{}_2F_1\left(\begin{matrix}1-k, -k\\ 2\end{matrix}\,\middle|\, z\right)\right],$$

$${}_4F_3\left(\begin{matrix}1-k, -k, 1-\frac{n}{2}+k, \frac{3}{2}-\frac{n}{2}+k\\ 2, 1-\frac{n}{2}, \frac{3}{2}-\frac{n}{2}\end{matrix}\,\middle|\, z\right)$$

$$= \frac{z^{n/2}}{(1-n/2)^{\bar{k}}\left(\frac{3}{2}-n/2\right)^{\bar{k}}}\,\frac{\mathrm{d}^k}{\mathrm{d}z^k}\left[z^{k-1/2}\,\frac{\mathrm{d}^k}{\mathrm{d}z^k}\left[z^{k+(1-n)/2}\,{}_2F_1\left(\begin{matrix}1-k, -k\\ 2\end{matrix}\,\middle|\, z\right)\right]\right].$$

Here $x^{\overline{m}}$ *denotes the rising factorial,* $x^{\overline{m}} = x(x+1)\ldots(x+m-1)$ [3, p. 48].

Proof. The first result follows by applying formula (5) from [7, p. 169]. The second result follows by using the same formula twice. □

Proposition 3.16.

$${}_2F_1\left(\begin{matrix}1-k, -k\\ 2\end{matrix}\,\middle|\, z\right) = \frac{2(1-z)^{k-1}}{k(k+1)}\,\frac{\mathrm{d}}{\mathrm{d}z'}\,P_k(z'),$$

where P_k *is the* k*th Legendre polynomial, and* $z' = (1+z)/(1-z)$.

Proof. First apply formula (9) in [7, p. 273]

$$\frac{\mathrm{d}^j}{\mathrm{d}z^j}\left[(1-z)^{a+j-1}\,{}_2F_1\left(\begin{matrix}a, b\\ c\end{matrix}\,\middle|\, z\right)\right] = (-1)^j\,\frac{a^{\bar{j}}(c-b)^{\bar{j}}}{c^{\bar{j}}}(1-z)^{a-1}\,{}_2F_1\left(\begin{matrix}a+j, b\\ c+j\end{matrix}\,\middle|\, z\right)$$

with $a = b = -k$, $c = 1$ and $j = 1$ to get

$${}_2F_1\left(\begin{matrix}1-k, -k\\ 2\end{matrix}\,\middle|\, z\right) = \frac{(1-z)^{k+1}}{k(k+1)}\,\frac{\mathrm{d}}{\mathrm{d}z}\left[(1-z)^{-k}\,{}_2F_1\left(\begin{matrix}-k, -k\\ 1\end{matrix}\,\middle|\, z\right)\right]$$

and then use the fact that

$$ {}_2F_1\left(\begin{matrix}-k,-k\\ 1\end{matrix}\,\middle|\, z\right)=(1-z)^k P_k\left(\frac{1+z}{1-z}\right) \qquad [12,\text{p. 466}]. \qquad \square $$

Now we can express $S_k^{(1)}(n)$ and $S_k^{(2)}(n)$ in terms of Legendre polynomials.

Proposition 3.17. *For all $n\geqslant 0$, $k\leqslant\lfloor(n-l)/2\rfloor$, $l=1,2$, we have*

$$ S_k^{(1)}(n)=\frac{1}{P_k'(1)(2k)!}\frac{\mathrm{d}^{2k}}{\mathrm{d}z^{2k}}\left[z^{2k-n}(1-z)^{k-1}P_k'\left(\frac{1+z}{1-z}\right)\right]\Bigg|_{z=1} $$

and

$$ S_k^{(2)}(n)=\frac{2^{2k}}{P_k'(1)(2k)!}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\left[z^{k-\frac{1}{2}}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\left[z^{k+(1-n)/2}(1-z)^{k-1}P_k'\left(\frac{1+z}{1-z}\right)\right]\right]\Bigg|_{z=1}. $$

Proof. The claim follows by combining Propositions 3.15 and 3.16. $\square$

3.4. Bivariate generating functions

The bivariate generating function of numbers $S_k^{(0)}(n)$ has been implicitly stated and studied in [9], and its generalization to the case $l>0$ is discussed in [10]. We derive it here explicitly, starting from the bivariate generating function for Narayana numbers [17, p. 238],

$$ N(x,t)=\sum_{n,k\geqslant 1}N(n,k)x^n t^k=\frac{1}{2x}\left[1-x-xt-\sqrt{(1-x-xt)^2-4x^2t}\right]. $$

Let us denote by $\tilde{S}_l(x,y)$ the bivariate generating function for the numbers $S_k^{(l)}(n)$, $n,k\geqslant 1$.

$$ \begin{aligned}\tilde{S}_l(x,y)&=\sum_{n,k\geqslant 1}S_k^{(l)}(n)x^n y^k=\sum_{n,k,p\geqslant 1}N(k,p)\binom{n-lp}{2k}x^n y^k\\ &=\sum_{k,p\geqslant 1}N(k,p)y^k\sum_{n\geqslant 1}\binom{n-lp}{2k}x^n=\sum_{k,p\geqslant 1}N(k,p)y^k\frac{x^{2k+lp}}{(1-x)^{2k+1}}\\ &=\frac{1}{1-x}\sum_{k,p\geqslant 1}N(k,p)\left[\frac{x^2y}{(1-x)^2}\right]^k[x^l]^p=\frac{1}{1-x}N(u,v),\end{aligned} $$

where $u=x^2y/(1-x)^2$, $v=x^l$.

In order to get the bivariate generating function for all $S_k^{(l)}(n)$, we add the generating function of the zeroth column, $S_{l,0}(x)=\sum_{n\geqslant 0}S_0^{(l)}(n)x^n=\sum_{n\geqslant 0}x^n=1/(1-x)$.

Theorem 3.18. *The bivariate generating function of numbers $S_k^{(l)}(n)$ is given by*

$$ S_l(x,y)=\frac{1}{2x^2y}\left[\Omega_l(x,y)-\sqrt{\Omega_l^2(x,y)-4x^2y}\right], $$

where

$$ \Omega_l(x,y)=(1-x)(1-y)-y\omega_l(x), $$

$$ \omega_l(x)=x-1-x^2\frac{1-x^l}{1-x}. $$

Let $E_l(n)$ be the expected number of h-bonds in a secondary structure of size n and rank l. We can express $E_l(n)$ in terms of the bivariate generating function in the following way [15]:

$$E_l(n) = \frac{[x^n](\partial S_l(x,y)/\partial y)|_{y=1}}{[x^n]S_l(x,y)|_{y=1}}. \tag{11}$$

Starting from formula (11) and using Darboux's theorem, one can derive the following result.

Theorem 3.19. *Let $E_l(n)$ be the expected number of h-bonds in a secondary structure of size n and rank l. Then, for $n \to \infty$*

$$E_l(n) \sim f_l(r_l)n,$$

where

$$f_l(x) = 2\,\frac{(x-1)\omega_l(x) - 2x^2}{(-\omega_l(x)+2x)w_l(x)},$$

$$w_l(x) = \frac{-\omega_l(x) - 2x}{1 - x/r_l},$$

$$\omega_l(x) = x - 1 - x^2\,\frac{1-x^l}{1-x}$$

and r_l is the smallest positive root of the equation $-\omega_l(x) - 2x = 0$, i.e. the positive root of the equation $x^{l/2+1} + 2x - 1 = 0$.

Corollary 3.20.

$$E_0(n) \sim \frac{n}{3},$$

$$E_1(n) \sim \frac{5-\sqrt{5}}{10}\,n,$$

$$E_2(n) \sim \frac{n}{4}.$$

Although the result for the case $l = 0$ is not very interesting in the context of secondary structures, we can rephrase it in terms of Motzkin paths.

Corollary 3.21. *The average number of* Up *steps in a Motzkin path of length n behaves as $n/3$ when $n \to \infty$.*

For the higher values of l, the expressions $f_l(r_l)$ become very complicated. The approximative values for $l = 3$, 4, and 5 are 0.236738, 0.23036, and 0.227971, respectively.

Acknowledgements

We are indebted to the referee for bringing to our attention the result of Proposition 2.3 and for other helpful suggestions.

References

[1] M. Aigner, Motzkin numbers, European J. Combin. 19 (1998) 663–675.
[2] D. Callan, Notes on Motzkin and Schröder numbers, 2000, preprint.
[3] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics, Addison-Wesley, Reading, 1988.
[4] C. Haslinger, P.F. Stadler, RNA structures with pseudo-knots: graph-theoretical, combinatorial and statistical properties, Bull. Math. Biol. 61 (1999) 437–467.
[5] I.L. Hofacker, P. Schuster, P.F. Stadler, Combinatorics of RNA secondary structures, Discrete Appl. Math. 88 (1998) 207–237.
[6] J. Kruskal, D. Sankoff, Time Warps, String Edits and Macromolecules, 2nd Edition, Addison-Wesley, Reading, 1999.
[7] Y.L. Luke, Mathematical Functions and Their Approximations, Academic Press, New York, 1975.

[8] T.V. Narayana, Lattice Path Combinatorics with Statistical Applications, Toronto University Press, Toronto, 1979.
[9] M. Nebel, Combinatorial properties of RNA secondary structures, J. Comput. Biol. 9 (2001) 541–573.
[10] M. 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.
[11] A.M. Odlyzko, Asymptotic enumeration methods, in: R. Graham, M. Grötschel, L. Lovász (Eds.), Handbook of Combinatorics, Vol. 2, Elsevier, Amsterdam, 1995.
[12] A.P. Prudnikhov, Yu.A. Brychkov, O.I. Marichev, Integrals and Series, Additional Chapters, Gordon and Breach, New York, 1998.
[13] W.R. Schmidt, M.S. Waterman, Linear trees and RNA secondary structure, Discrete Appl. Math. 51 (1994) 317–323.
[14] P. Schuster, P.F. Stadler, Discrete models of biopolymers, in: J. Crabbe, A. Konopka, M. Drew (Eds.), Handbook of Computational Chemistry and Biology, Marcel Dekker Inc., in press.
[15] R. Sedgewick, P. Flajolet, Introduction to the Analysis of Algorithms, Addison-Wesley, Reading, 1996.
[16] F. Soler, K. Jankowski, Modeling RNA secondary structures I, Math. Biosci. 105 (1991) 167–191.
[17] R.P. Stanley, Enumerative Combinatorics, Vol. 2, Cambridge University Press, Cambridge, 1999.
[18] P.R. Stein, M.S. Waterman, On some new sequences generalizing the Catalan and Motzkin numbers, Discrete Math. 26 (1979) 261–272.
[19] D. Veljan, On some primary and secondary structures in combinatorics, Math. Commun. 6 (2001) 217–232.
[20] G. Viennot, M. Vauchaussade de Chaumont, Enumeration of RNA secondary structures by complexity, Math. Med. Biol. Lecture Notes Biomath. 57 (1985) 360–365.
[21] M.S. Waterman, Secondary structures of single stranded nucleic acids, in: G.C. Rota (Ed.), Studies on Foundations and Combinatorics. Advances in Mathematics Supplementary Studies, Vol. I, Academic Press, New York, 1978, pp. 167–212.
[22] D.B. West, Introduction to Graph Theory, Prentice-Hall, Upper Saddle River, 1996.