2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 389–400

Minimal and maximal plateau lengths in Motzkin paths

Helmut Prodinger[1†] and Stephan Wagner[2]

[1] *Stellenbosch University, Department of Mathematics, 7602 Stellenbosch, South Africa.* hproding@sun.ac.za

[2] *Stellenbosch University, Department of Mathematics, 7602 Stellenbosch, South Africa.* swagner@sun.ac.za

The minimal length of a plateau (a sequence of horizontal steps, preceded by an up- and followed by a down-step) in a Motzkin path is known to be of interest in the study of secondary structures which in turn appear in mathematical biology. We will treat this and the related parameters *maximal plateau length*, *minimal horizontal segment* and *maximal horizontal segment* as well as some similar parameters in unary-binary trees by a pure generating functions approach—Motzkin paths are derived from Dyck paths by a substitution process. Furthermore, we provide a pretty general analytic method to obtain means and limiting distributions for these parameters. It turns out that the maximal plateau and the maximal horizontal segment follow a Gumbel distribution.

Keywords: Motzkin paths, singularity analysis, Mellin transform, bootstrapping, unary-binary trees, Gumbel distribution

1 Introduction

A *Motzkin path* of length n in the (x, y)-plane from $(0, 0)$ to $(n, 0)$ consists of steps $(1, 1)$ ("up"), $(1, -1)$ ("down"), and $(1, 0)$ ("level"), with the restriction that a Motzkin path must never go below the x-axis.

A *plateau* of length k is a sequence of k consecutive level steps, preceded by an up-step, and followed by a down-step. Let us denote $\mathcal{M}^k$ the set of all Motzkin paths where every plateau is at least k steps long. Note that $\mathcal{M}^0$ is the set of *all* Motzkin paths. It has been described in [DSV04, Neb01, VdC85] that the family $\mathcal{M}^k$ is of relevance in the study of *secondary structures*, and thus in turn in *mathematical biology.*

RNA-molecules can be represented as linear chains of four bases A, C, G, U, that are connected by *p-bonds*; this is known as the *primary structure* of a molecule. However, there is a second kind of bonds, so-called *h-bonds*, between pairs of bases in the chain. These follow three rules: each base can only participate in at most one h-bond; h-bonds may not cross; and the bases that form an h-bond may not be too close to each other in the primary structure. Together with these h-bonds, the linear chain of bases forms the *secondary structure*—Figure 1 shows an example.

Now it is known that there is a bijection between secondary structures and Motzkin paths, see again Figure 1. Here, the distance between every two bases that are connected by an h-bond is $> k$ if and only

†This material is based upon work supported by the National Research Foundation under grant number 2053748

Fig. 1: A secondary structure and its associated Motzkin path.

if the corresponding Motzkin path is in $\mathcal{M}^k$; this bijection has been used for enumeration purposes in the papers [DSV04, Neb01, VdC85].

We are led in a natural way to a *random variable* MinPlateau which assigns to each Motzkin path the *minimal plateau length*. We show how to study this parameter, by a pure use of generating functions and asymptotic techniques, without using any recursions, as for instance in [DSV04]. Thus, our approach is very much in line with the subtitle *Analytic Combinatorics* of this conference and the forthcoming book [FS07].

From a mathematical point of view, it is at least as interesting to study the random variable MaxPlateau, which assigns to each Motzkin path the *maximal plateau length.* Analytically, this is more challenging, as it resembles parameters like *height* or *maximal run length*, and the behaviour of the average is more intricate (we will see that MinPlateau = 0 for almost all Motzkin paths).

Furthermore, one can drop the restriction that the sequence of level steps must be rendered by an up- resp. down-step. Considering the respective lengths of *all* (maximal) sequences of horizontal steps, we are led to the parameters MinLevel and MaxLevel.

A related concept are unary-binary-trees. They are defined in a recursive way by saying that the empty tree is a unary-binary-tree, and that a root followed by one unary-binary-tree (a unary node), or a root followed by a left and a right unary-binary-tree (a binary node) are again unary-binary trees. They are a special instance of the *simply generated families of trees*, provided that one considers a leaf (empty node) as an internal node as well, see [MM78].

The analysis of the aforementioned parameters can be performed within a very general framework, which will be described in Section 4. It turns out that the average of MaxLevel and MaxPlateau is of logarithmic order, with fluctuating terms of lower order.

In order to obtain generating functions for our parameters, we use a substitution technique—Motzkin paths, for instance, can be obtained from *Dyck paths* (which contain no level steps) by the following substitution: each up-/down-step is replaced by an up-/down-step, followed by an arbitrary number of level steps. Additionally, one allows for an arbitrary number of level steps at the beginning of the paths. This is easy to model using generating functions—if one wants the parameter MaxLevel to be $\leq k$ for instance, one allows only substitutions with horizontal segments of length $\leq k$; likewise, if MinLevel should be $\geq k$, the substitution must use only horizontal segments of length $\geq k$. For the parameters related to the *plateau*, this can also be achieved, since one can easily set up a generating function for Dyck paths where an up-step followed by a down-step (a "peak") gets a special label.

In the case of unary-binary trees, there is also a substitution that produces them starting from binary trees (each leaf resp. internal node can be followed by an arbitrary sequence of unary nodes). In analogy to the Motzkin paths, one could be interested in the minimal resp. maximal lengths of these chains of unary nodes. To take the analogy further, the *plateau* instance would then correspond to the unary chains that grow out of the leaves (and not the other nodes).

The approximations that will be derived later in this paper exhibit the *Gumbel distribution* (or *extreme*

value distribution) as limiting distribution for the parameters MaxPlateau and MaxLevel (and the analogous parameters in unary-binary trees). The Gumbel distribution is usually defined via its distribution function

$$F(x) = e^{-e^{-x}}.$$

As shown in [LP06], *all* moments can be evaluated asymptotically in a semi-automatic fashion. Here we confine ourselves to give the averages in an explicit form.

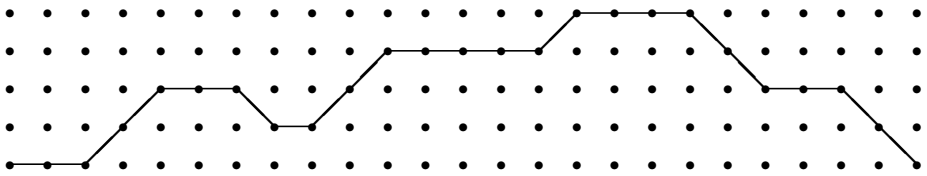

Fig. 2: A Motzkin path with MinPlateau = 2, MaxPlateau = 3, MinLevel = 1 and MaxLevel = 4.

2 Substitutions and generating functions

In the following, let us denote the generating function for Catalan paths by $C(z)$:

$$C(z) = \frac{1 - \sqrt{1-4z^2}}{2z^2}.$$

Furthermore, let $D(z,u)$ be the bivariate generating function for Catalan paths, where each peak is marked by uz instead of z^2. If the empty path is ignored, it follows that

$$D(z,u) = \sum_{k\geq 1} \Big(zu + z^2 D(z,u)\Big)^k = \frac{zu + z^2 D(z,u)}{1 - zu - z^2 D(z,u)}$$

or

$$D(z,u) = \frac{1 - zu - z^2 - \sqrt{1 - 2zu - 2z^2 + z^2u^2 - 2z^3u + z^4}}{2z^2}.$$

Now, the generating functions for our problems follow easily from the aforementioned substitution process. For instance, if we are interested in the generating function for Motzkin paths with the additional requirement that each plateau has length at least k, we have to substitute u by $\frac{z^{k+1}}{1-z}$ (and z by $\frac{z}{1-z}$) to obtain

$$\frac{1}{1-z} D\Big(\frac{z}{1-z}, \frac{z^{k+1}}{1-z}\Big) = \frac{1 - 2z - z^{k+2} - \sqrt{1 - 4z + 4z^2 - 2z^{k+2} + 4z^{k+3} - 4z^{k+4} + z^{2k+4}}}{2z^2(1-z)}.$$

Analogously, we obtain the following generating function for Motzkin paths where MaxPlateau is $\leq k$:

$$\frac{1}{1-z} D\Big(\frac{z}{1-z}, \frac{z(1-z^{k+1})}{1-z}\Big) = \frac{1 - 2z - z^2 + z^{k+3} - \sqrt{(1 - z^2 + z^{k+3})(1 - 4z + 3z^2 + z^{k+3})}}{2z^2(1-z)}.$$

The situation is even simpler in the case of the parameters MinLevel and MaxLevel: here, we just have to replace z in the generating function C. Hence, if we want MinLevel to be $\geq k$, we obtain

$$\Big(1+\frac{z^k}{1-z}\Big)C\Big(z\Big(1+\frac{z^k}{1-z}\Big)\Big)-C(z) \\ = \frac{1-z-\sqrt{(1-z)^2(1-4z^2)-4z^{k+2}(2-2z+z^k)}}{2z^2(1-z+z^k)}-\frac{1-\sqrt{1-4z^2}}{2z^2}.$$

Note that we have to allow level steps of length 0 (yielding the summand 1 in the above expression), and that we have to subtract the number of Catalan paths (which contain no level steps at all). Finally,

$$\frac{1-z^{k+1}}{1-z}\,C\Big(\frac{z(1-z^{k+1})}{1-z}\Big)=\frac{1-z-\sqrt{1-2z-3z^2+8z^{k+3}-4z^{2k+4}}}{2z^2(1-z^{k+1})}$$

is the generating function for Motzkin paths with MaxLevel $\leq k$.

3 The parameters MinPlateau and MinLevel in Motzkin paths

From an analytic point of view, these two parameters are simpler and less interesting, since MinPlateau is 0 for almost all Motzkin paths, whereas MinLevel is equal to 1. Hence, the mean also tends to these values. We will show this in detail in the case of plateaus only, the other case being analogous. Note that the generating function for Motzkin paths with MinPlateau ≥ 1 is given by

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

A simple application of the Flajolet-Odlyzko singularity analysis [FO90] shows that the number of Motzkin paths with n steps and the property that each plateau has length ≥ 1 is asymptotically given by

$$\sqrt{\frac{15+7\sqrt{5}}{8\pi}}\cdot n^{-3/2}\cdot\left(\frac{3+\sqrt{5}}{2}\right)^n,$$

whereas the overall number of Motzkin paths is given by the Motzkin numbers, which are asymptotically equal to

$$\sqrt{\frac{27}{4\pi}}\cdot n^{-3/2}\cdot 3^n.$$

Hence, MinLevel equals 0 for almost all Motzkin paths. These asymptotics have been given in [DSV04] as well. Now, in order to determine an average, we have to consider the sum

$$\sum_{k\geq 1}\frac{1-2z-z^{k+2}-\sqrt{1-4z+4z^2-2z^{k+2}+4z^{k+3}-4z^{k+4}+z^{2k+4}}}{2z^2(1-z)}.$$

We will show that the sum over all $k \geq 2$ yields a function which is analytic within the open circle of radius $\sqrt{2}-1$ around the origin, thus proving that the average of MinPlateau is asymptotically equal to

$$\sqrt{\frac{15+7\sqrt{5}}{54}} \cdot \left(\frac{3+\sqrt{5}}{6}\right)^n,$$

which tends to 0.

Since the coefficients of the Taylor series around 0 are positive for all summands, it suffices to prove that it is convergent for $z = \sqrt{2}-1$, implying absolute convergence for all values of z within the aforementioned circle. Denote the sum over all $k \geq 2$ by $S(z)$. Then we have

$$\begin{aligned} S(\sqrt{2}-1) &= \sum_{k\geq 2} \frac{(2+\sqrt{2})\Big(1-(\sqrt{2}-1)^k - \sqrt{1-6(\sqrt{2}-1)^k+(\sqrt{2}-1)^{2k}}\Big)}{4} \\ &\leq \sum_{k\geq 2} \frac{(2+\sqrt{2})\cdot 2(1+\sqrt{2})(\sqrt{2}-1)^k}{4} \\ &= \frac{\sqrt{2}+1}{2} < \infty, \end{aligned}$$

which finishes the proof.

In order to obtain the mean for MinLevel, one has to consider the sum

$$\sum_{k\geq 1}\left(\frac{1-z-\sqrt{(1-z)^2(1-4z^2)-4z^{k+2}(2-2z+z^k)}}{2z^2(1-z+z^k)} - \frac{1-\sqrt{1-4z^2}}{2z^2}\right).$$

In the same way as before, we see that the sum over $k \geq 2$ is analytic within the open circle of radius 0.396608 (the dominant singularity of the summand that corresponds to $k=2$) around the origin, and that the average of MinLevel thus tends to 1.

4 The parameters MaxPlateau and MaxLevel in Motzkin paths

Looking at the generating functions for these two parameters, we see that they have essentially the same form, namely $\frac{P(z,z^k)-\sqrt{R(z,z^k)}}{Q(z,z^k)}$ for some polynomials, where the dominating singularity (the smallest zero of $R(z,z^k)$) is decreasing in k and tending to a limit. Hence, rather than treat the two cases separately, we provide a general theorem and apply is to the two parameters as well as two similar parameters for unary-binary trees (in the subsequent section). The following lemma gives the essential asymptotic formula:

Lemma 1 *Let a generating function $f_k(z)$ be given by*

$$f_k(z) = \frac{P(z,z^k)-\sqrt{R(z,z^k)}}{Q(z,z^k)},$$

where $P(z,u), Q(z,u), R(z,u)$ are polynomials satisfying the following conditions:

- $r_0(z) := R(z,0)$ *has a simple positive real root* $\rho < 1$ *and no other roots in* $\{z : |z| \leq \rho\}$, $r_0(z)$ *is positive for* $z < \rho$ *and* $r_1(z) := R_u(z,0)$ *is positive for* $z = \rho$,
- $Q(z, z^k)$ *has no roots in* $\{z : |z| \leq \rho + c_1\}$ *for some positive constant* c_1 *and sufficiently large* $k \geq k_1$.

Then, the following asymptotic formula holds for all $k \geq k_0$*:*

$$\frac{[z^n]f_k(z)}{[z^n]f_\infty(z)} = \exp(-\delta n \rho^k)\big(1 + O(n^{-2} + k\rho^k + k^2\rho^k n^{-1} + k\rho^{2k}n)\big),$$

where $f_\infty := \frac{P(z,0)-\sqrt{R(z,0)}}{Q(z,0)}$ *and* $\delta = -\frac{r_1(\rho)}{\rho r_0'(\rho)} > 0$. *The implied constant doesn't depend on* k.

Remark 4.1 *Note that the O-term, in particular* $k\rho^{2k}n$*, is not small if* $k \leq -\frac{\log n}{2\log\rho} + \cdots$*; however, the asymptotic estimate still holds, i.e. we have*

$$\frac{[z^n]f_k(z)}{[z^n]f_\infty(z)} \ll \exp(-\delta n\rho^k) \cdot k\rho^{2k}n$$

in this case, which will be sufficient for the proof of our main theorem (in fact, the proof of this lemma shows that the estimate can even be refined).

Proof: First, we follow the lines of [HP03]: for suitable $0 < C < 1 - \rho$, there is no other root of r_0 than ρ inside the disk $\{z : |z| \leq \rho + 2C\}$, and we have

$$|R(z, z^k) - R(z,0)| = O((\rho + C)^k) < |R(z,0)|$$

for $|z| = \rho + C$ and $k \geq k_2$. By Rouché's Theorem, we conclude that $R(z, z^k)$ has exactly one simple root in the disk $\{z : |z| \leq \rho + C\}$ for sufficiently large k. Since

$$\operatorname{sign}(R(\rho, \rho^k)) = \operatorname{sign}(R_u(\rho, 0)) = \operatorname{sign}(r_1(\rho)) = 1$$

and

$$\operatorname{sign}(R(\rho + \tfrac{1}{k}, (\rho + \tfrac{1}{k})^k)) = \operatorname{sign}(R(\rho + \tfrac{1}{k}, 0)) = -1$$

for sufficiently large k, there must be a real root $\rho_k := \rho + \epsilon_k$ of $R(z, z^k)$ with $0 < \epsilon_k < \frac{1}{k}$. Applying the well-known bootstrapping method (compare Knuth [Knu78] for instance), we find that $\epsilon_k = O(\rho^k)$ and more precisely

$$\epsilon_k = -\frac{r_1(\rho)}{r_0'(\rho)}\rho^k + \frac{r_1(\rho)^2}{r_0'(\rho)^2}k\rho^{2k-1} + O(\rho^{2k}) = \delta\rho^{k+1}(1 + \delta k\rho^k + O(\rho^k)).$$

Now, note that we can write our generating function as

$$f_k(z) = \frac{P(z, z^k) - \sqrt{(1 - \frac{z}{\rho_k})s_k(z)}}{Q(z, z^k)}$$

for some polynomial s_k, and that ρ_k is the only singularity of f_k within the disk $\{z : |z| \leq \rho + \min(c_1, C)\}$. We expand f_k around ρ_k:

$$f_k(z) = \frac{p_k(z)}{q_k(z)} - \frac{\sqrt{s_k(\rho_k)}}{q_k(\rho_k)}\left(1 - \frac{z}{\rho_k}\right)^{1/2} + \frac{\rho_k\sqrt{s_k(\rho_k)}}{q_k(\rho_k)}\left(\frac{s_k'(\rho_k)}{2s_k(\rho_k)} - \frac{q_k'(\rho_k)}{q_k(\rho_k)}\right)\left(1 - \frac{z}{\rho_k}\right)^{3/2} + \cdots,$$

where $p_k(z) := P(z, z^k)$ and $q_k(z) := Q(z, z^k)$.

Applying the Flajolet-Odlyzko singularity analysis [FO90], we see that

$$[z^n]f_k(z) = \frac{\sqrt{s_k(\rho_k)}}{2\sqrt{\pi}q_k(\rho_k)} \cdot n^{-3/2} \cdot \rho_k^{-n}\left(1 + \left(\frac{3}{8} + \frac{3\rho_k}{2}\left(\frac{s_k'(\rho_k)}{2s_k(\rho_k)} - \frac{q_k'(\rho_k)}{q_k(\rho_k)}\right)\right)n^{-1} + O(n^{-2})\right)$$

holds uniformly in k. Furthermore, note that

$$\begin{aligned} s_k(\rho_k) &= -\rho_k \frac{d}{dz}R(z, z^k)\Big|_{z=\rho_k} \\ &= -\rho_k\big(R_z(\rho_k, \rho_k^k) + k\rho_k^{k-1}R_u(\rho_k, \rho_k^k)\big) \\ &= -\rho_k\big(R_z(\rho, 0)(1 + O(\epsilon_k + \rho_k^k)) + k\rho_k^{k-1}R_u(\rho, 0)(1 + O(\epsilon_k + \rho_k^k))\big) \\ &= -\rho R_z(\rho, 0)\big(1 + O(k\rho^k)\big) \\ &= -\rho r_0'(\rho)\big(1 + O(k\rho^k)\big) \end{aligned}$$

and similarly

$$\begin{aligned} s_k'(\rho_k) &= -\frac{\rho}{2}r_0''(\rho)\big(1 + O(k^2\rho^k)\big), \\ q_k(\rho_k) &= Q(\rho, 0)\big(1 + O(\rho^k)\big), \\ q_k'(\rho_k) &= Q_z(\rho, 0)\big(1 + O(k\rho^k)\big), \end{aligned}$$

so that we have

$$\begin{aligned} [z^n]f_k(z) = \frac{\sqrt{-\rho r_0'(\rho)}}{2\sqrt{\pi}Q(\rho, 0)} &\cdot n^{-3/2} \cdot \rho_k^{-n} \\ &\times \left(1 + \left(\frac{3}{8} + \frac{3\rho}{2}\left(\frac{r_0''(\rho)}{4r_0'(\rho)} - \frac{Q_z(\rho, 0)}{Q(\rho, 0)}\right)\right)n^{-1} + O(n^{-2} + k\rho^k + k^2\rho^k n^{-1})\right). \end{aligned}$$

Applying the singularity analysis to f_∞ yields, on the other hand,

$$[z^n]f_\infty(z) = \frac{\sqrt{-\rho r_0'(\rho)}}{2\sqrt{\pi}Q(\rho, 0)} \cdot n^{-3/2} \cdot \rho^{-n}\left(1 + \left(\left(\frac{3}{8} + \frac{3\rho}{2}\left(\frac{r_0''(\rho)}{4r_0'(\rho)} - \frac{Q_z(\rho, 0)}{Q(\rho, 0)}\right)\right)n^{-1} + O(n^{-2})\right),$$

from which we obtain

$$\frac{[z^n]f_k(z)}{[z^n]f_\infty(z)} = \left(\frac{\rho}{\rho_k}\right)^n\Big(1 + O(n^{-2} + k\rho^k + k^2\rho^k n^{-1})\Big),$$

and the statement of the lemma follows from the fact that

$$\left(\frac{\rho}{\rho_k}\right)^n = \left(1+\frac{\epsilon_k}{\rho}\right)^{-n} = \left(\exp\left(\delta\rho^k + \delta^2 k\rho^{2k} + O(\rho^{2k})\right)\right)^{-n},$$

where we are making use of the fact that the error term $\delta^2 k\rho^{2k} + O(\rho^{2k})$ is always positive if k_0 is taken large enough. □

The asymptotic formula of this lemma can now be used to deduce information about the behaviour of the corresponding random variable:

Theorem 4.2 *With the notation of Lemma 1, suppose that X_n is a random variable such that $P(X_n \leq k) = \frac{[z^n]f_k(z)}{[z^n]f_\infty(z)}$. Then X_n asymptotically follows a Gumbel distribution, and the mean of X_n is given by*

$$E(X_n) = \log_a n + \log_a \delta + \frac{\gamma}{\log a} + \frac{1}{2} + \phi(\log_a n + \log_a \delta) + O\left(\frac{\log^2 n}{n}\right), \tag{4.1}$$

where $a = \rho^{-1}$, $\delta = -\frac{r_1(\rho)}{\rho r_0'(\rho)}$, and ϕ is the periodic function that is defined by the Fourier series

$$-\frac{1}{\log a}\sum_{k\neq 0}\Gamma\left(-\frac{2k\pi i}{\log a}\right)e^{2k\pi i x}. \tag{4.2}$$

Remark 4.3 *If a is not too large, ϕ is a function of very small amplitude—for $a = 3$ (the case that is of interest for Motzkin paths), it is approximately $2.4 \cdot 10^{-4}$. Figure* **??** *shows a plot of ϕ in this case.*

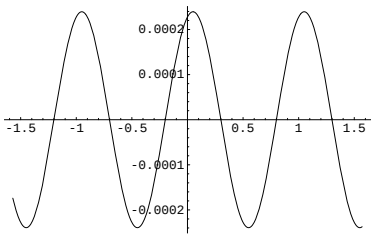

Fig. 3: Plot of the function ϕ in the case $a = 3$.

Proof: In order to obtain the mean of X_n, we have to determine the sum

$$E(X_n) = \sum_{k\geq 0} P(X_n > k) = \sum_{k\geq 0}(1 - P(X_n \leq k)) = \sum_{k\geq 0}\left(1 - \frac{[z^n]f_k(z)}{[z^n]f_\infty(z)}\right).$$

In view of the structure of the generating functions, we have $\frac{[z^n]f_k(z)}{[z^n]f_\infty(z)} = 1$ for $k > n$, hence the formula reduces to

$$E(X_n) = \sum_{k=0}^{n}\left(1 - \frac{[z^n]f_k(z)}{[z^n]f_\infty(z)}\right).$$

Now, let us replace the summand $1 - \frac{[z^n]f_k(z)}{[z^n]f_\infty(z)}$ by the simpler expression

$$1 - \exp(-\delta n \rho^k)$$

and estimate the error term. Note first that we have

$$\begin{aligned}\big|P(X_n \le k) - \exp(-\delta n\rho^k)\big| &\le P(X_n \le k) + \exp(-\delta n\rho^k)\\ &\le P(X_n \le k_0) + \exp(-\delta n\rho^{k_0})\\ &\ll n\exp(-\delta n\rho^{k_0})\end{aligned}$$

for $k \le k_0$, so the error term that comes from these summands is exponentially small and thus negligible. Furthermore, $1 - \exp(-\delta n\rho^k)$ is exponentially small for $k > n$. So we obtain, by Lemma 1,

$$\begin{aligned}E(X_n) = &\sum_{k=0}^{\infty}\big(1 - \exp(-\delta n\rho^k)\big)\\ &+ O\bigg(n\rho^n + n\exp(-\delta n\rho^{k_0}) + \sum_{k=k_0}^{n}\exp(-\delta n\rho^k)(n^{-2} + k\rho^k + k^2\rho^k n^{-1} + kn\rho^{2k})\bigg)\end{aligned}$$

or

$$E(X_n) = \sum_{k=0}^{\infty}(1 - \exp(-\delta n\rho^k)) + O\bigg(n^{-1} + \sum_{k=k_0}^{n}\exp(-\delta n\rho^k)(k\rho^k + kn\rho^{2k})\bigg).$$

For $k_0 \le k \le \frac{\log n}{2\log a}$, the error term is $O(n\exp(-\delta\sqrt{n}\,))$, for $\frac{3\log n}{\log a} \le k \le n$, it is $O\big(\frac{\log n}{n^3}\big)$. In the remaining interval, we have, for $k \le \frac{\log n}{\log a}$,

$$\exp(-\delta n\rho^k)(k\rho^k + kn\rho^{2k}) \ll n\log n\exp(-\delta n\rho^k)\rho^{2k} \ll \frac{\log n}{n},$$

and, for $k \ge \frac{\log n}{\log a}$,

$$\exp(-\delta n\rho^k)(k\rho^k + kn\rho^{2k}) \ll \log n\exp(-\delta n\rho^k)\rho^k \ll \frac{\log n}{n}.$$

So we finally arrive at

$$E(X_n) = \sum_{k=0}^{\infty}\big(1 - \exp(-\delta n\rho^k)\big) + O\Big(\frac{\log^2 n}{n}\Big).$$

The infinite sum is a standard example for an application of the Mellin transform—the asymptotic formula

$$\sum_{k=0}^{\infty}\big(1 - \exp(-x/a^k)\big) = \log_a x + \frac{\gamma}{\log a} + \frac{1}{2} + \phi(\log_a x) + O\Big(\frac{1}{x}\Big)$$

can be found in [FGD95, Szp01] for instance. Here, ϕ is the function given in (**??**). We only have to set $x = \delta n$ to obtain the desired formula (**??**).

Finally note that the expression $1-\exp(-\delta n\rho^k)$ is, apart from renormalisation, the distribution function of a Gumbel (extreme value) distribution, so that the limiting distribution of X_n is the Gumbel distribution. The same distribution is observed, for instance, for the *depth of tries*, as discussed in [LP06]. □

Let us apply this theorem to the parameters MaxPlateau and MaxLevel. For the maximum size of a plateau to be $\leq k$, we know the generating function to be

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

Hence, in the notation of Lemma 1,

$$R(z,u)=(1-z^2+uz^3)(1-4z+3z^2+uz^3),$$

$$r_0(z)=(1-z)^2(1+z)(1-3z) \quad \text{and} \quad r_1(z)=2z^3(1-z)^2.$$

It follows that $\rho=\frac{1}{3}$, $\delta=\frac{1}{18}$, and so the maximum size of a plateau follows a Gumbel distribution, where the mean is asymptotically given by

$$\log_3 n-\log_3 2+\frac{\gamma}{\log 3}-\frac{3}{2}+\phi(\log_3 n-\log_3 2).$$

Similarly,

$$\frac{1-z-\sqrt{1-2z-3z^2+8z^{k+3}-4z^{2k+4}}}{2z^2(1-z^{k+1})}$$

is the generating function for MaxLevel. We see that

$$R(z,u)=1-2z-3z^2+8uz^3-4u^2z^4, \quad r_0(z)=1-2z-3z^2 \quad \text{and} \quad r_1(z)=8z^3.$$

Therefore, $\rho=\frac{1}{3}$ again, $\delta=\frac{2}{9}$, and the mean of MaxLevel is asymptotically

$$\log_3 n+\log_3 2+\frac{\gamma}{\log 3}-\frac{3}{2}+\phi(\log_3 n+\log_3 2).$$

5 Unary-binary trees

These trees are defined by the equation $C=1+zC+zC^2$; they can be obtained from extended binary trees given by $B=y+zB^2$ (z marking internal nodes, y marking leaves) by the substitutions

$$z\longrightarrow\frac{z}{1-z}, \qquad y\longrightarrow\frac{1}{1-z};$$

compare e.g. [FP86].

Let us now consider the maximal length of a sequence of unary nodes; call the corresponding random variable X_n. Then the generating function of unary-binary trees where this parameter is $\leq k$ is obtained from the substitutions

$$z\longrightarrow\frac{z(1-z^{k+1})}{1-z}, \qquad y\longrightarrow\frac{1-z^{k+1}}{1-z},$$

yielding

$$f_k(z) = \frac{1 - z - \sqrt{1 - 6z + z^2 + 8z^{k+2} - 4z^{2k+3}}}{2z(1 - z^{k+1})}.$$

Note that

$$\frac{1 - z - \sqrt{1 - 6z + z^2}}{2z}$$

is the generating function of the *large Schröder numbers*, see [Deu01].

In the notation of Lemma 1, we have

$$R(z, u) = 1 - 6z + z^2 + 8z^2u - 4z^3u^2, \quad r_0(z) = 1 - 6z + z^2 \quad \text{and} \quad r_1(z) = 8z^2.$$

Consequently, $1/\rho = 3 + 2\sqrt{2}$, $\delta = 3\sqrt{2} - 4 = \sqrt{2}\rho$, and the average of the parameter X_n is asymptotically given by

$$\frac{\log n}{\log(3 + 2\sqrt{2}\,)} + \frac{1}{2\log_2(3 + 2\sqrt{2}\,)} + \frac{\gamma}{\log(3 + 2\sqrt{2}\,)} - \frac{1}{2} + \phi\Big(\frac{\log n}{\log(3 + 2\sqrt{2}\,)} + \frac{1}{2\log_2(3 + 2\sqrt{2}\,)}\Big).$$

Analogously, let the parameter Y_n be the maximal length of the chains emanating from the leaves. Then we have to perform the substitutions

$$z \longrightarrow \frac{z}{1 - z}, \qquad y \longrightarrow \frac{1 - z^{k+1}}{1 - z},$$

yielding

$$\frac{1 - z - \sqrt{1 - 6z + z^2 + 4z^{k+2}}}{2z}.$$

In the notation of Lemma 1, we have

$$R(z, u) = 1 - 6z + z^2 + 4z^2u, \quad r_0(z) = 1 - 6z + z^2 \quad \text{and} \quad r_1(z) = 4z^2.$$

Consequently, $1/\rho = 3 + 2\sqrt{2}$, $\delta = \frac{\rho}{\sqrt{2}}$, and the average of the parameter Y_n is asymptotically given by

$$\frac{\log n}{\log(3 + 2\sqrt{2}\,)} - \frac{1}{2\log_2(3 + 2\sqrt{2}\,)} + \frac{\gamma}{\log(3 + 2\sqrt{2}\,)} - \frac{1}{2} + \phi\Big(\frac{\log n}{\log(3 + 2\sqrt{2}\,)} - \frac{1}{2\log_2(3 + 2\sqrt{2}\,)}\Big).$$

If we don't distinguish between leaves and internal nodes, then it amounts to take $B = y + zB^2$ as before, but use the substitutions

$$z \longrightarrow \frac{z}{1 - z}, \qquad y \longrightarrow \frac{z}{1 - z},$$

with the result

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

Then we are in the "Motzkin-world" (as opposed to the "Schröder-world"). Indeed, applying the substitutions $z \longrightarrow \frac{z(1-z^{1+k})}{1-z}$, $y \longrightarrow \frac{z(1-z^{1+k})}{1-z}$ (maximal length of a unary chain $\leq k$), we obtain, apart from a

factor z, the same generating function (and thus the same asymptotics) as for the parameter MaxLevel in Motzkin paths. If, on the other hand, we only consider unary chains emanating from the leaves, we have to use the substitutions $z \longrightarrow \frac{z}{1-z}$, $y \longrightarrow \frac{z(1-z^{1+k})}{1-z}$, yielding the generating function

$$\frac{1-z-\sqrt{1-2z-3z^2+4z^{k+3}}}{2z}.$$

In the notation of Lemma 1, this means that

$$R(z,u) = 1-2z-3z^2+4z^3u, \quad r_0(z) = 1-2z-3z^2 \quad \text{and} \quad r_1(z) = 4z^3.$$

Consequently, $1/\rho = 3$, $\delta = \frac{1}{9}$, and the average of the parameter Y_n is asymptotically given by

$$\log_3 n + \frac{\gamma}{\log 3} - \frac{3}{2} + \phi(\log_3 n).$$

References

[Deu01] E. Deutsch. A bijective proof of the equation linking the Schröder numbers, large and small. *Discrete Math.*, 241:235–240, 2001.

[DSV04] T. Došlić, D. Svrtan, and D. Veljan. Enumerative aspects of secondary structures. *Discrete Math.*, 285:67–82, 2004.

[FGD95] P. Flajolet, X. Gourdon, and P. Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144:3–58, 1995.

[FO90] P. Flajolet and A. M. Odlyzko. Singularity analysis of generating functions. *SIAM J. Discrete Math.*, 3:216–240, 1990.

[FP86] P. Flajolet and H. Prodinger. Register allocation for unary-binary trees. *SIAM Journal on Computing*, 15:629–640, 1986.

[FS07] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2007?

[HP03] C. Heuberger and H. Prodinger. Carry propagation in signed digit representations. *European J. Combin.*, 24:293–320, 2003.

[Knu78] D. E. Knuth. The average time for carry propagation. *Nederl. Akad. Wetensch. Indag. Math.*, 40(2):238–242, 1978.

[LP06] G. Louchard and H. Prodinger. Asymptotics of the moments of extreme-value related distribution functions. *Algorithmica*, 46:431–467, 2006.

[MM78] A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.*, 30:997–1015, 1978.

[Neb01] M. Nebel. Combinatorial properties of RNA secondary structures. *J. Comput. Biol.*, 9:541–573, 2001.

[Szp01] W. Szpankowski. *Average case analysis of algorithms on sequences*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2001.

[VdC85] G. X. Viennot and M. Vauchassade de Chaumont. Enumeration of RNA secondary structures by complexity. *Math. Med. Biol. Lecture Notes Biomath.*, 57:360–365, 1985.